Measuring Behavior 2024

13th International Conference on Methods and Techniques in Behavioral Research, Aberdeen, Scotland

15 – 17 May 2024

Proceedings

Proceedings of Measuring Behavior 2024

13th International Conference on Methods and Techniques in Behavioral Research in Aberdeen, Scotland, May 15 – 17

DOI: 10.6084/m9.figshare.25897855

ISBN: 9789074821957

Volume Editors

Andrew Spink

Noldus Information Technology; andrew.spink@noldus.com

Gernot Riedel

University of Aberdeen; g.riedel@abdn.ac.uk

Khiet Truong

University of Twente; k.p.truong@utwente.nl

Lianne Robinson

University of Aberdeen; lianne.strachan@abdn.ac.uk

Table of Contents

Preface

Measuring Behavior 2024

One glance at these Proceedings, and it is clear that AI-analysis is taking off in our field. Whilst large language models like ChatGPT have taken the world by storm in the last year, specialized applications like computer-vision based techniques to track the behaviors of animals have gone from straightforward location tracking a few years ago to the detailed and deep insights into a range of behaviors and mental states today. For the first time this year, Measuring Behavior has no less than three sessions with 'AI' in their title, and many other sessions have papers including some aspect of AI-based measurement or analysis. The majority of talks present diverse new possibilities that AI brings to our field, although some also mention risks and challenges.

We also have more talks about measuring inner mental or affective states this year than in some previous years. That is the case both for human studies, for instance for measuring cognitive workload, and for animal studies measuring aspects such as pain and emotions. Both the development of new measurement methods and the increased ease of use of those methods have meant that they have been more widely adopted in recent years.

A third and very welcome trend is the attention given to animal welfare. Whilst there have always been talks on the subject at Measuring Behavior meetings, symposia have sometimes only taken place after strong promotion by the program committee. This year, we have two substantial sessions on animal welfare and a number of other talks on the topic as well.

Traditional conference elements such as sessions on behavioral tests, animal models of diseases, and tutorials and demonstrations about setups for measuring behaviors are still present and have led to a varied and interesting program with something for everyone. For many delegates, the meeting is a chance to meet up again with familiar faces from the community, and for others it is an opportunity to make new acquaintances for the coming years.

The previous Measuring Behavior was planned to take place in Krakow, Poland. Unfortunately, COVID intervened and after a couple of delays, we were forced to have the meeting online. Whilst that was still a fruitful meeting, we are very much looking forward to seeing each other face-to-face in Aberdeen this year. We hope that you will enjoy the meeting and your stay in Aberdeen.

Andrew Spink, Gernot Riedel & Khiet Truong

Chairs, Measuring Behavior 2024

Symposia

Symposium: Measuring behaviour & physiology in and around the cockpit

Assessing progress in flight performance in a virtual reality simulator

Ivo Stuldreher[1], Wietse Ledegang[1], Eric Groen[1]

[1] Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, The Netherlands; ivo.stuldreher@tno.nl

Introduction

Currently, student pilots of the Royal Netherlands Air Force (RNLAF) receive most of their initial flight training in a turbo-prop training aircraft. The RNLAF considers introducing Virtual Reality (VR) flight simulators as an additional training means. One of the advantages of VR compared to real aircraft is that it allows for the recording of flight parameters which potentially can be used for objective evaluation of the student's performance. In the current study we explored objective performance measures to objectively assess the progress in flight performance of student pilots in VR.

Methods

The simulator environment (multiSIM B.V.) consisted of a fixed-base cockpit of a PC-7 aircraft and control devices with control loading (see Figure 1a). The cockpit and outside visual environment were presented in a VARJO-Aero VR headset with built-in eye-tracker. With this simulator a simplified two-day training program was developed by an RNLAF flight instructor, comprising three VR sessions. In each session, three basic flight maneuvers were practiced in a fixed order, being straight-and-level flight (SAL), a speed change (SC) and a level turn (LT). Each maneuver was practiced three times in a row during 210-second runs, followed by a fourth run in which the maneuver was performed while simultaneously performing an auditory memory task. The four runs of SAL were followed by four runs of SC and four runs of LT. We here only discuss the results of runs without the memory task, totaling to nine runs for each maneuver.

Fifteen military cadets (12 males and 3 females; 23.7 ± 2.4 years old) of the Royal Military Academy participated in the study, and performed all three VR sessions. These participants had negligible flying experience. As a reference, six RNLAF pilots, who recently completed their initial flight training, performed only the first VR session. All participants signed an informed consent prior to participating. During the VR sessions, the participants did not receive any feedback on their performance, and learning was based on standardized instruction material, which was studied at the beginning of each new VR session

For the data analysis, flight parameters were extracted from the VR system, computed as the deviation or 'error' from a certain desired value, such as the error from the desired altitude (which was always 5000 ft), desired airspeed, desired bank angle, or desired heading. These desired values were known to participants. To investigate which parameters reflected flight performance of each maneuver, we selected those parameters which significantly changed with repetition or 'run', according to an ANOVA analysis. To compare performance across flight parameters, these measures were normalized in relation to the largest error observed across all participants and a zero error. These normalized performance measures range from zero to one, where a performance of zero corresponds to the largest error observed across participants and a performance of one corresponds to no error with respect to the desired target value. The normalized performance measures were then averaged to obtain an overall performance measure.

The validity of this overall performance measure was compared with subjective ratings of an RNLAF flight instructor. After the experiment, the instructor rated a selected set of recorded runs for overall performance, basic aircraft control, lookout performance and multi-tasking performance. These ratings ranged between 1-3 (unsatisfactory), 4-6 (fair), and 7-9 (good). The set included eleven runs per flight maneuver, consisting of three recordings from three students (i.e., nine runs per maneuver), and two runs from a pilot, cumulating to 33 recordings in total.

Finally, the participants' progress in flight performance was estimated by fitting two linear functions to the overall normalized performance for each flight maneuver separately, as depicted in Figure 1b. Here, the first linear function represents learning speed, where the intercept corresponds to the start level and slope to the learning rate. In the second function the intercept refers to the end level performance. This end level was defined as the normalized flight performance in run 11. That leaves two free parameters, the time to end level and the start level, which were estimated through a non-linear least squares method. Although we are aware that learning is no linear process, this simplified approach allows still allows for assessing learning rate and end level, as good learners should both learn fast and reach a high end level.

Results

For the SAL condition there was a significant effect of run on airspeed, roll, altitude and heading. For the SC condition there was a significant effect of run on altitude, airspeed and heading. For the LT condition there was a significant effect of run on altitude, roll and side slip. Subsequently, each of these performance parameters were normalized and combined into an overall mean normalized performance for each flight maneuver separately. ANOVA analyses revealed a significant effect of run on the normalized performance metric for each flight maneuver, indicating learning took place and approaching the performance level of RNLAF pilots.

Positive and significant correlations were found between the normalized performance metric, that was aggregated across conditions, and instructor ratings on 'overall performance' ($r = .76$, $p <.001$), 'basic aircraft control' ($r = .70$, $p < .001$) and 'multi-tasking' ($r = .59$, $p <.005$), but not with the instructor rating for 'lookout performance' ($r = .37$, $p = .087$). Figure 3c depicts how the instructor ratings of unsatisfied, fair and good correspond to the objective flight performance measure.

The learning curve fits showed rather good results, with average R^2 of .93 for the SAL maneuver, .69 for the LT maneuver, and .83 for the SC maneuver. Figure 3c depicts the progression in flight performance and function fit, averaged over participants and averaged over the three flight maneuvers.

Conclusion

In this study we explored how the progress in manual flight performance of student pilots can be described with objective measures extracted from a VR flight simulator. The results suggest that normalization of the relevant flight parameters from each flight maneuver allows for the computation of an overall performance measure. The high correlation with instructor gradings suggests that, for the limited set of maneuvers, the student's progress in manual flying skills can objectively be assessed in the VR flight simulator. Although flying an aircraft requires more than just manual flying skills, this finding suggests that the non-pilot participants acquired manual flying skills to a level approaching that of more experienced pilots. Furthermore, it seems that, for these relatively elementary flight maneuvers, progress in flight performance can be adequately quantified by a learning curve composed of two linear functions. Future work should consider a broader set of flight maneuvers that would better reflect all aspects of flight training.

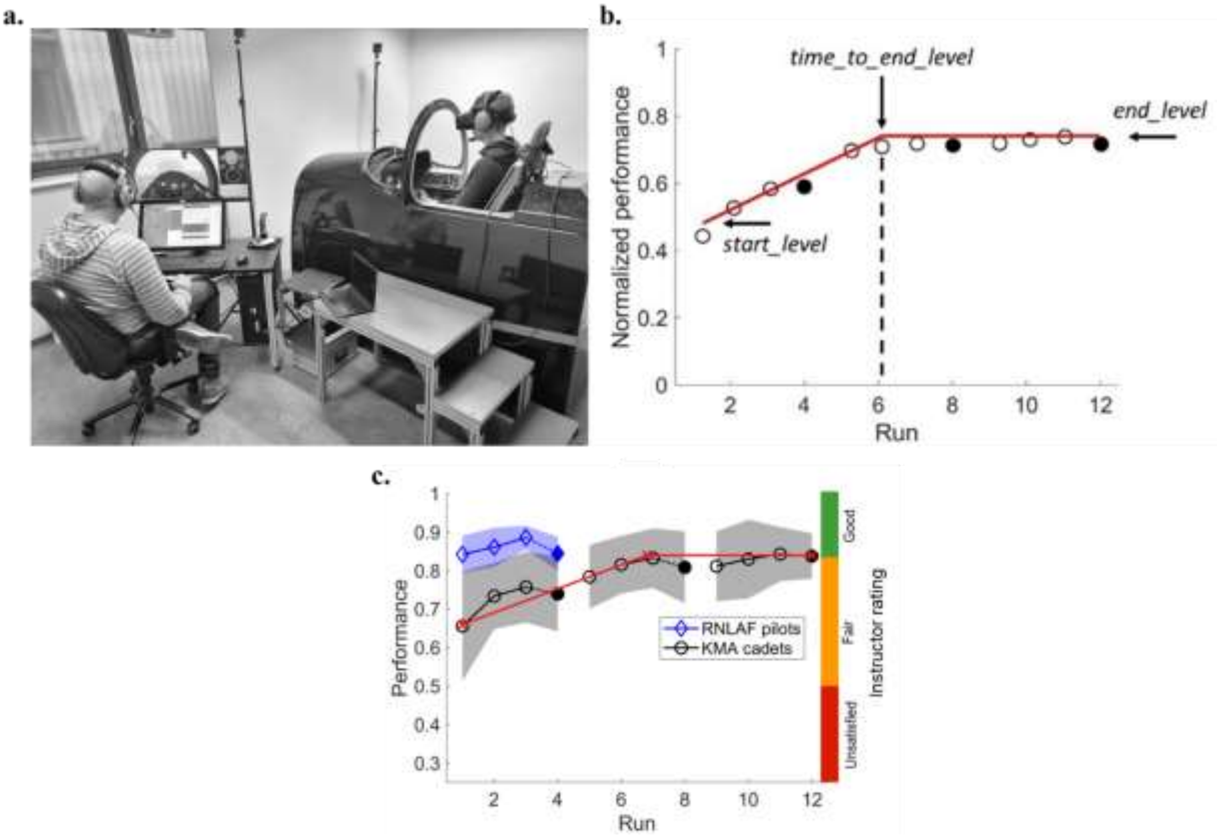


Figure 1 a. Setup of the VR simulator during the experiment, with a test leader behind the instructor station and a subject inside the cockpit mock-up, wearing the VR- headset and headphone. b. Visualization of the characterization of a learning curve (red line) by two linear functions. The data points reflect mock data for illustration purpose only, where open markers reflect data points that are considered in curve fitting and filled markers reflect data points that are not considered in curve fitting. The first function describes the learning effect (i.e., the increment in performance over time), whereas the second curve describes the end level of the individuals maximum performance. The start_level describes the experience of the participant, the time_to_end_level describes the number of runs needed to reach end_level performance. c. Overall learning curve averaged over the three flight maneuvers in relation to the instructor grading.

EEG characterization of dynamic complex processes and rare events to understand operator's activity in aeronautical context

B. Somon[1], A. Campagne[2], M. Salomone2 and B. Berberian[1]

1. ONERA/DTIS, 13661 Salon Cedex Air, France. {bertille.somon ; bruno.berberian}@onera.fr
2. Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France. {salomomi; aurelie.campagne}@univ-grenoble-alpes.fr

Introduction

In recent years, more and more researchers have questioned the relevance of lab-based controlled experiments to characterize operational everyday life activity. Several criticisms arise not only at the conceptual, but also at the methodological level. At the conceptual level, researchers have pointed out that the complexity, stakes and engagement level required in lab-based experiments is usually very reduced compared to what can be observed in real-life environments. Moreover, laboratory experimental paradigms mainly assessed in a controlled manner the effect of a specific factor on a specific cognitive function, but few of them evaluated the applicability of these results in more operational contexts, and considered the interactions and covariations between cognitive mechanisms which can arise in such contexts [1]. Researchers have tried to prevent this by developing novel controlled experimental paradigms such as the Multi-Attribute Task Battery (MATB-II; [2]) or adaptation of the Overcooked! game [3], which tend to mimic everyday-life activities and introduce increased levels of complexity. Notably, this complexity is based on the concomitant use of several tasks which challenge various cognitive functions. At the methodological level, experimental designs, especially in cognitive neuroscience, tend to assess cognitive functions and their variations at a discrete event-based level. Most studies require the repetition of tens of stimuli to: i) reduce the impact of the intra- and inter-individual variability; ii) ensure the statistical power and reproducibility of the results [4]; and iii) ensure the effect of a specific variation on a specific function [5]. Even though some research has focused on continuous [6] or block instead of event-based analyses of activity, these measures are currently often used to monitor very global operator states, such as fatigue, vigilance, or cognitive workload. Additionally, event-related potentials (ERP) estimation remains the most employed electroencephalographic (EEG) analysis method and requires the experimenters to introduce task-irrelevant events or specific stimuli in various types and modality, likely to bias the task in hand. This has been applied in a few brain-computer interface (BCI) or operational research, where cues [7] have been used.

This question of continuous or dynamic monitoring is particularly relevant in complex and/or risky operational situations. Recent advanced technologies and the advent of Artificial Intelligence (AI) have led to increase the automation level of systems. We now interact daily with automated systems in various activity sectors, particularly in the field of aeronautics and rail transports and, more recently in the field of the automotive with the growing accepted use of automated cars [8]. Although the benefits of such automated systems are undoubtable [9], several studies have also demonstrated that there are drawbacks to the change in operator activity driven by this automation [10]. Operators who previously performed specific actions now shifted to a supervisory role and therefore had to adapt their routines and skills, leading to new unforeseen difficulties in human-system interactions. These difficulties, such as a loss of motor skills, attention decrements, increased complacency and lack of system understanding, are coined the out-of-the-loop (OOL) performance problem [11]. Although this phenomenon has been widely studied, especially in the human factors for aeronautics community, it remains very difficult to characterize, and even more to compensate. Two difficulties arise: a theoretical and a technical one. At the theoretical level, the OOL has been described in terms of situation awareness or high-level constructs and concepts, providing low explainability regarding the whys and wherefores of this phenomenon. The folk-model-like description of the OOL prevents us from identifying properly the cognitive mechanisms implicated in its emergence and anticipate their variations [12]. Still, the OOL has been identified to emerge in highly complex environments in which the operator remains a supervisor of the automated system. In this context, when removed from the control loop, the operator is often unable to detect errors from the automated system and compensate them whenever necessary. Thus, we have proposed that the performance (or error) monitoring mechanism, widely studied by the cognitive neuroscience community, might be a good concept to characterize the OOL [12].

Additionally, we believe that using the methods of the cognitive neuroscience community, such as cerebral activity recording methods (e.g., EEG), can help us have a better understanding and characterization of these applied difficulties. Nevertheless, we have been faced with the technical difficulty related to the OOL characterization. Studies on the neurophysiological correlates of performance monitoring activities have focused mainly on event-related potentials time-locked to participants' responses [13]. However, the long time-scale of emergence of the OOL phenomenon requires more continuous measures to be performed. Additionally, the activity usually studied in performance monitoring is the Error-Related Negativity (ERN; [14]), which is very difficult to reconcile with the very reliable systems used in the aeronautics context. Finally, these activities are time-locked specifically to a participants' responses. In the context of the OOL, the system is highly automated and requires no response from the participant which limits the use of these measures. In this short paper, we describe the framework and the experimental protocols we used as well as the brain activity analysis approach we adopted to characterize the supervisory activity of an automated system and the out-of-the loop phenomenon, taking into account the difficulties associated with the study context and previously described. The limits of this approach and adaptations to improve characterization of the brain activity underlying the cognitive mechanisms involved are also explained.

Rationale

Concerning the performance monitoring framework, we have seen through a review of the literature that, even though performance monitoring associated to system supervision was not a very studied topic, there were more and more studies related to understanding the brain activity associated with another human agent's error detection [15]. These studies partly emerged from the Joint Action community and social neuroscience, in order to understand when, how and why we detect the errors committed by another human being [16]. One interesting aspect of these studies is the fact that researchers evidenced performance monitoring brain components based on EEG measures during supervision and error observation [13]. This knowledge led us to develop an experimental paradigm in order to compare the brain activities associated with human and artificial agent error monitoring. These activities not being fully characterized, we had to proceed in a step-by-step approach, from a lab-based to a more applied environment, which was our target context. In a first experiment, we used a well-known experimental paradigm in the cognitive neuroscience community – the Eriksen flanker task – that we adapted as an aeronautics-based supervision task [17]. In a second step, we tried to transfer the observed results and to adapt the EEG analysis to a more dynamic and ecological task, consisting of the supervision of an aeronautics-based conflict avoidance simulator [18]. Additionally, we also assessed through these studies, the effect of time-on-task on performance monitoring brain correlates and their evolution during the OOL phenomenon.

Experimental studies

In this section, we will present briefly the paradigms we have used to characterize brain activity related to supervision activity and its evolution across time during the OOL phenomenon.

Material and methods

In a first experiment (see [17] for full details), we recorded the EEG activity of 17 healthy participants (12 men; 27.5+/-4.78 y.o.) during another human agent or system supervision in a modified supervision arrowhead version of the Flanker task [19] [20]. During this task, participants had to supervise the decisions of another agent (human or automated system) on the orientation of a central target arrow on a screen among or without distracters. Participants performed a total of 40 blocks of 72 trials over 2 experimental sessions. Each trial (see Figure 1a) presented a stimulus for 10ms followed by a short reaction time window after which the other agent's response was displayed. Finally, the question "Error?" was displayed, and the participant had one second to tell whether the agent's response was correct or not. The participant supervised a human agent (introduced to him at the beginning of the task) for half of the blocks and an artificial agent for the other half. The type of agent performing the task was provided at the beginning of each block. In reality, all blocks were performed by a computer, which accuracy was 66.6% and reaction times were based on the participant's own reaction times to the same flanker task performed previously. This ploy was used in order to ensure the same number of error and correct trials, as well as prevent effects of trust or differential learning across participants. Additionally, two levels of difficulty were introduced for the flanker task, as a follow up of a previous experiment [20], but will not be discussed here.

For the two experimental sessions, participants' subjective (difficulty levels), behavioral (error detection rate and d') as well as neurophysiological (75 Ag-AgCl active electrode EEG) activity were recorded. The EEG signal was amplified using an actiCHamp system (Brain Vision, LLC), digitized at a 24-bit rate and sampled at 1kHz, with a 0.05 µV resolution. Classical signal pre-processing tools (i.e., band-pass filtering, artifact removal with ICA) were used in order to clean the signal (see [17] section 2.3.2. for a full description). In this first study, consistently with the literature on error supervision, event-related potentials were time-locked to the other agent's response display and analyzed at fronto-central sites [21]. In accordance with the literature, mean peak amplitude of the N2 and P3 ERPs (likened to the observational versions of the error-related negativity and positivity [13]) were computed and analyzed with a repeated-measures analysis of variance with accuracy – error vs. correct – difficulty – easy vs. difficult – and agent type – human vs. system – as within-subject factors. Going further, given a potential spatio-temporal dynamic in supervisory activity and inter-individual variances, a more robust cluster-based permutation test [22] (i.e., requiring no a priori on ERP location or latency) was performed.

In the second experiment (see [18] for full details), we recorded the EEG activity of 20 healthy participants (7 women; 27.75+/-1.42y.o.) performing a dynamic conflict avoidance simulator supervision task. Participants had to assess whether simulator choices to avoid obstacles by turning right or left was correct or erroneous depending on the context surrounding the aircraft. Participants completed a total of 12 experimental blocks of 25 trials over 2 sessions. Each trial (see Figure 1b) started with a first initialization phase lasting 2 to 5 seconds during which the aircraft flew straight, then conflicting obstacles arrived and were detected by the simulator, then the simulator's avoidance decision was displayed. At this moment the simulation was frozen, providing no new evidence for the participant to identify the correct or erroneous response. After 1 second the participant was asked to provide his response stating whether the simulator's response was correct or erroneous. Once the participant had given his response, the simulation was started again and the simulator performed the avoidance. Finally, feedback was displayed to the participant, indicating the correctness of the simulator's response. As described more precisely in [18], and as a follow up of the previous experiment, there were two difficulty levels corresponding to obstacles being aligned or dispersed throughout the entire screen (respectively easy and difficult condition). Across the entire sessions, the accuracy of the simulator was set at 60%, in order to ensure a sufficient number of trials to be analyzed for both accuracy conditions.

For the two sessions, participants' subjective (difficulty levels), behavioral (hit rate) as well as neurophysiological (64 Ag-AgCl active electrode EEG) activity were recorded. The signal was acquired and preprocessed with similar parameters as in the first experiment. (see [18] section 2.3.3. for a full description). In this second study, a first classical analysis on ERPs was performed in order to see the transferability of previous results to more ecological and dynamic tasks. Given the low transferability of previously observed results to this dynamic task, we completed this classical analysis with time-frequency analyses allowing to assess more dynamically brain activity variations. Both analyses were time-locked to the display of the system's response. To avoid a priori assumptions about the supervisory brain activity in the studied dynamic environment, we performed a cluster-based permutation test to identify the specific spatio-temporal clusters differentiating the various experimental conditions. EEG analyses were averaged across participants depending to the system accuracy – error vs. correct – task difficulty – easy vs. difficult – but also to the moment of the experiment – beginning vs. end – as within-subject factors. For all analyses, only the participant correct evaluations of the simulator's responses were analyzed. The impact of time-on-task (moment of the experiment) on the performance monitoring activity was assessed by averaging activity of the first two blocks (beginning) and the last two blocks (end) of each session.

Results

In the first experiment (see [17] for a full description of results), despite an effect of the difficulty level on both the subjective feedback ($diff_{easy}$= 3.35 ± 0.46, $diff_{difficult}$= 4.60 ± 0.58; t[14] = 3.073, $p < .01$) and the stimuli detectability (d'_{easy}=2.42, $d'_{difficult}$=1.57; F(1,16) = 17.24, $p < .001$, $\eta^2_P = .52$), this effect was not observed on the error detection rate (F(1,16) = 0.48, p = .5). Interestingly there was no difference either of the type of agent (human vs. System) on neither the error detection rate, nor the d' measures.

At the physiological level (see Figure 2a), we identified a significant N2-P3 component at the FCz electrode associated with error detection. The P3 was present only for error detection (main effect of accuracy on peak P3 amplitude; F(1,16) = 69.42, p < .005, η^2_P = .81). Additionally, we observed a significant effect of difficulty on both the N2 (F(1,16) = 4.60, p < .05, η^2_P = .22) and P3 (F(1,16) = 15.77, p < .005, η^2_P = .50) components. This classical ERP analysis revealed no significant effect of the type of agent supervised on these performance monitoring components. Yet, another more robust analysis (cluster-based permutation test) revealed a significant difference of activity with a broad positive component, differentiating significantly between human agent and system error detection. Interestingly, this component had the same characteristics as the P3 in terms of shape and latency.

In the second experiment (see [18] for a full description of results), similarly to the first experiment, participant subjectively reported a significant difference between the easy and difficult conditions ($diff_{easy}$=1.07 ± 0.13, $diff_{difficult}$=2.12 ± 0.15; t[17] = −5.0675, p < .005) but there was no effect on the objective performances (t[17] = −0.35, p = .73) with the average hit rate equal to 97.61 ± 0.57%. Additionally, the moment of the experiment (beginning vs. end) showed no effect on behavioral performances.

At the physiological level (see Figure 2b), the classical ERP analysis revealed no cleared component associated with the observation of system's correct or erroneous responses. Cluster-based permutation tests identified a small negative posterior cluster differentiating between error-related and correct-related brain activity. However, the shape, location and latency of this cluster were difficult to associate to usual performance monitoring ERP. In contrast, the use of a time-frequency decomposition (see Figure 2c), better suited to the temporal variances of brain activity, allowed us to identify a brain cluster in the theta frequency band whose activity was significantly increased in response to system errors compared to correct system responses. Additionally, theta activity has been modulated by the moment of the experiment within a cluster, with a higher left-lateralized fronto-temporo-parietal theta activity at the beginning compared to the end of the experiment, regardless of system response and also following system errors. Finally, several clusters associated with low alpha activity also showed increased amplitude at the beginning of the experiment, and decreased amplitude for correct responses compared to system errors at the beginning.

Discussion and way forwards

The goal of these works was to characterize the brain activity variations related to system performance monitoring activity and their evolution during the OOL phenomenon. The use of brain measurements appears of major interest to better understand and characterize supervision activity with regard to subjective and behavioral data, likely to be less sensitive to modulation factors such as task difficulty, as illustrated by our two studies and the literature, even in more applied contexts [23] [24]. In both experiments, task difficulty modulated subjective feedback as well as physiological activity related to performance monitoring (for both ERPs and time-frequency measures) but did not show any effect on supervision performances. Likewise, in the second experiment, the moment of the experiment (beginning vs. end) showed a significant effect on brain markers related to performance monitoring activity and vigilance state (respectively in the theta and alpha frequency bands) but did not show any impact on performances.

However, although current EEG measures – especially ERPs – constitute good biomarkers to characterize performance monitoring activity in lab-based environments, as also illustrated in other prior studies [24], their transferability to more operational and dynamic contexts, remains difficult. In this context (second study), we were unable to reproduce the ERP results related to the supervision activity of system's erroneous and correct responses highlighted in the first study. Greater temporal variability in the timing of system error detection and greater uncertainty in identifying correct and erroneous responses in this dynamic and more complex environment to process may help to justify this result, given the sensitivity of ERP markers related to supervisory activity with task difficulty. Nevertheless, in the second study, we observed specific error-related variations in the frequency domain, as well as their evolution across time. Based on this result, part of the difficulty we encountered in translating the ERP results in the second study could come from the fact that operational environments would involve, compared to more controlled laboratory environments (as in the first study), several cognitive functions with variable temporality depending on the dynamic evolution of the stimulation environment, which would make the analysis of the neurophysiological correlates associated with each function more difficult [5]. If we wish to better understand the mechanisms at play during the supervision activity and the precursors of its degradation during the OOL phenomenon in more operational and dynamic environments, it appears necessary i) to apply and develop new methods for processing and analyzing EEG signals and ii) to associate other metrics (e.g. data fusion) to better identify and individualize the cognitive functions involved and the associated neural correlates.

Concerning the EEG signal analysis methods, we have seen through our two studies that it is possible to extract new information with new signal processing or analysis methods. During the first experiment the use of a statistical analysis method without *a priori* spatio-temporal of the EEG activity, namely the cluster-based permutation test, allowed us to observe a difference in amplitude between human and system error detection, on what we can link to the P3 component. Given the functional role of the P3 regarding information processing, this result would suggest the establishment of some form of complacency during system supervision. This complacency is a phenomenon that usually occurs during system supervision and would be considered a precursor to the OOL phenomenon [12]. Additionally, the time-frequency processing method also provided more information during the more dynamic experiment. Indeed, the complacency result observed in the first experiment echoes with the time frequency results observed during the second experiment, where we observed an increased theta activity at the beginning compared to the end of the experiment, especially for error detection. Theta activity has generally been associated to cognitive control and performance monitoring [25]. However, this analysis method does have its limitations. As mentioned earlier, in an ecological context, there is more variability in the cognitive processes involved and their temporality. This can make their detection difficult to the benefit of more massive activities representing global operator states (such as vigilance, fatigue, cognitive workload). In this study, the clusters we observed could be composed of several dozen electrodes. These clusters, which spread over several brain regions, could therefore reflect more general states, making it more difficult to characterize complex cognitive processes. Complementary solutions could improve our understanding of the data EEG and performed analyses.

The use of methods such as blind source separation methods or feature identification algorithms (such as Independent/Principal/Canonical Component Analysis – ICA, PCA, CCA – for instance) could also be an interesting avenue to disentangle various sources of activity and therefore might allow to better decipher the different cognitive mechanisms related to the supervision activity and the precursor of the OOL phenomenon. The use of methods allowing to improve the spatial resolution of the surface EEG data such as a surface Laplacian transformation [24], estimating current source density, could also contribute to this objective. The application of this method on the data from a previous study allowed to better distinguish the influence of task difficulty on the different response-locked and feedback-locked performance monitoring ERPs (see [20] for full details). Nevertheless, all these methods do not completely eliminate the difficulty of properly labelling the cognitive process associated with a given brain activity. Thus, these methods could be combined with source reconstruction algorithms (e.g., beamforming), which have the advantage of associating cerebral activity with a specific brain region or structure. In addition to reducing the size of clusters, these analyses can be used to refine the identification of brain activity and thus the characterization of the cognitive processes involved. Finally, we have seen in the second experiment that the data was much noisier in a dynamics experimental context compared to a controlled one. One possibility could be to use more recent data cleaning algorithms, such as the artificial subspace reconstruction (ASR) or neural networks that have proven very useful recently [26].

Concerning the data fusion, one of the objectives is to use additional metrics allowing in particular to better characterize the cognitive processes involved during the supervision activity to better dissociate them on the EEG patterns. Monitoring ocular activity constitutes a candidate of interest in this perspective. In the literature, ocular activity is identified as a window into vigilance, attentional processes as well as complacency [27]. Throughout visual scene exploration, ocular activity characterizes attentional spreading as well as preferential treatment or neglect of specific elements [28]. Ocular activity might thus be a good candidate to identify and characterize the dynamics of exploration and information processing of a visual scene, leading to decision making. This tool would therefore provide additional information for understanding the cognitive operations of operators during activity. This contribution would in turn nourish the understanding and characterization of cerebral markers obtained with EEG during dynamic tasks. The EEG-eye tracking coupling also presents a major methodological interest in the processing of EEG signals, in particular in resolving the problem of spatio-temporal overlap of neuronal activities during the close succession of eye fixations during the visual exploration of information during a dynamic and complex task. Thus, we are currently setting up another experiment aiming at using joint eye-tracking-EEG activity to time-lock performance monitoring and have a better characterization of it related brain components in a dynamic conflict avoidance task (see the ASTRID ANR Project EMOOL). Additionally, recent studies have proposed that the fixation rate could be a good marker for detecting the OOL phenomenon [29]. Notably, variations of trust and complacency tend to modify the ocular activity of operators in highly automated environments [27].

Conclusion

This set of studies aimed to question the neurocognitive mechanisms underlying performance supervision activity and the nature of the precursors at the origin of the out-of-the-loop phenomenon, a widely used concepts which lacks neurocognitive foundations to be characterized more thoroughly, and thus anticipated and countered. We have seen that with this type of concept rising from interactions with highly automated and very reliable systems, it is difficult to define experimental protocols allowing the use of neurophysiological methods classically used in the literature. After showing difficulties in reproducing lab-based measures in more operational contexts, we have faced the difficulty of the number of trials, the dynamic of the task and the time-locking events. We propose promising avenues to address, step by step, these difficulties (signal processing tools, innovative, analyses, physiological data fusion, etc.) and aim for our final goal.

Ethical Statement

Studies presented in this paper were approved by a local ethics committee (CERNI Grenoble, n°IRB00010290-2016-09-13-12 and n°IRB00010290-2017-07-04-20-CERNI_AvisConsultatif-2017-06-13-04) and conducted according to the principles expressed in the revised Declaration of Helsinki.

References

1. P.A. Hancock, "Whither Workload? Mapping a Path for Its Future Development," in *Human Mental Workload: Models and Applications*, Dublin, Ireland, 2017, pp. 3-17.
2. Yamira Santiago-Espada, Robert R. Myer, Kara A. Latorella, and James R., Jr. Comstock, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide.," NASA Langley Research Center, Technical Memorandum 2011.
3. M. Carroll et al., "On the Utility of Learning about Humans for Human-AI Coordination," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 5174-5185.
4. Guiomar Niso et al., "Good scientific practice in EEG and MEG research: Progress and perspectives.," *NeuroImage*, vol. 257, no. 119056, August 2022.
5. Steven J. Luck, *An introduction to the event-related potential technique.*. Cambridge, Massachusetts; London, England: MIT press, 2014.
6. Cameron D. Hassall, Yan Yan, and Laurence T. Hunt, "The neural correlates of continuous feedback processing," *Psychophysiology*, vol. 60, no. 12, p. e14399, 2023.
7. Raphaëlle Nina Roy, Stéphane Bonnet, Sylvie Charbonnier, and Aurélie Campagne, "Efficient workload classification based on ignored auditory probes: a proof of concept.," *Frontiers in Human Neuroscience*, vol. 10, no. 00519, October 2016.
8. Carl Benedikt Frey and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological Forecasting and Social Change*, vol. 114, pp. 254-280, January 2017.
9. Earl L. Wiener and Renwick E. Curry, "Flight-deck automation: Promises and problems.," *Ergonomics*, vol. 23, no. 10, pp. 995-1011, 1980.

10. S. W. A. Dekker and D. D. Woods, "MABA-MABA or abracadabra? Progress on human–automation co-ordination.," *Cognition, Technology & Work*, vol. 4, pp. 240-244, 2002.
11. Mica R. Endsley and Esin O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors*, vol. 37, no. 2, pp. 381-394, June 1995.
12. B. Berberian, B. Somon, A. Sahaï, and J. Gouraud, "The out-of-the-loop Brain: A neuroergonomic approach of the human automation interaction.," *Annual Reviews in Control*, vol. 44, pp. 303-315, 2017.
13. Markus Ullsperger, Adrian G. Fischer, Roland Nigbur, and Tanja Endrass, "Neural mechanisms and temporal dynamics of performance monitoring.," *Trends in Cognitive Sciences*, vol. 18, no. 5, pp. 259-267, March 2014.
14. Anja Riesel, Anna Weinberg, Tanja Endrass, Alexandria Meyer, and Greg Hajcak, "The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks.," *Biological Psychology*, vol. 93, no. 3, pp. 377-385, July 2013.
15. Bertille Somon, Aurélie Campagne, Arnaud Delorme, and Bruno Berberian, "Performance Monitoring Applied to System Supervision," *Frontiers in Human Neuroscience*, vol. 11, no. 360, July 2017.
16. Margherita Adelaide Musco, Elisa Zazzera, Eraldo Paulesu, and Lucia Maria Sacheli, "Error observation as a window on performance monitoring in social cntexts? A systematic review," *Neuroscience and Biobehavioral Reviews*, vol. 105077, February 2023.
17. Bertille Somon, Aurélie Campagne, Arnaud Delorme, and Bruno Berberian, "Human or not human? Performance monitoring ERPs during human agent and machine supervision.," *NeuroImage*, vol. 186, pp. 266-277, 2019.
18. Bertille Somon, Aurélie Campagne, Arnaud Delorme, and Bruno Berberian, "Brain mechanisms of automated conflict avoidance simulator supervision.," *Psychophysiology*, vol. 60, no. 2, p. e14171, 2023.
19. Barbara A. Eriksen and Charles W. Eriksen, "Effects of noise letters upon the identification of a target letter in a nonsearch task," *Perception & Psychophysics*, vol. 16, no. 1, pp. 143-149, 1974.
20. Bertille Somon, Aurélie Campagne, Arnaud Delorme, and Bruno Berberian, "Evaluation of performance monitoring ERPs through difficulty manipulation in a response-feedback paradigm.," *Brain Research*, vol. 1704, pp. 196-206, 2019.
21. Stefanie Enriquez-Geppert, Carsten Konrad, Christo Pantev, and René J. Huster, "Conflict and inhibition differentially affect the N200/P300 complex in a combined go/nogo and stop-signal task," *NeuroImage*, vol. 51, no. 2, pp. 877-87, June 2010.
22. Eric Maris and Robert Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data.," *Journal of neuroscience methods*, vol. 164, no. 1, pp. 177-190, 2007.
23. Luca Longo and M. Chiara Leva, Eds., *Human Mental Workload: Models and Applications*. Dublin, Ireland: Springer, 2017.
24. Liesbet Van der Borght, Femke Houtman, Boris Burle, and Wim Notebaert, "Distinguishing the influence of task difficulty on error-related ERPs using surface Laplacian transformation.," *Biological Psychology*, vol. 115, pp. 78-85, March 2016.
25. J. F. Cavanagh and M. J. Frank, "Frontal theta as a mechanism for cognitive control.," *Trends in Cognitive Sciences*, vol. 18, no. 8, pp. 414-421, August 2014.
26. S. Blum, N. S. J. Jacobsen, M. G. Bleichner, and S. Debener, "A Riemannian Modification of Artifact Subspace Reconstruction for EEG Artifact Handling.," *Frontiers in Human Neuroscience*, vol. 13, no. 00141, April 2019.
27. Gal Ziv, "Gaze Behavior and Visual Attention: A Review of Eye Tracking Studies in Aviation.," *The International Journal of Aviation Psychology*, vol. 26, no. 3-4, pp. 75-104, July 2016.

28. Pieter J.A. Unema, Sebastian Pannasch, Markus Joos, and Boris M. Velichkovsky, "Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration.," *Visual Cognition*, vol. 12, no. 3, pp. 473-494, 2005.
29. G. Di Flumeri et al., "Brain–Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems.," *Frontiers in Human Neuroscience*, vol. 13, no. 00296, September 2019.

Figures

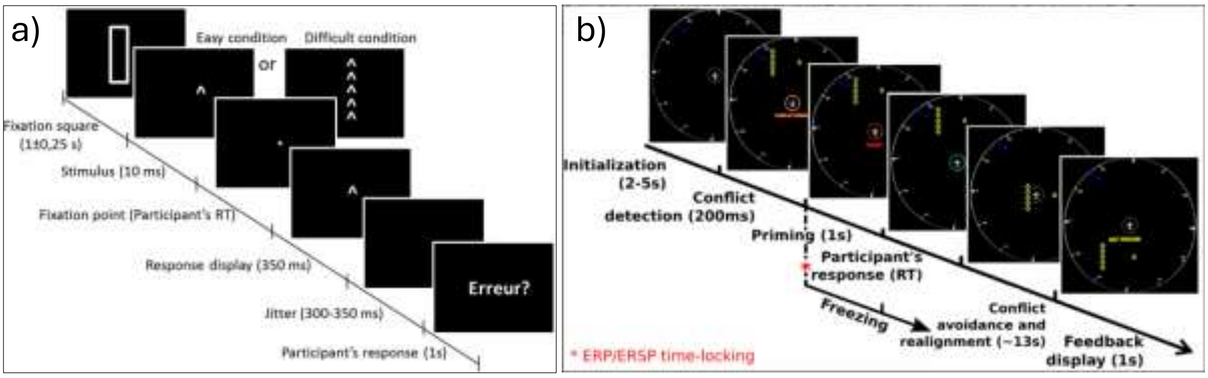

a)
Easy condition
Difficult condition
or
Fixation square (1±0,25 s)
Stimulus (10 ms)
Fixation point (Participant's RT)
Response display (350 ms)
Jitter (300-350 ms)
Participant's response (1s)
Erreur?
b)
Initialization (2-5s)
Conflict detection (200ms)
Priming (1s)
Participant's response (RT)
Freezing
Conflict avoidance and realignment (~13s)
Feedback display (1s)
* ERP/ERSP time-locking

Figure 1. Experimental designs of a) Experiment 1 (from [17]) - one trial of the supervised flanker task: participants had to determine whether the agent executing the task (human or artificial) performed correctly or not in the flanker task by comparing the orientation of the central target arrowhead of the task stimulus and the response given by the other agent; b) Experiment 2 (from [18])– one trial of the conflict avoidance simulator supervision task where participants had to determine whether the simulator was selecting the right direction to avoid obstacles (yellow circles) or not.

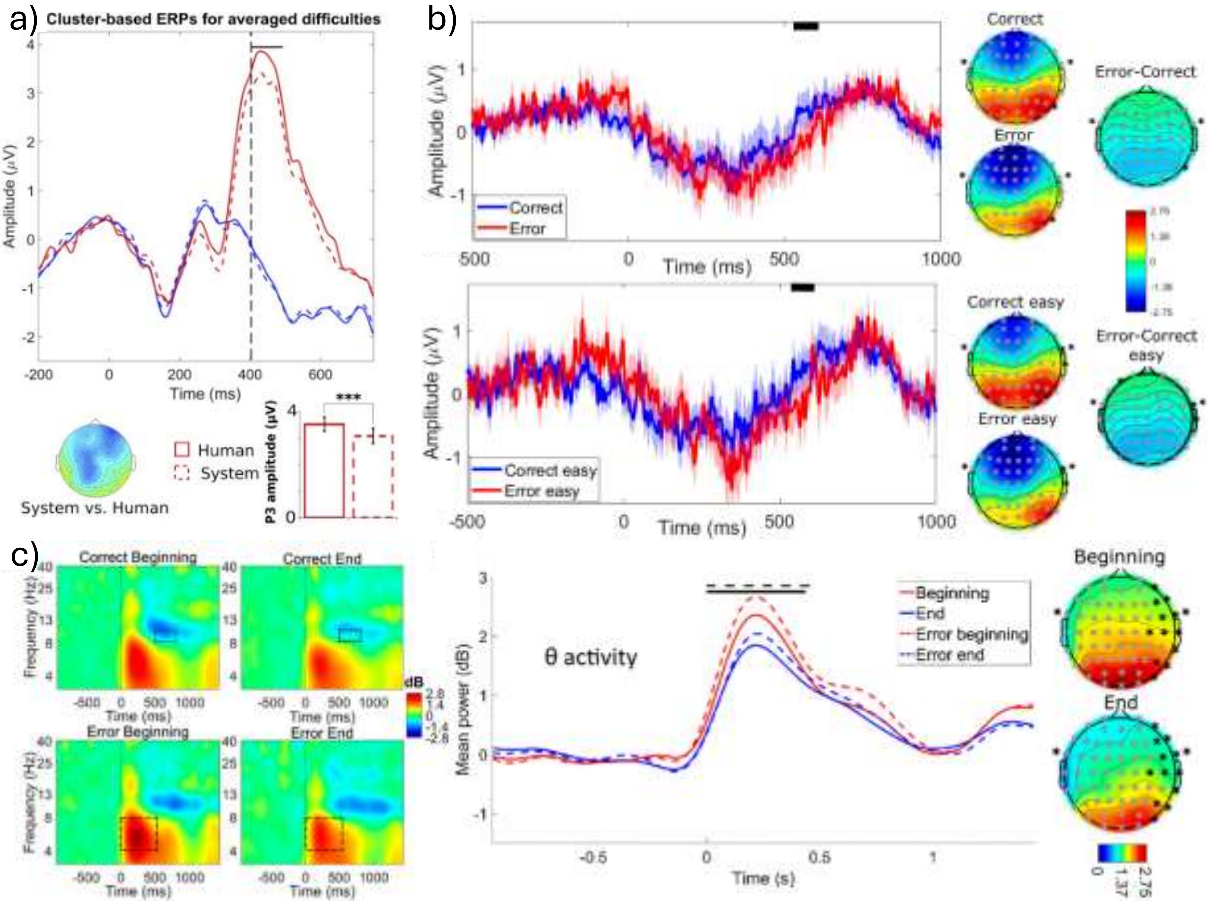


Figure 2. Electroencephalographic activity observed during a) Experiment 1 (see [17]), and b) & c) Experiment 2 (see [18]). a) Results of the cluster-based permutation test from [17] showing a P300-like cluster differentiating significantly between the activity associated to human (plain line) and system (dotted line) error (red) detection. The topography of the difference wave and average amplitude of the cluster in each condition are displayed below. b) Time-courses and topographies (with difference waves; extracted from [18]) of the clusters differentiating between system error detection (red) and correct responses detection (blue) for both difficulties grouped (top panel) and for the easy condition only (bottom panel). c) Time-frequency plots (left), time course of theta activity (middle) and cluster topographies (right) of correct system response and error detection during

supervision of the conflict avoidance simulator at the beginning (left time-frequency plots, red lines and top topography) and the end (right time-frequency plots, blue lines and bottom topography) of the experiment.

A Window into the Mind? On Usefulness and Challenges of Neurophysiological Measurements in the Cockpit

A. Hamann[1] and N. Carstengerdes[1]

1 Institut für Flugführung, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) Braunschweig, Germany. anneke.hamann@dlr.de

Abstract

Neurophysiological measurements seem a powerful tool for investigating pilots' cognitive states. Such measurements need to be reliable and valid in order to gain interpretable and robust results. Recent research, however, suggests that this may not be as easy as it seems. In this paper, we give a short overview of commonly used neurophysiological measurements, and problems that need to be overcome to make these measurements the valid, reliable source of information we wish for.

Introduction: Aviation, Automation, Adaptation

The history of aviation is one of rapid increases in technological advances and ever higher levels of automation [8]. These developments go hand in hand with decreasing numbers of crew in the cockpit, following the assumption that higher levels of automation result in tasks that can be achieved by fewer human pilots [1]. Historically, the task allocation between human and machine has followed simple guidelines based on the capabilities of the actors [12], resulting in a static allocation. Later on, dynamic allocations arose in the form of adaptable automation which the human can changes based on their current needs, or adaptive automation in which the task allocation is changed by the machine, triggered by pre-defined situational, environmental or task-related factors [26, 31]. With the emergence of more sophisticated machine learning algorithms, the aviation industry and research alike now hope to develop concepts of adaptive automation that are tailored to the pilots and their current needs, and are thus more flexible and more supportive than traditional systems [8]. In order to do so, however, the system needs information about the pilots, their current state and needs. In their quest for detailed information about the human operator, researchers have turned to physiological measurements [2, 3]. Peripheral measurements like electrocardiography or electrodermal activity are often used to indicate stress or high arousal [2]. Neurophysiological measurements like electroencephalography (EEG) are used to gain more detailed insights into the humans' brain activity and cognitive concepts like mental workload or mental fatigue [2]. With the emergence of wearable functional near-infrared spectroscopy (fNIRS) devices, this interest in neurophysiological measurements has peaked again, indicated by a growing body of research in the last years [35]. EEG and fNIRS are promising measurements for the use in the cockpit because of their low intrusiveness, possibility to gain data continuously, and of course their objectivity. Thus, they have huge advantages compared to self-report or performance data.

Yet, there is one question that needs to be answered before sophisticated adaptive assistance systems should be fed with (neuro-) physiological data: How precise are the information we can gain from said measurements?

Mental Workload and the Validity of EEG Measures

To give an example and highlight the problems with valid neurophysiological measurements, in this paper we will focus on one concept: Mental workload. Mental workload is one of the most frequently researched concepts in aviation, despite controversies about its definition and conceptualisation [10, 27]. Generally, mental workload is defined as the part of one's cognitive resources needed to accomplish a task [25]. The more difficult the task, the more cognitive resources one needs to allocate to its accomplishment. If one has no spare resources left, task performance will decline and errors will occur. In aviation, it is generally believed that a medium level of mental workload is ideal [22]: Too low mental workload can border on boredom and can prompt disengagement from the task, causing loss of attention and out-of-the-loop phenomena when the pilot is suddenly asked to actively engage again. On the other hand, too high mental workload can lead to declined performance and errors that could lead to catastrophic outcomes. In sum, it would be ideal to keep the pilots' mental workload at a comfortable medium level to ensure optimal performance. Future adaptive assistance systems could help with this, if the pilots' current mental workload level could be assessed.

There is a large and growing body of research on mental workload assessment using EEG. Researchers usually focus on frequency band analyses for this, i.e. the decomposition of the EEG signal into its frequency bands. Changes in the composition of the signal are used to assess changes in cognitive states, e.g. between an easy and a difficult task. This way, data can be collected continuously and without the need to insert additional stimuli into the cockpit environment to elicit responses (as is done in analyses of event-related potentials). Indeed, the literature suggests a classical, "tell-tale" pattern of changing signal compositions for mental workload: Increasing mental workload is usually indicated by an increase in frontal theta activity, accompanied by a decrease in parietal alpha activity [9, 11, 16]. Yet, upon closer inspection, this "tell-tale" pattern cannot be found in every publication. While this is to be expected in research, the reasons for failing to detect the expected patterns should concern us.

For example, there are other concepts that elicit very similar cognitive activity, such as mental fatigue. Mental fatigue is described as a sense of weariness usually induced by long monotonous, yet demanding tasks [6, 15]. Not unlike mental workload, it is also based on the idea of resource depletion. The longer one needs to focus on a task, the more cognitive resources one spends until these resources are used up. If the individual does not or cannot take a break to replenish their cognitive resources, they will experience a subjective feeling of fatigue and a general unwillingness to spend further effort [15]. In EEG measurements, mental fatigue usually manifests in increasing frontal theta activity, accompanied by increasing parietal alpha activity [33, 34, 36], the same regions involved in mental workload. This is not just inconvenient when one wants to differentiate the concepts. There is research indicating that confounding mental workload and mental fatigue (when a strenuous task is performed over a long period of time) decreases the accuracy of mental workload classification [30]. Why? Because decreasing alpha activity with mental workload and increasing alpha activity with mental fatigue may cancel each other out and result in no detectable changes [30]. And mental fatigue is not the only concept that can interfere with a mental workload assessment. Frequent task switching, for example between navigation and communication tasks, also affects the alpha frequency band [29] and could be responsible for "missing" changes in parietal alpha activity [18]. On top of that, emotional responses, especially negative ones, impact frontal theta activity [16] and have the potential to interfere with the assessment as well.

These examples highlight why valid mental workload assessment may be achievable in the laboratory, but becomes rather difficult rather fast then turning towards realistic tasks. An algorithm trained to spot the typical, "tell-tale" mental workload pattern from laboratory studies can be confused easily when another factor comes into play in real-world settings. And yet, very few studies consider these problems and explicitly control for confounding factors or discuss limitations of their findings in this regard [4, 17]. It seems that an EEG-based system alone lacks validity. Maybe adding another data source could solve the problem?

Combining EEG and fNIRS

Combining fNIRS and EEG measurements has certain advantages. The higher temporal resolution of EEG and the higher spatial resolution of fNIRS measurements complement each other and can lead to more accurate results of when and where exactly changes happen, if analysed accordingly. And there is also the possibility to analyse convergent validity: If both measurements point towards the same result, it is more likely that the finding is true and not influenced by a confounding factor.

fNIRS measures the cortical activation based on changing oxygen concentrations in the cortical blood flow. Overly simplified, increasing brain activity increases the consumption of oxygen and thus the flow of oxygenated blood in the activated region, while the levels of deoxygenated blood in said region decrease simultaneously. Because of the different optical properties of oxygenated and deoxygenated blood, fNIRS can make this process visible and give an indication of the changes in cortical activation (for a more thorough explanation see [17]). The majority of studies indicate increasing (pre-) frontal cortical oxygenation (more oxygenated, less deoxygenated blood) with increasing mental workload [5, 13, 14]. There is one problem, however. The literature also points towards increasing (pre-) frontal cortical activation with increasing mental fatigue [7, 23, 24]. This could indicate that frontal cortical activation, as captured by fNIRS, is an indication of general demand, but rather unspecific. Unfortunately, most studies are performed on frontal and prefrontal regions, so there are only few published findings on parietal activation to this date [19]. In sum, while fNIRS is an interesting addition to EEG, the results gained from the measurement cannot differentiate mental workload from other types and sources of demand. This is highly problematic for the development of assistance systems based on neurophysiology. Assistance for a fatigued pilot could look very different from assistance for an overloaded pilot, and a system unable to tell the difference may adapt in the wrong direction. Moreover, if both effects indeed cancel each other out, an overloaded and fatigued pilot may not be offered any assistance because the system could no longer detect “unwanted” patterns of activation.

Our Systematic Approach to Achieving Validity

What can we as researchers do to overcome this problem of validity? The simple answer: Be aware of the limitations of the methods we apply, and investigate their limits systematically. In order to find valid neurophysiological measures of mental workload, we did a series of studies, systematically focusing on mental workload [18] and mental fatigue [19] while controlling for the influence of the respective other. In the following, we detail our approach. The described experiments were approved by the ethics commission of the German Psychological Society (DGPs) and conducted in accordance with the declaration of Helsinki.

Internal Validity: Inducing a Concept while Controlling for Confounds

If one wants to disentangle the neurophysiological correlates of mental workload from those of other concepts, one needs to make sure to induce only mental workload and control for other influences. In our recent research [18], we took great care to induce mental workload in four levels by means of increasing the difficulty of an adapted n-back task that was tailored to the flight context. We controlled for influences of mental fatigue by keeping the duration of the task to approx. 45 minutes, by randomizing the difficulty levels and by means of statistical analyses. Moreover, we used self-report and performance measures to ensure we actually did induce four distinct levels of mental workload.

Convergent Validity: Combining and Comparing Measurement Methods

We performed simultaneous fNIRS and EEG measurements to investigate the convergent validity between the measurements. In order to do so, we used compatible devices and chose a montage in which EEG electrodes and fNIRS optodes covered the same areas. There is software available for such purposes, like the fNIRS Optodes’ Location Decider fOLD [37], and we highly recommend making use of such tools as well as documenting the exact measurement locations in order to foster replicability.

Sensitivity and Specificity: Testing Measures on Different Concepts

In order to see which neurophysiological changes were unique to mental workload and could not be mistaken for mental fatigue, we also needed to research mental fatigue. Thus, we conducted a subsequent experiment in which we induced mental fatigue and controlled for confounding with mental workload [19]. By comparing the results of both studies, we would be able to see which neurophysiological measures were sensitive to mental workload, i.e. would vary with increasing mental workload, and specific to it, i.e. would not vary with increasing mental fatigue. In order to make the results comparable we designed the second experiment to be as similar as possible to the first. we used the same measurement equipment and montage, the same laboratory and flight simulator and the same cognitive task. We induced mental fatigue by increasing time on task (90 minutes) and kept mental workload constant by applying only one moderate difficulty level derived from our first study (details on our methods can be found in [18, 19]). We also chose the same measures, and data processing and analysis steps we had used in our first study.

Our Results: As Valid as Can Be?

As our current aim is to highlight the methods rather than the results, we will only briefly discuss our results here and refer the interested reader to the individual publications [18–20] as well as a comparative in-depth discussion in [17]. In short, we analysed and compared the fNIRS data (oxygenated blood HbO, deoxygenated blood HbR) and EEG frequency bands (frontal theta, and parietal alpha and beta), as well as two commonly used mental workload indices (Task Load Index TLI [32]; Engagement Index EI [28]) [20] between both experiments.

An overview of our findings is shown in Figure 3. We found that frontal measures (i.e. cortical oxygenation, theta band power and TLI) were sensitive to changes in mental workload, and that by combining EEG and fNIRS data, we could differentiate all four induced levels of mental workload, more than with the separate measurements. However, the frontal EEG measures proved sensitive to increasing mental fatigue as well, thus lacking specificity to mental workload. fNIRS may prove a viable option to differentiate the concepts, but our results were somewhat inconclusive (for a discussion see [17]). Parietal activity (i.e. alpha and beta band power and EI) lacked sensitivity to changing mental workload altogether.

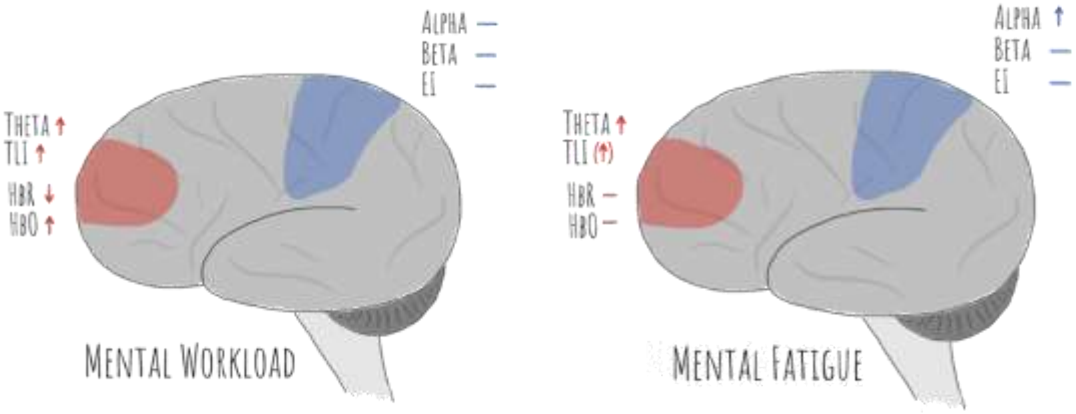


Figure 3. Schematic comparison of the results of our previous studies for mental workload and mental fatigue [18–20]. Red indicates frontal cortical areas, blue indicates parietal areas. Significant increases are marked ↑; significant decreases are marked ↓; no significant changes are marked – ; brackets indicate significant overall effect but no significant post-hoc tests. Figure taken from Hamann, 2023, p. 55 [17].

Conclusion

Even though we took great care to control for confounding, and could thus compare "pure" effects of mental workload and mental fatigue, it was quite difficult to tell the concepts apart. What is evident is that increasing frontal cortical activation indicates increasing demand, but cannot be used to explain its cause exactly, even under laboratory conditions. Unfortunately, in the cockpit it is unlikely to find pure mental workload. During long-haul flights, mental fatigue may well arise. Rostering and long shifts can lead to sleepiness on top of mental workload. And frequent task switching is inherent to piloting an aircraft. Thus, neurophysiological measurements may not be ideal for measuring mental workload "in the wild" and for differentiating cognitive concepts. We may be able to capture increasing or decreasing demand in a pilot, but not where exactly this demand originates from.

This may sound a little disappointing, but does not mean we should abandon neurophysiological measurements altogether. The fact alone that we can gain insights into a pilot's current cognitive demand and monitor changes should be impressive enough. And maybe this ability is already sufficient for our purposes. In an aircraft, there is an abundance of other sources of information readily available. Covariates like the duration of the mission, weather, aircraft system health indications and even the pilots' inputs into the system could bridge the gap between cognitive demand and its origin [21]. Increasing frontal activation in combination with system failure messages may be a good indication of mental workload, while the same pattern of frontal activation after multiple uneventful hours of flight is likely due to mental fatigue. Such a multimodal approach and combination of the overall state of both pilot (via neurophysiology) and aircraft (via covariates) could help to achieve the vision of an adaptive assistance system that is tailored to the current needs of the pilot without relying too much on the power of one measurement. In the end, the important part is to be aware of the capabilities and limitations of our methods instead of just assuming validity, and to apply them carefully where they are suitable and useful.

References

1. Billings, C. E. (1991). Toward a human-centered aircraft automation philosophy. *The International Journal of Aviation Psychology* **1**(4), 261–270. doi:10.1207/s15327108ijap0104_1.
2. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews* **44**, 58–75. doi:10.1016/j.neubiorev.2012.10.003.
3. Boumann, H., Hamann, A., Biella, M., Carstengerdes, N., & Sammito, S. (2023). Suitability of physiological, self-report and behavioral measures for assessing mental workload in pilots. In D. Harris & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics* (Vol. 14017, pp. 3–20, Lecture Notes in Computer Science). Cham: Springer Nature Switzerland.
4. Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience* **9**, 136. doi:10.3389/fnins.2015.00136.
5. Causse, M., Chua, Z. K., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports* **7**(1), 5222. doi:10.1038/s41598-017-05378-x.
6. Charbonnier, S., Roy, R. N., Bonnet, S., & Campagne, A. (2016). EEG index for control operators' mental fatigue monitoring using interactions between brain regions. *Expert Systems with Applications* **52**, 91–98. doi:10.1016/j.eswa.2016.01.013.
7. Chuang, C.-H., Cao, Z., King, J.-T., Wu, B.-S., Wang, Y.-K., & Lin, C.-T. (2018). Brain electrodynamic and hemodynamic signatures against fatigue during driving. *Frontiers in Neuroscience* **12**, 181. doi:10.3389/fnins.2018.00181.
8. CIEHF (2020). *The human dimension in tomorrow's aviation system: White Paper* .
9. Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., et al. (2019). Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions. *Sensors* **19**(6). doi:10.3390/s19061324.
10. Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work* **6**(2), 79–86. doi:10.1007/s10111-003-0136-9.
11. Dussault, C., Jouanin, J.-C., & Guezennec, C.-Y. (2004). EEG and ECG changes during selected flight sequences. *Aviation, Space, and Environmental Medicine* **75**(10), 889–897.
12. Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system* . Washington, DC, USA: National Research Council.
13. Gateau, T., Ayaz, H., & Dehais, F. (2018). In silico vs. over the clouds: On-the-fly mental state estimation of aircraft pilots, using a functional near infrared spectroscopy based passive-BCI. *Frontiers in Human Neuroscience* **12**, 187. doi:10.3389/fnhum.2018.00187.
14. Geissler, C. F., Domes, G., & Frings, C. (2020). Shedding light on the frontal hemodynamics of spatial working memory using functional near-infrared spectroscopy. *Neuropsychologia* **146**, 107570. doi:10.1016/j.neuropsychologia.2020.107570.
15. Grandjean, E. (1979). Fatigue in industry. *British Journal of Industrial Medicine* **36**(3), 175–186. doi:10.1136/oem.36.3.175.

16. Grissmann, S., Faller, J., Scharinger, C., Spüler, M., & Gerjets, P. (2017). Electroencephalography based analysis of working memory load and affective valence in an N-back task with emotional stimuli. *Frontiers in Human Neuroscience* **11**, 616. doi:10.3389/fnhum.2017.00616.
17. Hamann, A. (2023). *A systematic investigation of EEG and fNIRS measures for the assessment of mental workload in the cockpit*. Dissertation. Technische Universität Dresden, Dresden, Germany. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-869293.
18. Hamann, A., & Carstengerdes, N. (2022). Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights. *Scientific Reports* **12**(1), 6449. doi:10.1038/s41598-022-10044-y.
19. Hamann, A., & Carstengerdes, N. (2023). Assessing the development of mental fatigue during simulated flights with concurrent EEG-fNIRS measurement. *Scientific Reports* **13**(1), 4738. doi:10.1038/s41598-023-31264-w.
20. Hamann, A., & Carstengerdes, N. (2023). Don’t Think Twice, It’s All Right? – An examination of commonly used EEG indices and their sensitivity to mental workload. In D. Harris & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics* (Vol. 14017, pp. 65–78, Lecture Notes in Computer Science). Cham: Springer Nature Switzerland.
21. Hinss, M. F., Brock, A. M., & Roy, R. N. (2022). Cognitive effects of prolonged continuous human-machine interaction: The case for mental state-based adaptive interfaces. *Frontiers in Neuroergonomics* **3**. doi:10.3389/fnrgo.2022.935092.
22. Martins, A. P. G. (2016). A review of important cognitive concepts in aviation. *Aviation* **20**(2), 65–84. doi:10.3846/16487788.2016.1196559.
23. Nguyen, T., Ahn, S., Jang, H., Jun, S. C., & Kim, J. G. (2017). Utilization of a combined EEG/NIRS system to predict driver drowsiness. *Scientific Reports* **7**, 43933. doi:10.1038/srep43933.
24. Nogueira, M. G., Silvestrin, M., Barreto, C. S. F., Sato, J. R., Mesquita, R. C., Biazoli, C., et al. (2022). Differences in brain activity between fast and slow responses on psychomotor vigilance task: an fNIRS study. *Brain Imaging and Behavior*, 1–12. doi:10.1007/s11682-021-00611-8.
25. O’Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance* . New York: John Wiley & Sons.
26. Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems: Progress report* . Warminster, PA, USA.
27. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making* **2**(2), 140–160. doi:10.1518/155534308X284417.
28. Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology* **40**(1-2), 187–195. doi:10.1016/0301-0511(95)05116-3.
29. Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., & El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* **123**, 111–120. doi:10.1016/j.ijpsycho.2017.10.004.
30. Roy, R. N., Bonnet, S., Charbonnier, S., & Campagne, A. (2013). Mental fatigue and working memory load estimation: Interaction and implications for EEG-based passive BCI. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2013**, 6607–6610. doi:10.1109/EMBC.2013.6611070.
31. Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance*: *Theory and Applications* (1st ed., pp. 37–63, Human Factors in Transportation). London: CRC Press.
32. Smith, M. E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Human Factors* **43**(3), 366–380. doi:10.1518/001872001775898287.

33. Tran, Y., Craig, A., Craig, R., Chai, R., & Nguyen, H. (2020). The influence of mental fatigue on brain activity: Evidence from a systematic review with meta-analyses. *Psychophysiology* **57**(5), e13554. doi:10.1111/psyp.13554.
34. Trejo, L. J., Knuth, K., Prado, R., Rosipal, R., Kubitz, K., Kochavi, R. L., et al. (2007). EEG-based estimation of mental fatigue: Convergent evidence for a three-state model. In D. D. Schmorrow & L. M. Reeves (Eds.), *International Conference on Foundations of Augmented Cognition, Beijing, China* (Vol. 4565, pp. 201–211, Lecture Notes in Computer Science, Vol. 4565). Berlin: Springer. doi:10.1007/978-3-540-73216-7_23.
35. van Weelden, E., Alimardani, M., Wiltshire, T. J., & Louwerse, M. M. (2022). Aviation and neurophysiology: A systematic review. *Applied Ergonomics* **105**, 103838. doi:10.1016/j.apergo.2022.103838.
36. Wascher, E., Rasch, B., Sänger, J., Hoffmann, S., Schneider, D., Rinkenauer, G., et al. (2014). Frontal theta activity reflects distinct aspects of mental fatigue. *Biological Psychology* **96**, 57–65. doi:10.1016/j.biopsycho.2013.11.010.
37. Zimeo Morais, G. A., Balardin, J. B., & Sato, J. R. (2018). fNIRS Optodes' Location Decider (fOLD): a toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports* **8**(1), 3341. doi:10.1038/s41598-018-21716-z.

Unlocking Cost-Effective Insights: Leveraging Webcam Metrics for Cognitive Workload Assessments

Maykel van Miltenburg, Carmen van Klaren, Chihab Amghane

Royal Netherlands Aerospace Centre – NLR

Maykel.van.Miltenburg@nlr.nl, Carmen.van.Klaren@nlr.nl, Chihab.Amghane@nlr.nl

Introduction

Assessing cognitive workload through eye activity analysis has become integral across various disciplines. High-resolution eye-tracking systems, utilizing sophisticated infrared technology and precise cameras, offer detailed ocular data but demand substantial costs, calibration, and specialized setup [1]. Conversely, leveraging webcams combined with specific algorithms has emerged as a cost-effective and feasible alternative for assessing eye-related parameters [2,3]. This study delves into the potential correlation between eye-derived metrics from webcams, particularly blinks and percentage eye closure (PERCLOS), and corresponding metrics from a high-resolution eye-tracking system. Our objective centers on evaluating whether webcam metrics can reliably mirror metrics obtained from an advanced eye-tracking system.

PERCLOS, a well-established metric in drowsiness detection studies, represents the percentage of eye closure over a certain period. P60 is defined as the time during which the eyes are closed for a minimum of 60%, and this duration tends to increase with the onset of fatigue [4,5]. This effect can occur as a result of prolonged periods of low arousal, but it is not expected to differentiate task complexity during short periods of active engagement [6]. Additionally, our study delves into the measurement of blinks, a fundamental aspect of eye behavior. Blinks are integrant in understanding cognitive workload with numerous studies revealing an association between increased subjective or task-induced workload and a decrease in blink frequency [7–9]. However, the aim is comparing eye-derived metrics obtained via webcam and eye-tracker, rather than its sensitivity to task complexity.

To contribute to the exploration of cost-effective and versatile methodologies for workload assessment, our experimental design integrates the Multi-Attribute Task Battery II (MATB-II) and the n-back task, executed concurrently under varied complexity levels. While the n-back task offers simplicity and predictive power, criticisms regarding its ecological validity have arisen [10–13]. The MATB-II provides a multifaceted multitasking environment tailored for operator performance assessment, mirroring real-world complexities [14]. Its utilization allows controlled variations, making it instrumental in workload studies [15–17].

By comparing webcam-derived metrics and outputs from high-resolution eye-tracking systems within multitasking scenarios, this study aims to contribute to the exploration of cost-effective and versatile methodologies for workload assessment.

Research questions and hypotheses

Our study investigates the following two research questions with corresponding hypotheses:

1. How does the manipulation of task complexity, across varying conditions from easy to hard, influence performance metrics such as reaction time in MATB-II, hit rate in the n-back task, and physiological markers including total blink count and PERCLOS?

    - Hypothesis 1a: As task complexity increases, an increase in reaction time is anticipated.
    - Hypothesis 1b: As task complexity increases, a decrease in hit rates is expected.
    - Hypothesis 1c: As task complexity increases, a reduction in total blink count is anticipated.
    - Hypothesis 1d: As task complexity varies, no effect in PERCLOS is expected.

2. To what extent is there a relationship between eye-derived metrics, particularly total blink count from webcam recordings, and those derived from the high-resolution eye-tracking system?

    - Hypothesis 2a: A significant correlation between the total blink count derived from webcam recordings and those obtained from the high-resolution eye-tracking system is expected, indicating the feasibility of webcam-based metrics as a cost-effective alternative for eye-tracking assessments in workload studies.

    - Hypothesis 2b: A significant correlation between PERCLOS from webcam and eye-tracking system is unlikely due to the expectation that task complexity will not affect PERCLOS. However, it is being investigated as an exploratory outcome measure.

Method

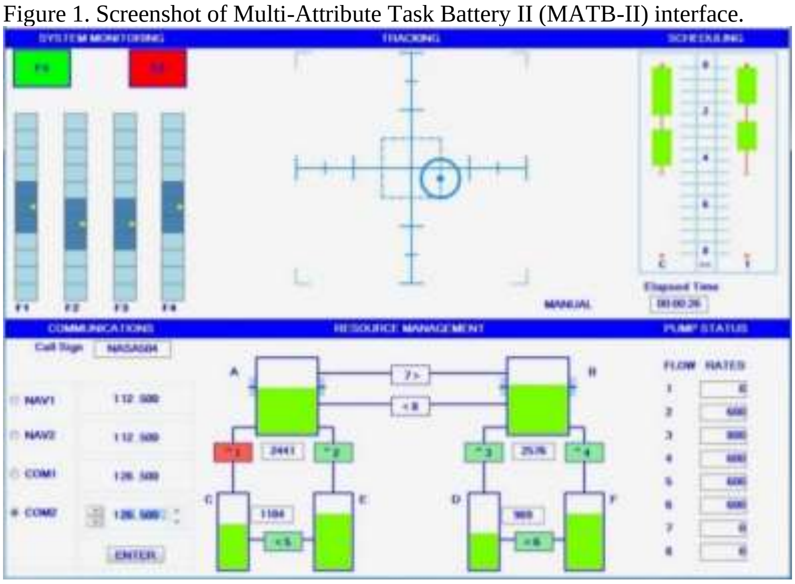
Participants

A total of 12 healthy participants (8 male, 4 female; 9 right-handed, 1 ambidextrous, 2 left-handed) aged between 23 and 55 (*M* = 30.17, *SD* = 8.75) participated in this study. Informed consent was obtained from each participant, and all received a compensation for their participation. The study complied with the tenets of the Declaration of Helsinki.

Tasks

Figure 1 shows an example display of the multitask environment (MATB-II) with the four subtasks outlined. For this study, the MATB-II test included three of the four subtasks, namely a psychomotor tracking subtask (in which participants had to keep the target inside a central square with a joystick; top center), a system monitoring (SYSMON) subtask (in which participants were tasked with clicking on the green light when it went out, clicking on the red light when it appeared, and clicking on the scales when the central indicators deviated significantly from center; top left) and a resource management (RESMAN) subtask (where participants had to manage fuel levels in tanks by (de)activating pumps from other tanks; bottom center) [18]. The original MATB-II test also included a communication subtask (bottom left) which was replaced by an auditory n-back task.

Figure 1. Screenshot of Multi-Attribute Task Battery II (MATB-II) interface.



The auditory n-back task is a widely used task to measure and manipulate cognitive workload [19,20]. For each presented letter, participants have to decide whether it was a repetition of the letter, "n" letters before [21]. The auditory n-back task contained a random sequence of pre-recorded letters presented auditory in intervals of 2.5 seconds. For each trial of 10 letters, there was a hit. Participants were instructed to respond verbally by saying "Yes" after a hit.

The tracking task was constant for all conditions. Settings for the other MATB-II subtasks and n-back task varied across the five difficulty levels:

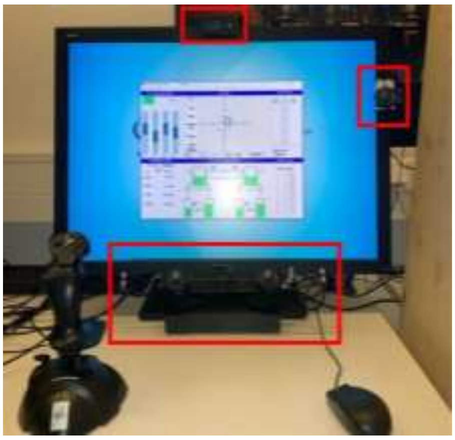
Extremely easy: 10 SYSMON events (intervals of 30s), 6 failures in RESMAN, 0-back task

3. Easy: 15 SYSMON events (intervals of 15s), 6 failures in RESMAN, 1-back task
4. Medium: 20 SYSMON events (intervals of 12s), 6 failures in RESMAN, 1-back task
5. Hard: 30 SYSMON events (intervals of 11s), 11 failures in RESMAN, 1-back task
6. Extremely hard: 30 SYMON events (intervals of 10s), 11 failures in RESMAN, 2-back task

Equipment

All tests were administered on a standard Windows 10 laptop with a 27-inch LED-backlit LCD monitor. Participants used an optical mouse and a joystick to provide responses during the (sub)tasks. Video data were captured during the task with a webcam (Logitech C505e) mounted in front of the participant (see Figure 2) at a sampling rate of 30 Hz. The resolution of the webcam footage is 720p (HD). Simultaneously, a three camera based eye-tracking system (Smart Eye Pro 10.2) was used to capture and record participants' eye activity at a sampling frequency of 60 Hz.

Figure 2. Experimental setup of the eye-tracker (top right and bottom center) and webcam (top center).



Design and procedure

In this within-participant design experiment, six conditions were implemented. The experiment started with a short briefing about the study and participants were asked to complete a questionnaire gathering information about age, gender, handedness, and prior experience with the MATB-II test. After which, received a brief presentation delivering instructions for both the MATB-II tasks and the n-back task. Subsequently, participants engaged in a practice session for the MATB-II task, lasting approximately 3 minutes. During this phase, participants had the opportunity to familiarize themselves with the MATB-II test, ask questions, and the experimenter ensured participant's comprehension of task functions. Following the practice session, a brief 5-minute baseline test was administered, during which eye-tracking data and webcam footage were recorded. Subsequently, five conditions, varying in task load and complexity, were executed. The recording of the eye-tracking, along with webcam footage, started just before initiating each new condition. Participants received a break of approximately 2 minutes after completing each condition before proceeding to the next. The order of the conditions was randomly determined. After each condition, the participants completed a short questionnaire regarding subjective workload (based on the NASA Task Load Index [22]). The experiment was conducted in a room with controlled lighting, and the experimenter remained outside the central field of vision of the participant. The total experiment took 60 minutes.

Behavioral performance measures

Within our study, only the performances on the system monitoring (SYSMON) subtask and n-back task were analyzed as a performance outcome measure. Reaction times (RTs) and hit rate were selected as primary outcome measures for, respectively, MATB-II SYSMON subtask and n-back task. For each condition, RTs (in ms) and hit rate (in %) were averaged. Hit rates were calculated by dividing the number of correct responses to target trials by the total number of target trials. Miss trials, defined as trials in which no response was given before the next stimulus appeared (n-back) or within 15 seconds (SYSMON), and commission errors (RTs <100 ms) were excluded (SYSMON: 0.1% data loss).

Subjective cognitive workload measure

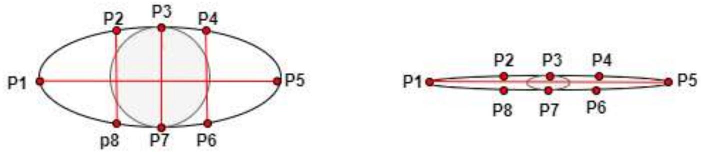
The NASA Task Load Index (NASA-TLX) consists of six bipolar subscales: mental workload, physical workload, temporal workload (time pressure), performance, effort, and frustration. The scale ranges from 1 to 100; higher scores on the NASA-TLX indicate a higher degree of subjective workload. The overall workload index was measured by averaging the results on these subscales (without weighting factors). Responses were given through a slider to give a brief estimation of workload. This was incorporated in the MATB-II. Corwin et al. **[23]** found in two studies that the NASA-TLX demonstrated high validity and reliability.

Eye-derived measures

In this section, the approach used to extract data from the webcam footage and eye-tracking will be explained. First, the methodology behind detecting regions of interest on the face will be explained. This is followed by a description of the used metrics and the rationale behind them. Finally, the implementations of these metrics will be explained.

In our study, we used the landmark method to determine eye coordinates necessary for calculating the eye aspect ratio in the webcam footage. The Landmark detection approach served to determine the location of facial components such as the mouth, cheeks and eyes [3]. The FaceMesh solution from the MediaPipe library [24] was used to determine these facial landmarks. Subsequently, the state of the eyes (whether they were open or closed, and to what extent) could be calculated in the following step. The region of interest were the eyes and a set of coordinates *P={P1,P2,P3,P4,P5,P6,P7,P8}* describing the points located on each eye (see Figure 3).

Figure 3. Overview of the landmark positions extracted with FaceMesh on the contour of the eye used to determine state of the EAR. The left image portrays the eye in an open state and the image on the right in a closed state.



The Eye aspect Ratio (EAR) is a measure of the ratio between the width and the height of the eye [2]. There are multiple approaches to calculating the EAR based on the two or three dimensional landmark locations extracted from the face of the subject. The three dimensional landmark coordinates are useful when the subject is expected to make a lot of head movements while being monitored [1]. In our study, the decision was made to use the conventional 2D landmark location approach, since the subjects were not expected to frequently move the head and the camera was mounted frontally. Furthermore, instead of using 6 points on the eye, the decision was made to use 8 points in order to reduce measurement error (Eq 1). The EAR is calculated for each eye separately and then averaged to obtain the EAR for both eyes.

$$EAR = \frac{\|P_2 - P_8\| + \|P_3 - P_7\| + \|P_4 - P_6\|}{3\|P_1 - P_5\|} \text{ (Eq 1)}$$

PERCLOS is defined as the percentage of eye closure over a certain period of time [25]. The PERCLOS metric is used frequently in drowsiness detection studies where it is defined as P60, P70 or P80, indicating the percentage of time the eyes were at least 60, 70 or 80 percent closed, respectively [25,26]. For the webcam footage, the EAR was used to determine the PERCLOS, since an open eye corresponds to a higher EAR and a closed eye corresponds to a lower EAR. For the eye-tracker data, the eyelid opening (as expressed in meters) instead of the EAR is used with the same approach. In order to express the EAR (or eyelid opening in case of the eye-tracker) as a percentage of eye closure, both minimum and maximum EAR values were determined for each participant.

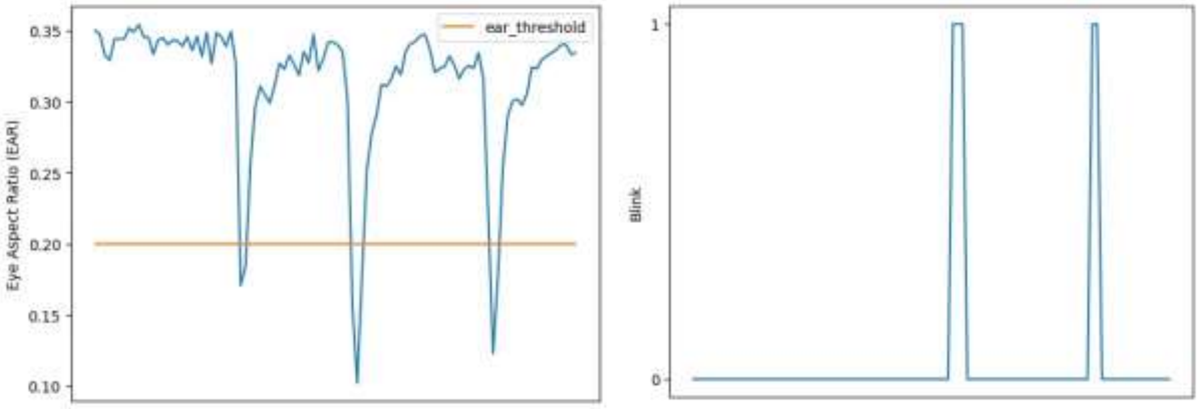
The peaks of the EAR over time for each participant were extracted and averaged over the hundred (for the eye-tracker 200 values due its sampling rate) lowest and highest EAR values resulting in EAR_min and EAR_max. The EAR_min and EAR_max values were used to scale the data where an EAR value of EAR_max would be scaled to 1 and a value of EAR_min to 0 (Eq 2). The resulting value is used as an approximating of the percentage of the eye closure, i.e. Eye Closure Percentage, in order to determine the PERCLOS. Thus, in this experiment, the PERCLOS was calculated by determining the total time the Eye Closure Percentage was equal to or higher than a certain percentage (60, 70 or 80) divided by the total time (Eq 3). The PERCLOS was calculated for each eye separately and average in order to obtain the PERCLOS for both eyes combined. Finally, the eye closure percentage values of the timesteps where a blink occurred were omitted during calculations for the PERCLOS.

$$Eye\ Closure\ Percentage = (1 - \frac{EAR - EAr_{min}}{EAr_{max} - EAR_{min}}) * 100 \text{ (Eq 2)}$$

$$P_X = \frac{total\ time\ \ eye\ closure\ percentage \geq X}{total\ time} * 100 \text{ where X} = \{60{,}70{,}80\} \text{ (Eq 3)}$$

For the eye-tracker, blink is identified using its dedicated acquisition software (Smart Eye Pro 10.2). For the webcam footage, blink is defined as an EAR being lower than 0.2 for at least 100 ms and at most 400 ms [2]. The threshold value for the EAR as well as the duration for a blink were based on earlier studies [2,27]. Since the EAR values were calculated both for the eyes separately as well as combined, the decision was made to determine blinks both for the eyes combined and separately. In Figure 4, an example of detected blinks within the webcam footage during a window of three seconds is shown. The left image shows EAR and the threshold used to determine whether the eye was closed or not. If the EAR was below the threshold for a consecutive period of time (100 ms – 400 ms) then it is counted as a blink. The right image shows that only the last two peaks of the upper image are actually counted as blinks.

Figure 4. Example of detected blinks during a window of three seconds with a threshold value of 0.2 (ear_threshold).



Statistical analysis

Analyses were performed with linear mixed-effects model analyses using LME in R (R Core Team, 2017) with participant as random factor. The data was checked for outliers and normality. Following the fitting of the linear mixed effects models, residuals were extracted for each eye-derived metric. Pearson's correlation coefficient was then computed between these residuals, considering participant variability. A two-tailed t-test was performed to indicate whether the correlation was statistically significant.

Results

Firstly, the effect of condition on the performance and eye-derived metrics is presented. Secondly, the correlation between the eye-derived metrics from the webcam and the eye-tracker is investigated.

Performance and subjective workload

The expectation is that the performance on the tasks decreases as task complexity rises. To determine if the different conditions affect performances on both the SYSMON subtask and n-back task, two behavioral outcome measures were analyzed, respectively the reaction time and hit rate. The mean declarative score on reaction time and hit rate is presented in Table 1.

Table 1. Overview of descriptive statistics (*M, SD, CL*) of condition on reaction time and hit rate (with .95 as confidence level).

	MATB-II				n-back			
Condition	*M*	*SD*	*Lower CL*	*Upper CL*	*M*	*SD*	*Lower CL*	*Upper CL*
	Reaction time (in ms)				Hit rate (in %)			
Extremely easy	1571	431	1230	1912	94.5	9.4	87.0	101.9
Easy	1596	574	1255	1937	81.7	15.5	74.3	89.2
Medium	1499	225	1158	1839	89.2	13.7	81.8	96.7
Hard	1922	710	1581	2263	87.4	10.5	79.9	94.8
Extremely hard	2144	806	1803	2485	81.6	14.0	74.2	89.1

Condition had a significant effect on reaction time and hit rate which details $(F, p, \eta_p{}^2)$ are presented in Table 2. Post-hoc analysis indicates that reaction time during the most mentally challenging condition (i.e. extremely hard) increased significantly with 573 ms when compared to the least mentally challenging condition (i.e. extremely easy). Hit rate decreased significantly with 13% when compared. There was no effect of condition order on performance, which demonstrates that participants became neither better nor worse in their performance.

Table 2. Overview of effects $(F, p, \eta_p{}^2)$ of condition on reaction time and hit rate.

Performance measurement	*F*	*p*	$\eta_p{}^2$
Reaction time	3.433	.016	.238
Hit rate	2.720	.042	.198

The results of NASA Task Load Index (NASA-TLX) showed a significant increase in perceived workload with increasing task demands. This underscores the effectiveness of our approach in inducing varying levels of cognitive workload across experimental conditions, though details are not presented here.

Eye-derived metrics

The expectation is that the total blinks decreases as task complexity increases. No effect on P60 is expected. Descriptive statistics for the eye-derived metrics, total blinks and P60 from both the webcam and eye-tracker are presented in Table 3. Condition order had an effect on total blinks from the eye-tracker only and was controlled for.

Table 3. Overview of descriptive statistics (*M, SD, CL*) of condition on total blinks and P60 (with .95 as confidence level).

	Webcam				Eye-tracker			
Condition	*M*	*SD*	*Lower CL*	*Upper CL*	*M*	*SD*	*Lower CL*	*Upper CL*
	Total blinks				Total blinks			
Extremely easy	41.0	33.6	19.3	62.7	27.4	17.4	9.7	45.1
Easy	39.5	32.2	18.0	61.0	29.6	29.0	12.0	47.2
Medium	38.0	25.9	16.3	59.7	26.5	16.8	8.8	44.2
Hard	35.5	22.0	13.8	57.2	21.8	15.3	4.0	39.5
Extremely hard	58.8	46.6	37.2	80.3	40.8	35.4	23.1	58.4
	P60				P60			
Extremely easy	5.7	9.4	1.5	9.8	10.0	12.2	5.1	4.9
Easy	3.7	5.3	.0	7.8	7.7	8.6	2.8	2.6
Medium	5.6	7.7	1.4	9.7	6.7	8.9	1.9	1.6
Hard	5.0	5.8	.8	9.1	7.2	6.4	2.3	2.0
Extremely hard	3.9	4.3	.0	8.0	7.6	5.1	2.7	2.4

Condition had a significant effect on total blinks. No significant effect of condition on P60 was measured. The details $(F, p, \eta_p{}^2)$ are presented in Table 4. Post-hoc analysis indicates that the total number of blinks from the webcam and eye-tracker during the most mentally challenging condition (i.e. extremely hard) increased significantly with, respectively 19 and 14 blinks when compared to the least mentally challenging condition (i.e. extremely easy).

Table 4. Overview of effects $(F, p, \eta_p{}^2)$ of condition on each eye-derived metric.

Eye-derived metric		F	p	$\eta_p{}^2$
Total blinks	Webcam	3.867	.009	.275
	Eye-tracker	4.164	.006	.292
P60	Webcam	1.106	.366	.091
	Eye-tracker	.610	.658	.053

Pearson’s correlation analysis revealed a strong positive correlation between the residuals of the total blinks from the webcam and total blinks from the eye-tracker across participants (*Pearson's r* = .787). This indicates a consistent relationship between the two metrics even after controlling for individual participant differences. A two-tailed t-test indicated that this correlation was statistically significant, *t*(10) = 10.111, *p* < .001. The Pearson’s correlation coefficient between the residuals of P60 from the webcam and P60 from the eye-tracker was found to be .226 which did not reach statistical significance, *t*(11) = 1.413, *p* = .185. The results of P70 and P80 showed similar results, though their details are not presented.

Discussion

The current study delved into the potential of utilizing webcam technology as a cost-effective alternative for assessing cognitive workload by comparing eye-derived metrics with those obtained from a high-resolution eye-tracking system. The results offer valuable insights into the practicality and limitations of this approach, shaping discussions around its applicability, benefits, and avenues for future research.

The observed effects of varying task complexity on performance metrics align with existing literature on cognitive workload. As task complexity increased, participants exhibited prolonged reaction times and decreased hit rates, underscoring the validity of the experimental design in inducing different levels of cognitive workload. These findings are consistent with the expected impact of cognitive load on task performance and contribute to the understanding of how individuals respond to multitasking scenarios.

The study investigated eye-derived metrics, focusing on total blinks and PERCLOS, utilizing a webcam-based approach. While total blinks exhibited a significant increase under higher cognitive workload conditions, P60 did not show significant effects of task complexity. As expected, the correlation analysis between webcam and eye-tracker-derived total blinks revealed a strong positive relationship, suggesting that webcams may serve as a reliable and cost-effective tool for capturing specific eye-related parameters. Contrary to the hypothesis, results showed a positive relation between total blinks and task load and complexity. A possible explanation could be that the introduction of the auditory n-back task might have reduced visual concentration. Findings from a study employing exclusively auditory oddball tasks showed an increase in blink frequency as task load increases, in contrast to the decrease observed in visual tasks [28]. Another study combining a driving task with an auditory task resulted in similar outcomes [29].

Limitations and future research

The study has several limitations that warrant consideration, particularly regarding PERCLOS. The choice of webcam framerate, potential impacts of eyewear on landmark detection, and the temporal aspects of blink detection introduce variability and potential sources of measurement error.

The PERCLOS is defined as the percentage eye lid closure over time. In our work an extension of the EAR was used to determine the PERCLOS. In order to determine a minimum and maximum EAR, the peaks were averaged over a certain time window. The minimum and maximum EAR would correspond to eyes closed and eyes opened. While an ideal scenario would involve a calibration session, determining EAR at rest, eyes open, and eyes closed, the absence of such a calibration step introduces potential advantages aligned with the objective of our study. The exclusion of individualized calibration sessions, while posing challenges, may streamline the applicability of our methodology in real-world settings.

The webcam used in our study captured footage at a conventional framerate of 30, which may limit the precision of EAR determination and blink detection. Employing a higher framerate could potentially enhance the accuracy of our measurements. Additionally, the inclusion of participants wearing glasses introduced a factor that might impact the landmark detection algorithm due to reflections.

Another limitation arises from the criteria used to define blinks. In our methodology, an EAR lower than the threshold for at least 100 ms and at most 400 ms was considered a blink. However, the time between blinks was not explicitly specified. This introduces the theoretical possibility that two blinks separated by a single timeframe could be erroneously counted as one blink. The EAR value in the intermediate timeframe may be susceptible to measurement error, adding a layer of complexity to the interpretation of blink data. These considerations contribute to the nuanced nature of our blink-related findings.

Future research could consider incorporating calibration steps to ascertain the range of EAR values, explore higher framerates to potentially enhance precision of the PERCLOS calculations, and refine methodologies to mitigate the influence of eyewear on landmark detection. Addressing these aspects will contribute to the robustness and validity of webcam-based eye-tracking methodologies.

Conclusion

Despite the limitations, this study underscores the practicality of utilizing webcams in assessing cognitive workload, especially in scenarios where high-resolution eye-tracking systems pose logistical challenges. The positive correlation between webcam and eye-tracker-derived total blinks highlights the feasibility of webcams as a cost-effective tool for capturing specific eye-related parameters. The integration of low-cost tools like webcams into research and applied settings may pave the way for more accessible and versatile methodologies in workload studies, revolutionizing data collection and analysis. However, cautious interpretation is advised due to the identified limitations, and future research should address these challenges to enhance the reliability and validity of webcam-based methodologies.

References

1. Kraft, D., Hartmann, F., & Bieber, G. (2022). Camera-based Blink Detection using 3D-Landmarks. *Proceedings of the 7th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, 1–7. https://doi.org/10.1145/3558884.3558890

2. Soukupova, T., & Cech, J. (2016, February). Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia* (p. 2).

3. Wu, Y., & Ji, Q. (2019). Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision*, **127**(2), 115–142. https://doi.org/10.1007/s11263-018-1097-z

4. Sommer, D., & Golz, M. (2010). Evaluation of PERCLOS based Current Fatigue Monitoring Technologies. Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. *IEEE Engineering in Medicine and Biology Society*. Conference, 2010, 4456–4459. https://doi.org/10.1109/IEMBS.2010.5625960

5. Zhang, J., Chen, Z., Liu, W., Ding, P., & Wu, Q. (2021). A Field Study of Work Type Influence on Air Traffic Controllers' Fatigue Based on Data-Driven PERCLOS Detection. *International Journal of Environmental Research and Public Health*, **18**(22), Article 22. https://doi.org/10.3390/ijerph182211937

6. Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2020). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training. *Human Factors*, **62**(8), 1365–1386. https://doi.org/10.1177/0018720819874544

7. Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surgical Endoscopy*, **26**(10), 2746–2750. https://doi.org/10.1007/s00464-012-2268-6

8. Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics,* **41**(5), 656–669. https://doi.org/10.1080/001401398186829

9. Charles, R., & Nixon, J. (2017, April 27). Blink counts can differentiate between task type and load. https://dspace.lib.cranfield.ac.uk/handle/1826/12547

10. Adams, J. A. (1987). Criticisms of Vigilance Research: A Discussion. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **29**(6), 737–740. https://doi.org/10.1177/001872088702900612

11. Al-Shargie, F. (2019). Vigilance Enhancement Using Traditional Methods: A Review [Preprint]. engrXiv. https://doi.org/10.31224/osf.io/mypt7

12. Kibler, A. W. (1965). The Relevance of Vigilance Research to Aerospace Monitoring Tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **7**(2), 93–99. https://doi.org/10.1177/001872086500700202

13. Koelega, H. S. (1993). Stimulant drugs and vigilance performance: A review. *Psychopharmacology*, **111**(1), 1–16. https://doi.org/10.1007/BF02257400

14. Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide (No. L-20031).

15. Caldwell, J. A., Caldwell, J. L., Brown, D. L., & Smith, J. K. (2004). The Effects of 37 Hours of Continuous Wakefulness On the Physiological Arousal, Cognitive Performance, Self-Reported Mood, and Simulator Flight Performance of F-117A Pilots. *Military Psychology*, **16**(3), 163–181. https://doi.org/10.1207/s15327876mp1603_2

16. Carlozzi, N., Horner, M., Kose, S., Yamanaka, K., Mishory, A., Mu, Q., Nahas, Z., Wells, S., & George, M. (2009). Personality and Reaction Time after Sleep Deprivation. *Current Psychology* (New Brunswick, N.J.), **29**, 24–33. https://doi.org/10.1007/s12144-009-9068-8

17. Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, **38**(2), 311-322.

18. Comstock, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research (NAS 1.15:104174). https://ntrs.nasa.gov/citations/19920007912

19. Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., & Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering*, **9**(4), 045008. https://doi.org/10.1088/1741-2560/9/4/045008

20. Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, **55**(4), 352–358. https://doi.org/10.1037/h0043688

21. He, D., Donmez, B., Liu, C. C., & Plataniotis, K. N. (2019). High Cognitive Load Assessment in Drivers Through Wireless Electroencephalography and the Validation of a Modified N-Back Task. *IEEE Transactions on Human-Machine Systems*, **49**(4), 362–371. https://doi.org/10.1109/THMS.2019.2917194

22. Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **50**(9), 904–908. https://doi.org/10.1177/154193120605000909

23. Corwin, W. H., & Biferno, M. H. (1989). Assessment of Crew Workload Measurement Methods, Techniques and Procedures. DOUGLAS AIRCRAFT CO LONG BEACH CA, 1, 237.

24. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. https://doi.org/10.48550/ARXIV.1906.08172

25. United States. Federal Motor Carrier Safety Administration. Technology Division. (1998). PERCLOS: A Valid Psychophysiological Measure of Alertness As Assessed by Psychomotor Vigilance. https://doi.org/10.21949/1502740

26. Abe, T. (2023). PERCLOS-based technologies for detecting drowsiness: Current evidence and future directions. *SLEEP Advances*, **4**(1), zpad006. https://doi.org/10.1093/sleepadvances/zpad006

27. Dewi, C., Chen, R.-C., Jiang, X., & Yu, H. (2022). Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks. *PeerJ Computer Science*, **8**, e943. https://doi.org/10.7717/peerj-cs.943

28. Magliacano, A., Fiorenza, S., Estraneo, A., & Trojano, L. (2020). Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neuroscience Letters*, **736**, 135293. https://doi.org/10.1016/j.neulet.2020.135293

29. Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task Performance and Eye Activity: Predicting Behavior Relating to Cognitive Workload. **78**(5).

Symposium: Advances in using AI to assess animal behaviour and welfare

Transdisciplinary Initiatives to Collaborate on the Responsible Use of AI for Animal Welfare

Mona F. Giersberg & T. Bas Rodenburg

Animals in Science and Society, Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University

Abstract

Artificial Intelligence-based methods have the potential to expand human capacities to measure and monitor animal welfare indicators more accurately and continuously. However, developing or adapting these methods requires the expertise of relatively unrelated academic fields, such as veterinary medicine and computing sciences. Therefore, collaborative structures are needed that offer perspectives beyond individual research projects. One example of such initiatives is the AI & Animal Welfare Lab at Utrecht University, which we present in this paper.

Introduction

Animals are part of our lives in different contexts. We keep cats as companions, cows for milk production and service dogs to assist us with daily tasks. In our society, we increasingly prioritise the welfare of these animals. At the same time, there has been a paradigm shift within animal welfare science. Current concepts of animal welfare go beyond a mere prevention of harm and focus on enabling positive states and experiences [1]. Implementing this new understanding into practice also calls for novel techniques and tools to measure and monitor animal welfare. Observing an animal's behaviour can provide useful information about its welfare. However, indicators for positive states and experiences may be only visible in the long term or are too subtle to be detected by the traditional methods of ethology. To detect and respond to (un-)desired behaviour in time, we need to expand human capacities and monitor the animals under our care more accurately and continuously. This can be achieved by using Artificial Intelligence (AI)-based methods, such as computer vision.

Developing these methods for or adapting them to an animal welfare context requires the expertise of relatively unrelated academic fields, such as veterinary medicine and computing sciences. In addition, to ensure that AI-based methods are socially robust and contribute to the challenges stakeholders face in practice, it is essential to involve non-academic partners from industry, public organisations and governments [2]. Thus, what is needed is a transdisciplinary approach. By this, a project team can move beyond discipline-specific perspectives to create AI applications for animal welfare monitoring which are both scientifically sound and problem-solving oriented. However, this approach comes with its own challenges. In general, it takes time and efforts to establish mutual trust and collaboration among a diverse group of research partners. Furthermore, in a multidisciplinary team the determination of the underlying meaning of core-concepts, such as animal welfare, can be difficult. The same concept may have different connotations for scientists from different disciplines [3] and may be viewed from a descriptive, normative or hybrid perspective [4]. If the team is not aware of these aspects and if they are not discussed openly, it can be challenging to agree on the interpretation of core-concepts, which often has direct consequences for further research and development activities [4].

For real collaboration on AI-based methods for animal welfare monitoring, structures are needed that offer perspectives beyond individual research projects. In addition to content-related work, these structures should be aimed at community building and include regular moments of reflection on team processes. This will allow for building trust and foster long-term collaborations among research partners. Only by linking discipline specific expertise with process-related structures, will it be possible to make use of technological advances to ultimately improve the lives of animals in practice. One example of such collaborative initiatives is the AI & Animal Welfare Lab at Utrecht University, which we present in this paper.

Aim of the AI & Animal Welfare Lab

The AI & Animal Welfare Lab is one of the currently 15 theme-based AI Labs of Utrecht University. The overall aim of the Labs is to tackle societal challenges by generating state-of-the-art knowledge on AI and data science and bridging the gaps between research, education and professional practice. Within the AI & Animal Welfare Lab our ultimate goal is to improve animals' lives in practice by using the full potential of AI in a responsible way. What is unique about the Lab is that the animal is at the centre. Our aim is to understand the animal's perspective as best as possible and to take it seriously. This means that we strive to implement AI with added value for the animal itself and not with the motivation to use animals more efficiently for human purposes (e.g. for food production, in sports or as service animals). From this perspective our key researchers from veterinary medicine, biology and computing and information sciences run different projects together, in which they work in close collaboration with companies, government agencies and NGOs. To realise our ambition to serve not only as a research- but also as an education hub, we involve master and PhD students in each project and position them towards future careers in academia or industry. The AI-based solutions we develop and apply during these collaborative projects are related to different focus areas which are described below.

Thematic focus

The different contexts in which animals are kept often require different expertise and methods to assess and monitor animal welfare. At the same time, methods developed for one specific context may be transferable or adaptable to other situations or animal species. Within the AI & Animal Welfare Lab, our work is structured in four thematic focus areas (Figure 1). The lines between these focus areas are fluid, which means that most key researchers work on several themes.

Our first thematic focus area is positive animal welfare. As mentioned above, it has become increasingly important to not only prevent welfare problems, but also to promote positive states in the animals we keep. However, detecting, measuring or quantifying rewarding behaviours and positive emotions can be challenging. Besides other behaviours, vocalisation is a useful key indicator. Both the type and the characteristics (e.g. amplitude, frequency) of vocalisations can provide valuable information on how an animal is feeling. In the Lab, we use for instance methods based on clustering and active learning, behaviour classification and feature learning to exploit data on positive welfare indicators.

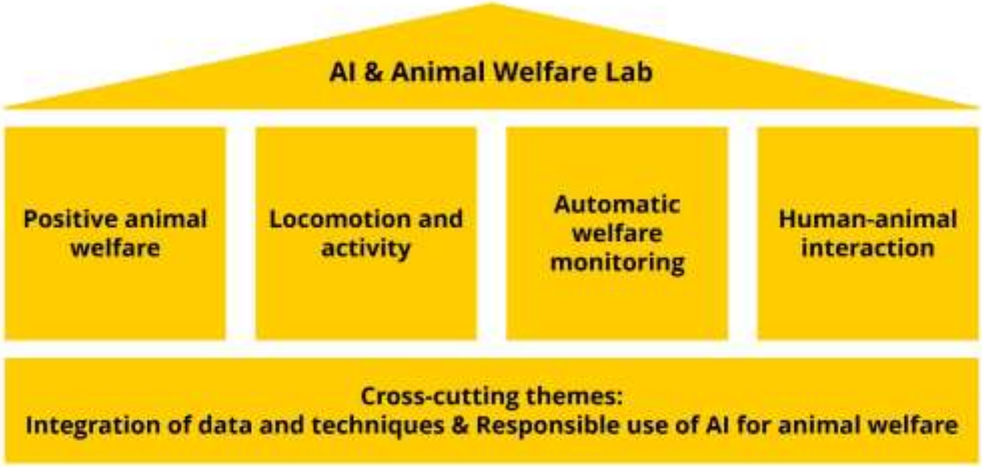


Figure 1. The four thematic focus areas and two cross-cutting themes of the AI & Animal Welfare Lab.

The second focus area is locomotion and activity. At the moment, our team focuses on equine lameness. Lameness affects a huge number of horses and often involves a loss of quality of life. Early and accurate diagnosis is crucial but traditional visual examination is subjective and depends on the background of the assessor. Within the AI & Animal Welfare Lab, researchers develop the method of quantitative gait analyses. For this, various techniques are explored, such as optical motion capture, inertial measurement unit sensors, force-/pressure plates and surface electromyography, which are combined with advanced data analysis and processing methods like machine learning. We plan to adapt and transfer these tools to other quadrupedal animals like cows, but also to bipedal animals like broiler chickens, in which gait problems and low levels of activity can be an issue.

Third, we focus on integrated automatic welfare monitoring in various settings. In contrast to the focus areas mentioned above, we do not study and develop tools for individual welfare indicators within this theme. Instead, we aim to reliably measure and monitor a set of indicators over time in a certain setting (e.g. on a farm, in a shelter). Our efforts focus on the combination and integration of different types of routinely collected data, for instance by climate sensors or accelerometers, to assess the overall welfare status of an animal or a group of animals throughout (part of) their lives. This can aid the development of early warning systems which enable caretakers to intervene in time. To achieve this, we employ machine learning techniques, such as transfer learning.

Our final thematic focus area concerns human-animal interactions. Here, we focus on animals working for human needs, for instance at the police or in animal assisted interventions with people with PTSD. It is important that these situations and tasks are not detrimental to the animal's welfare. Therefore, it is important to be able to define, recognise and quantify for instance behavioural indicators of stress during such human-animal interactions. Similar to the other thematic areas, we use different object recognition and tracking algorithms to get more information on and to better understand the situations sketched above from an animal welfare perspective.

In addition to these four focus areas we work on two cross-cutting themes which are relevant for all activities of the Lab. The first one concerns the integration of data and techniques. Common challenges in this context include questions of data ownership, privacy, security and data exchange. In addition, we often face analytical difficulties when trying to combine heterogenous data (e.g. data recorded by different sensors on a farm). Besides this, it is questionable, for instance in terms of data ownership, whether it is even desirable to integrate data into a single database. Therefore, researchers of our Lab test new approaches on animal welfare monitoring data, such as federated learning, which means the training of decentralised heterogenous data residing in different locations, companies or institutions. Second, we prioritise the responsible use of AI for animal welfare in all contexts. The use of AI in general is accompanied by several ethical challenges, such as transparency, responsibility and discrimination. However, implementing AI for animal welfare monitoring is even more complex. Animals are recognised as sentient beings who can interact and build relationships with humans. Animals being subjected to technologies which have the potential to fundamentally change our notions and practices of animal handling and care may therefore be reason for new societal concerns. For a successful implementation of AI-based solutions in the context of animal welfare monitoring, it is important to stay ahead of societal concerns and to be able to respond to them. This can only be achieved by analysis of and systematic reflection on the normative questions that may emerge when human-animal relations are mediated, enhanced or disrupted by AI. Therefore, we included the socio-ethical dimension of AI from the very start of the Lab. This makes it possible to use AI applications in a responsible way in both their research and practical implementation phase.

Future work

Within the AI & Animal Welfare Lab we continuously expand our collaborative network with new internal and external (non-academic) partners who are involved in our various student and PhD projects. This is accompanied by regular reflection on the thematic focus of our collaborative initiatives and the Lab as a whole. Does the use of a certain AI-based method really serve the animal? In addition, we aim to make full use of the AI Labs network at Utrecht University and deepen our collaboration with the other labs to share ideas and information on the developed and applied AI solutions. We further seek to team up with similar transdisciplinary initiatives on an international level to share methods and techniques, and probably also serve as an exchange platform for students.

Acknowledgements

We would like to thank the whole team of the AI & Animal Welfare Lab for their valuable contributions to the thematic areas and for shaping the overall focus of the Lab. All research is team sport.

References

1. Mellor DJ. Positive animal welfare states and reference standards for welfare assessment. *N Z Vet J* (2015) 63:17–23. doi:10.1080/00480169.2014.926802

2. Hadorn GH, Biber-Klemm S, Grossenbacher-Mansuy W, Hoffmann-Riem H, Joye D, Pohl C, Wiesmann U, Zemp E. The emergence of transdisciplinarity as a form of research. *Handb Transdiscipl Res* (2008)19–39. doi:10.1007/978-1-4020-6699-3_2

3. Kampourakis K. On the Meaning of Concepts in Science Education. *Sci Educ* (2018) **27**:591–592. doi:10.1007/s11191-018-0004-x

4. Giersberg MF, Bolhuis JE, Rodenburg TB, Meijboom FLB. 80. How smart should resilience be? On the need of a transdisciplinary approach to transform pig production systems. in *Transforming food systems: Ethics, innovation and responsibility (EurSafe 2022)*, eds. D. Bruce, A. Bruce (Edinburgh: Wageningen Academic Publishers), 513–518. doi:10.3920/978-90-8686-939-8_80

Using Multi-Directional Computer Vision for Automated Leg Health Scoring in Broilers

I. Fodor[1], E.D. Ellen[1], M. Taghavi[1], B. de Klerk[2], M. Jacobs[3], A.C. Bouwman[1] and M. van der Sluis[1]

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, the Netherlands.
malou.vandersluis@wur.nl

[2]Cobb Europe, Boxmeer, the Netherlands.

[3]FR Analytics B.V., Wierden, the Netherlands.

Background

Impaired walking ability is commonly seen in broilers and can cause welfare problems [1]. Genetic selection for improved leg health (e.g., in relation to tibial dyschondroplasia or hock burn) is possible [2], but requires detailed observations of individual birds' leg health. Commonly, different leg health aspects are assessed visually. However, such manual observations are time-consuming and subjective. Therefore, automated approaches for scoring leg health and walking ability could have great added value for broiler breeding programs. In this study, video data of broilers walking individually in a corridor were collected, initially with only a back-view camera. We analyzed the resulting videos using a pose estimation model to study different walking characteristics in broilers, to assess whether automated back-view pose estimation can serve as a proxy for manual gait scoring in broilers (*study A*; recently described in more detail in [3]). However, back-view videos of broilers for pose estimation are challenging to obtain under practical conditions, given that this requires a clear view of individual birds walking, while birds are commonly housed in large groups, which can lead to occlusion by other birds. Top-view video recordings might therefore be more feasible to obtain and have less occlusion by other birds, but we cannot record the same poses from top-view video as we can from back-view video. However, there might be correlations between 1) manually determined leg health scores and top-view derived features, and/or 2) between back-view and top-view derived poses. This could provide us with top-view recorded proxies for leg health scoring in broilers, making automated leg health scoring more feasible in practice. Therefore, in the second part of this study (*study B*) we will 1) examine the relationship between manually determined leg health scores and computer vision derived poses/features from top-view video recordings, and 2) determine correlations between computer vision derived poses/features from back- and top-view video recordings of broilers.

Methods

Ethical statement

Data were collected on two broiler farms in the Netherlands, under control of Cobb Europe (Boxmeer, the Netherlands). Cobb Europe complies with the Dutch legislation on animal welfare (study A). Study B is not considered to be an animal experiment under the Law on Animal Experiments, as confirmed by the local Animal Welfare Body (July 11, 2022, Lelystad, the Netherlands).

Study A

Video data were collected for 87 broilers of 33 days old, walking in a corridor individually. The schematic setup of the corridor is shown in Figure 1. An Intel RealSense D415 camera was placed at the start of the walkway, providing a back-view of the broilers walking. Birds were individually placed at the start of the walkway and were then able to walk to the end of the walkway, that ended in their home pen. Before being placed in the corridor, the birds' body weights were recorded (Table 1).

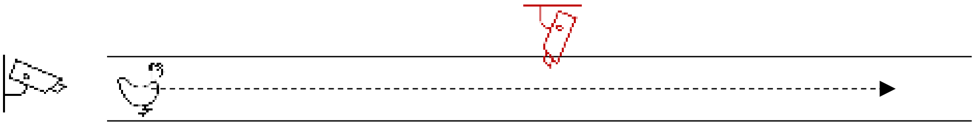


Figure 1. Schematic view of the corridor used in study A. In study B, an additional view was added in a similar setup (red-colored camera in the figure).

Table 1. Body weights and gait scores of the birds in the trials of study A and study B (including standard deviations).

Study	Age of birds (days)	Body weight (g)	Gait score distribution[1]
A	33	2424 ± 177	Score 0 to 1 = 2 birds; score 1 to 1.5 = 13 birds; score 1.5 to 2 = 33 birds; score 2 to 2.5 = 20 birds; score 2.5 to 3 = 11 birds; score 3 to 3.5 = 4 birds; score 3.5 to 4 = 1 bird; scores 4 and higher: 0 birds
B	34	2292 ± 242	Score 0 = 0 birds; score 1 = 74 birds; score 2 = 39 birds; score 3 = 8 birds; score 4 = 2 birds; score 5 = 1 bird

In study A, gait scores were scored by multiple observers and mean gait scores were used. Therefore, these birds are binned in half-score classes with non-inclusive lower and inclusive upper bounds.

From the back-view videos, eight keypoints were detected using a pretrained broiler pose estimation deep learning model: head, neck, knees (left and right), hocks (left and right) and feet (left and right). A more detailed description of the pretrained model can be found in [4]. This model was subsequently retrained using data on the broilers in the walkway at a different age (14 and 21 days old). In the subsequent pose estimation, three birds were excluded because they did not walk during the trial. The pose estimation focused on two main components of walking: the double support phase (standing with both feet on the ground) and the steps at maximum leg lift. From these two phases, seven pose features were derived: hock joint lateral angles, the medial angle of the shank and the horizontal line, the normalized tibiotarsus length, the normalized shank length, the hock-knee distance ratio, the hock-feet distance ratio (all during the double support phase), and the normalized step height (during the steps at maximum leg lift phase). The birds' gait scores were manually determined from the videos, by four experienced observers. They scored the birds' walking ability on a scale from zero to five, using the system described in [5], which is similar - but not equal - to the commonly implemented system described in [6]. In these scoring systems, lower scores represent better walking ability. Each bird received a final gait score that was the mean of the four independent gait scores (**Table 1**). Because of the limited sample size (n = 84), birds were subsequently categorized into having a good gait (scores ≤ 2) or a suboptimal gait (scores > 2).

Study B

Video data were collected for 124 broilers walking through a corridor individually, at 34 days old. The schematic setup of the corridor is shown in **Figure 1**, and was placed inside the home pen, having the birds' regular bedding as floor material. Reolink RLC-510A cameras were placed at the start (back-view) and above (top-view) the walkway. Birds were individually placed at the start of the walkway and were then able to walk to the end of the walkway, that ended in their home pen. Before being placed in the corridor, birds were individually scored for hock burn and footpad dermatitis and their body weights were recorded (**Table 1**). While the birds were walking through the corridor, their gait scores were observed live and recorded manually by a single experienced observer, on the earlier-described scale from zero to five.

The two different video angle views will initially be analyzed independently of the other view at the same timepoint. For the back-view video, the existing pose estimation model of study A will be implemented, after some additional training in the slightly adapted environment. Using the resulting keypoint coordinates, we will determine several pose features including step height, hock joint lateral angles, tibiotarsus and shank length, and hock-knee and hock-feet distance ratios during walking or while the bird is standing still. For the top-view video recordings, we will implement tracking by detection. The resulting bounding box locations and properties over time will allow us to assess 1) the extent of swinging from side to side (lateral movement) through tracking of the bounding box center point over time in the corridor and 2) wing extension during walking through alterations in the width of the bounding box during walking. Although no results are available yet at this point in time, it will be studied whether the walking characteristics recorded from the different video angles differ between birds with different gait scores, to assess how well these recordings can predict birds' gait. Furthermore, we will examine whether the walking characteristics that are collected from the different video views for each bird are correlated. Together, this will provide insight into whether top-view video recordings, alone or in combination with back-view recordings, can serve as a proxy for leg health scoring in broilers.

Results and discussion

The train and test errors of the model in study A were 2.11 and 4.02 px (frame resolution: 1,280 x 720 px), respectively, when considering only the keypoints with a likelihood ≥ 0.6. When comparing the pose estimations from birds with a good gait versus birds with a suboptimal gait (while accounting for body weight), it was observed that birds with a suboptimal gait had sharper hock joint angles (150.7° versus 152.9°, p = 0.042), a lower hock-feet distance ratio (i.e., the feet of these birds were relatively more spread out than the hocks, compared to the broilers with a good gait; 0.84 versus 0.87, p = 0.013), and a lower relative step height (33.2% versus 37.7%, p = 0.002), indicating that they did not lift their feet as high as the birds with a good gait. These differences in walking characteristics between birds with a good versus a suboptimal gait have potential to be used in automated gait scoring approaches. In part B of this study it will be examined whether top-view video recordings can also be used to distinguish between birds with different gait scores, for improved ease of practical implementation of automated broiler gait scoring in the future.

References

1. Knowles, T.G., Kestin, S.C., Haslam, S.M., Brown, S.N., Green, L.E., Butterworth, A., Pope, S.J., Pfeiffer, D., Nicol, C.J. (2008). Leg disorders in broiler chickens: prevalence, risk factors and prevention. *PLoS One*, 3, e1545.

2. Kapell, D.N.R.G., Hill, W.G., Neeteson, A-.M., McAdam, J., Koerhuis, A.N.M., Avendaño, S. (2012). Twenty-five years of selection for improved leg health in purebred broiler lines and underlying genetic parameters. *Poultry Science*, 91, 3032-3043.

3. Fodor, I., van der Sluis, M., Jacobs, M., de Klerk, B., Bouwman, A.C., Ellen, E.D. (2023). Automated pose estimation reveals walking characteristics associated with lameness in broilers. Poultry Science, 102, 102787.

4. Doornweerd, J.E., Kootstra G., Veerkamp, R.F., Ellen, E.D., van der Eijk, van de Straat, T., Bouwman, A.C. (2021). Across-species pose estimation in poultry based on images using deep learning. *Front. Anim. Sci*. 2:791290.

5. Kestin, S.C., Knowles, T.G., Tinch, A.E., Gregory, N. G. (1992). Prevalence of leg weakness in broiler chickens and its relationship with genotype. *Vet. Rec.* 131:190–194.

6. van der Sluis, M., Ellen, E.D., de Klerk, B., Rodenburg, T.B, de Haas. Y. (2021). The relationship between gait and automated recordings of individual broiler activity levels. *Poult. Sci.* 100:101300.

Identifying and tracking group housed hens using ArUco marker backpacks

A. van Putten, M. F. Giersberg and T. B. Rodenburg

Animals in Science and Society, Department of Population Health Sciences, Faculty of Veterinary Medicine,

Utrecht University, Utrecht, The Netherlands

Introduction

The demand for improving animal welfare in livestock production has increased the need for novel approaches to monitor and evaluate welfare in large commercial flocks of laying hens. Current measurements of (mostly negative) health indicators focus on only one aspect of the concept of animal welfare and do not capture sufficient variation over time and between animals. Novel approaches are required to track individuals in group housing over time. Technology for measuring welfare-related indicators has been developed within the field of Precision Livestock Farming (PLF). Advancements in animal tracking technologies have primarily focused on indicators of negative health such as plumage damage or drops in production. A more broad array of welfare indicators could however be measured using such techniques. Automatic tracking of hens has been proposed as a method to monitor individual behavior and activity, as well as social interactions between animals [1]. Initial tracking using computer vision has been promising [2]. Identification of animals is still required, especially in applications for research and animal breeding. Since cameras are already installed to use computer vision, the use of computer readable markers enables data collection from one location [3]. These computer readable markers allow for both identification of animals and the tracking of location and orientation. When recognized, the positions of every ArUco marker are returned for every frame, allowing for precise tracking of location, activity and orientation for each individual. The use of ArUco markers thus allows for identification and position tracking of animals.

Method

The experiment was carried out with the same pullets as described by Kliphuis et al., (2023) and Manet et al., (2023), with the same annotation method as described by Guo et al. (2022). The research project was approved by the central authority for scientific procedures on animals (Centrale Commissie Dierproeven (CCD), the Hague, the Netherlands) under the number AVD1080020198685. Two commercial hybrids, ISA Brown and Dekalb White, were equipped with the backpacks during the 9th week and we collected data at the 13th week of age. We used ArUco markers printed on a previously tested laminated paper backpack design [6]. These were fitted on commercial layer hybrids in an experimental farm with 20 pens housing around 10 animals each. We examined 65 videos of 2 minutes each by running a basic ArUco recognition python script using OpenCV, as well as by doing manual bounding box annotations using the program CVAT. These double checked manual annotations provided our golden standard for examining the ArUco tracking results. Distances in pixel coordinates from the middle of the ArUco markers were used to calculate two novel individual phenotypes: The sum of distance moved and the average distance from one hen to all other hens.

Results and discussion

For the purpose of identification, only one read of the ArUco in a tracked video has to be recorded. For the Dekalb White hens we found 94.0% of the identities of hens which were visible on the video and for the ISA Brown hens we found 61.5%. The suboptimal fit of the backpacks on the larger ISA Brown hens resulted in blocking of the view on the markers compared to the Dekalb White hens. The longer feathers of the ISA Brown hens could be seen blocking the marker from the view of the camera. This did not occur in the Dekalb White hens. Therefore, ArUco is a feasible method for identification if animals are tracked for a minimum of 2 minutes and the design of the ArUco backpack matches the size of the birds.

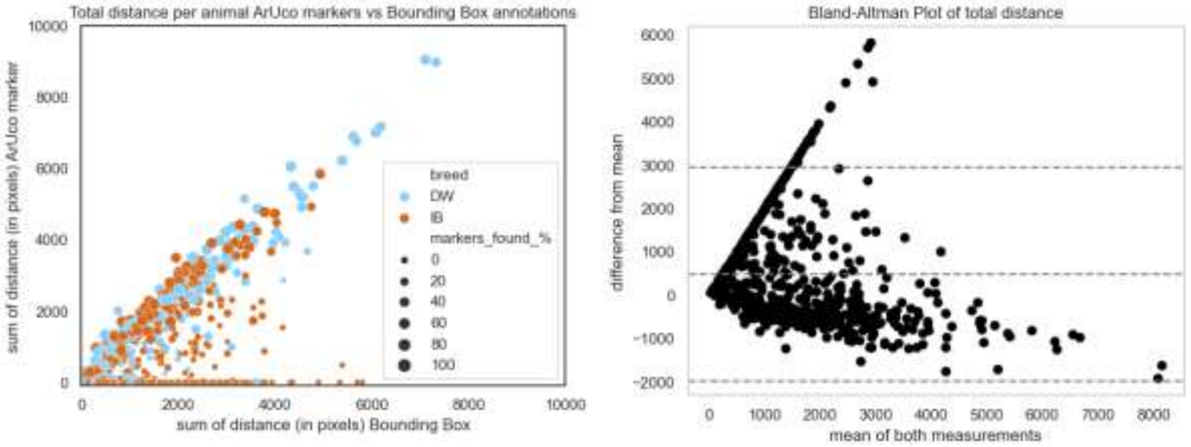
The use of ArUco marker positions was tested for tracking individual phenotypes of movement and positions. With the annotated positions, we could calculate the percentage collected of tracks compared to the theoretical maximum possible. By multiplying the number of frames with the number of animals present in the pen, we calculated the fraction of markers measured against the maximum possible number of markers. With the annotated data available, we were also able to find the correctly tracked fraction based on the animals which were visible. These results can be found in table 1 with the calculation method included.

Table 1. Tracking percentage of ArUco marker backpacks for identification and position tracking for two layer hybrids.

description	Calculation method	ISA Brown %	Dekalb White %	Total %
Expected markers found	$\frac{ArUco\ markers\ on\ visible\ animal}{(nr\ visible\ annotations \times fps)} \times 100\%$	36.3%	76.0%	52.6%
ArUco found compared to maximum	$\frac{ArUco\ markers\ on\ visible\ animal}{(nr\ animals \times nr\ frames)} \times 100\%$	30.1%	62.4%	43.5%
Identification rate	$\frac{nr\ linked\ ArUco\ markers}{(nr\ observed\ animals)} \times 100\%$	61.5%	94.0%	75.2%

When we compare the phenotype of distance moved to the corresponding bounding box annotations, we see a clear correlation (pearson 0.77, p<0.0001) for the available 52.6% marker tracks to the corresponding bounding boxes. This correlation is visible in the correlation plot in Figure 1a.

In order to understand the difference between the tracked data and our golden standard, we plotted the difference from the mean against the mean values in the Bland-Altman plot in Figure 1b. This type of mean-difference plot allows us to spot the agreement between the two methods. . A clear increase can be seen in deviation at higher distances moved in the Bland-Altman plot. Therefore distance moved should be corrected for. The animals without ArUco tracks are now included through manual linking to bounding box information, which will not be possible in future research. Now these points show the amount of missing data to be expected as well as the similar distribution of data compared to the points with ArUco marker tracks.



1a.

1b.

Figure 1. The sum of distance moved per animal as measured by manual annotation and through marker tracking in a correlation plot (1a) and Bland-Altman plot (1b).

Conclusion

ArUco markers on printed backpacks could provide a standalone solution for collecting novel phenotypes in the laying hen production sector. We do however suggest a smaller research setting with group housed animals as the optimal use case. Missing data is one of the greatest challenges that needs to be overcome through backpack design improvements. Even in a situation with a good design of the backpacks, for instance in the Dekalb White hens, 76.0% of the ArUco markers were detected compared to the maximum possible number of marker detections. We need to consider correcting these measurements, for which we could improve design, data filtering, data extraction and measurement type. When combined with computer vision technology, ArUco markers provide a convenient method of identifying animals and providing additional information on location and orientation.

References

1. Ellen, E.D., van der Sluis, M., Siegford, J., Guzhva, O., Toscano, M.J., Bennewitz, J., Van Der Zande, L.E., Van Der Eijk, J.A.J., De Haas, E.N., Norton, T., Piette, D., Tetens, J., de Klerk, B., Visser, B., and Bas Rodenburg, T. (2019). Review of sensor technologies in animal breeding: Phenotyping behaviors of laying hens to select against feather pecking. *Animals*, **9** (3).
2. Guo, Q., Sun, Y., Min, L., van Putten, A., Knol, E., Visser, B., Rodenburg, T., Bolhuis, J., Bijma, P., and N. de With, P. (2022). Video-based Detection and Tracking with Improved Re-Identification Association for Pigs and Laying Hens in Farms. *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 69–78.
3. Alarcón-Nieto, G., Graving, J.M., Klarevas-Irby, J.A., Maldonado-Chaparro, A.A., Mueller, I., and Farine, D.R. (2018) An automated barcode tracking system for behavioural studies in birds. *Methods in Ecology and Evolution*, **9** (6), 1536–1547.
4. Manet, M.W.E., Kliphuis, S., Nordquist, R.E., Goerlich, V.C., Tuyttens, F.A.M., and Rodenburg, T.B. (2023) Brown and white layer pullet hybrids show different fear responses towards humans, but what role does light during incubation play in that? *Applied Animal Behaviour Science*, **267**.
5. Kliphuis, S., Manet, M.W.E., Goerlich, V.C., Nordquist, R.E., Vernooij, H., Brand, H. van den, Tuyttens, F.A.M., and Rodenburg, T.B. (2023) Early-life interventions to prevent feather pecking and reduce fearfulness in laying hens. *Poultry Science*, **102** (8).
6. van der Eijk, J.A.J., Verwoolde, M.B., de Vries Reilingh, G., Jansen, C.A., Rodenburg, T.B., and Lammers, A. (2019) Chicken lines divergently selected on feather pecking differ in immune characteristics. *Physiology and Behaviour*, **212**.

AI in rose-coloured glasses: how close to an individual animal can (should) we come?

Oleksiy Guzhva[1]

[1] Swedish University of Agricultural Sciences, Department of Biosystems and Technology, Box 190, 234 22 Lomma, Sweden, oleksiy.guzhva@slu.se

In the diverse world of animal behavior studies, understanding social interactions plays a crucial role in the conceptual framework of survival and evolution of species ranging from insects to primates, including humans. Exploring the biological underpinnings of these natural interactions, particularly over extended periods and among multiple individuals, is a daunting but critical task. Yet, this domain of science is riddled with complexities, as it demands rigorous and objective quantification of a myriad of behavioral parameters. Challenges include the simultaneous monitoring of multiple individuals, accounting for complex factors such as object occlusion, collision, and maintaining continuity and accuracy of tracking. Moreover, the daunting task of choosing the most fitting model for each unique behavioral study further complicates this field.

Current automatic tools for behavioral analysis, though diverse, typically focus on specialized aspects such as movement tracking or high-level attributes like body poses. However, these tools often provide a piecemeal view, limiting the scope for a comprehensive behavioral understanding. Moreover, the requirement for programming skills to operate many of these tools has restricted their accessibility to broader audiences. In response, there has been a drive toward developing user-friendly tools that allow for holistic behavioral assessments and quantification, but the journey is far from complete.

Recent advancements in multi-object tracking and automated behavioral analysis, powered by new generations of GPUs and semi-automated CNN frameworks (e.g., DeepLabCut, AniPose, and DeepEthogram), represent significant strides in the field. These tools have begun to mitigate some of the traditional challenges associated with manual behavioral observations, such as the limitations imposed by dataset size and the extensive time required to adapt workflows for different studies or species. Nonetheless, a crucial aspect often overlooked in discussions around these advanced tools is the definition and adequacy of the "few examples" required to train the underlying neural networks for effective performance.

Deep learning, a subset of machine learning methodologies, has revolutionized fields beyond behavioral science, including industrial image recognition and autonomous vehicles. Its reliance on large volumes of labeled data is well-known, yet the gathering of such data, especially in specialized fields like animal behavior, is frequently a resource-intensive task. This has led to the exploration of semi-supervised learning, where unlabeled data complement small labeled datasets to enhance model accuracy. However, challenges such as significant distribution mismatches between labeled and unlabeled data can severely impact the performance of semi-supervised models, a situation further exacerbated in the unique “no studies are alike” context of behavioral observation studies.

In addressing these challenges, recent body of research suggests that for effective classification models in behavioral studies, a minimum threshold of around 1,000 samples per behavioral class might be required to achieve parity with manual observations. This perspective is supported by a review of 167 articles, which found limited attempts at pre hoc or post hoc sample size determination specifically for machine learning applications in classification studies [1]. Notably, many pre hoc methods, such as those proposed by researchers at MIT, are tailored for neural networks and suggest a prohibitively large number of samples for even modest network structures [2]. The absence of widespread, robust methods for sample size determination in the field is a pressing issue, yet it also presents an opportunity for innovation in how ethograms are created, and behavioral sequences are selected and annotated [3, 4].

The quest for the optimal sample size in behavioral studies is not just a numerical challenge; it's a multifaceted problem involving the temporal context of behaviors, the variance of behavioral manifestations across species, and the need for definitions and standards that can be applied consistently across diverse research settings. Studies indicate that as sample sizes increase, effect sizes and classification accuracies also increase, provided the datasets have high discriminative power between classes. This leads to a conundrum where the pursuit of larger samples for better accuracy must be balanced against the practicalities and costs of data collection, emphasizing the adage that sometimes "better data beats more data [5, 6]".

This work aims to engage scientists from different backgrounds in a rigorous, hands-on discussion with several demo examples comparing automated behavioural assessment frameworks in terms of actual sub-sample size, annotation strategy and potential pitfalls when creating/selecting ethograms and beyond.

In conclusion, the field of automated behavioral analysis stands at a crossroads, with significant advances on one hand and persistent challenges on the other. As the community continues to innovate and adapt, the journey towards understanding the complex tapestry of animal behaviors through technological lenses promises to be as enriching as it is essential. The future of this field lies in the successful amalgamation of interdisciplinary research, advanced computational techniques, and a nuanced understanding of the biological, ethical, and logistical dimensions of studying animal behavior.

References

1. Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., … & Tyrrell, P. N. (2019). Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Canadian Association of Radiologists Journal, 70(4), 344-353.

2. Baum, E. B., & Haussler, D. (1989). What size net gives valid generalization?. Neural computation, 1(1), 151-160.

3. Vapnik, V. (2000). The nature of statistical learning theory. Springer.

4. Rokem, A., Wu, Y., & Lee, A. (2017). Assessment of the need for separate test set and number of medical images necessary for deep learning: a sub-sampling study.

5. Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?

6. Rajput, D., Wang, W. J., & Chen, C. C. (2023). Evaluation of a decided sample size in machine learning applications. BMC bioinformatics, 24(1), 48.

AI-PigNet: Insights into the social interaction of pigs through automated data and social network analysis

Saif Agha[1], Eric Psota[2], Simon P. Turner[3], Craig R. G. Lewis[4] and Andrea Doeschl-Wilson[1]

[1]The Roslin Institute, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, UK

[2] PIC North America, Hendersonville, TN, USA

[3]Animal and Veterinary Sciences Department, Scotland's Rural College, West Mains Road, Edinburgh EH9JG, UK

[4] PIC, C/Pau Vila no. 22, Sant Cugat del Valles, 08174 Barcelona, Spain.

Abstract

The aim was to explore the feasibility to use data generated from on-farm AI monitoring systems to construct social networks that describe the social dynamics of growing pigs. Results show that automated position and activity data exhibited a matching accuracy of over 97% with human observations. Social Network Analysis (SNA) of the automated data provide valuable insights into social structures and informative behavioural phenotypes. These can be employed for improving animal performance, welfare, and health.

Background

Social interactions are valuable indicators of animal behaviour and should be considered for simultaneously improving livestock performance, health, and welfare. Social network analysis (SNA), applied to data extracted from manually decoded video-recordings of group housed pigs, has shown great potential in identifying novel behavioural phenotypes that quantify the direct and indirect role of each animal in pen level aggression and resulting skin lesions [1,2]. Furthermore, these novel phenotypes are heritable and favourably correlated with other economically important traits [3,4]. In practice, however, it is often not feasible to monitor animal behaviour on a large scale and over a long period of time by human observers. Automatic monitoring systems and cutting-edge AI technologies e.g., deep learning (DL), hold great promise for facilitating the capture of social interactions and measuring animal behaviour. The overall aim of this study was to explore the scope and feasibility to use data generated from on-farm computer monitoring systems to construct social networks and identify informative behavioural phenotypes that describe the social structure within pens of growing pigs.

Material and Methods

Data was derived from automated recording systems, that provide 2D camera images and video records of contacts and social interactions of >1000 growing pigs grouped in different pens, at the commercial farms of the PIC breeding company. The automated system provides data for the ear tag, the number of seconds that has elapsed since the start of video recording, a binary indicator that the identified pig was eating or drinking with the specific time and the posture of each animal, e.g., lying lateral, lying sternal, sitting, standing. In addition, the system incorporates in-house developed DL routines to provide the 2-dimensional X-Y coordinates of the pig's front and rear torso, defined as the distance (in meters) from the top-left corner of the pen [5].

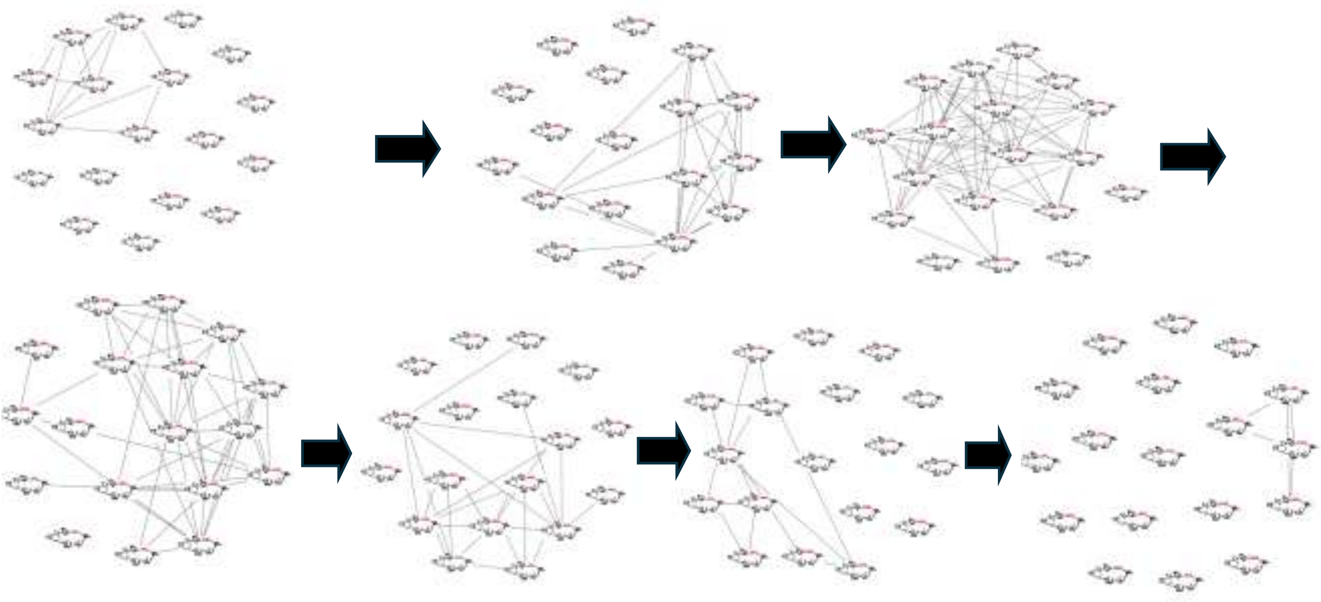
Pigs were marked with livestock paint, making it possible to visually validate the accuracy of the posture and activity of the pigs and the coordinates of their front and rear, by comparing annotations provided by the automated system with their observable locations and postures in the video records. By leveraging various R packages *spatsoc*, *asnipe*, *sna*, and *igraph*, a method was designed to handle the tracking data files generated by the automated system. The method systematically reads the X-Y coordinates of the animals, specifically front and rear, along with the corresponding timestamps provided by the automated system. It then assigns to each group a temporary group-time variable for subsequent analyses of social interactions in that group over a specified time period. Subsequently, the method employs a spatial analysis to identify animals in direct contact with each other, based on their proximity. In this study, proximity was identified between animals, within the same time group, by a 0.5-meter threshold, allowing for the characterization of animal interactions and movements during specific time intervals. SNA was applied to obtain SNA centrality measures, at the group- and individual-level to quantify the social structure for each group. Furthermore, the evolution of social interactions within each pen over time were investigated by the Multi-Dimensional Scaling (MDS) projection based on the Hamming distances of social interaction patterns between networks.

Results

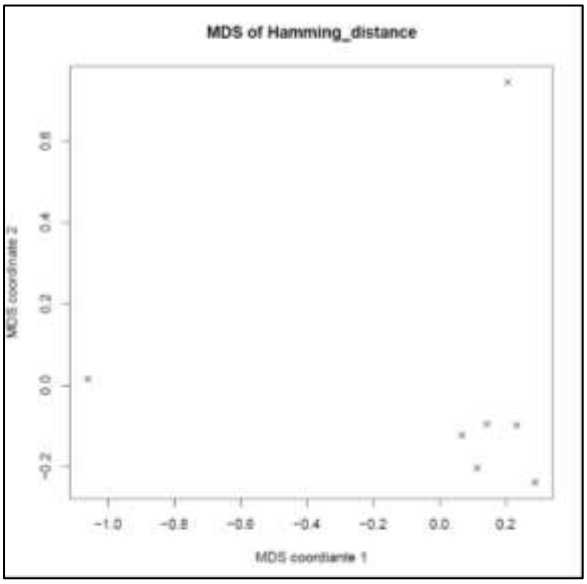
Our results showed that the automated data, derived from the DL algorithm, demonstrated a matching accuracy for the position and activity of the pigs of over 97% compared to human observation. Notably, the validation of the coordinates for the front and rear, involving the ordering of animals based on the X and Y coordinates, revealed no errors in the derived data, further affirming the robustness of the DL methods. This comparison served as a crucial step in ensuring the reliability and precision of the derived data. Furthermore, social networks were successfully constructed from the automated data for each identified group-time (Figure 1a). Additionally, the MDS projection based on the Hamming distances of social interaction patterns between these group time, e.g., networks, within pen was obtained, highlighting the evolution of social interactions within each pen over time (Figure 1b). For each identified group, SNA traits at the group-level were calculated e.g., group-degree, group-density, group-closeness, group-betweenness, group-eigenvector centrality and largest clique size within each identified group. Furthermore, individual SNA traits were calculated for each animal within each identified group e.g., degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and clustering coefficient.

Figure 1. Social networks of each of identified group-time showing the evolution of social interactions over time within a pen (a) and the Multi-Dimensional Scaling (MDS) projection, performed based on the Hamming distances between networks, is showing the similarity/dissimilarity of social behaviour patterns between groups. Closer points in the MDS indicate a smaller Hamming distance between the corresponding networks, implying that these networks exhibit similar behaviour patterns (b).

(a)



(b)



Conclusion

The validation process undertaken to assess the accuracy of position and activity measurements for the on-farm automated monitoring system has yielded promising results. This study emphasizes the feasibility of integrating automated monitoring systems, advanced AI technologies, and SNA in capturing social interactions and measuring animal behaviour in commercial farms.

References

1. Foister, S., Doeschl-Wilson, A., Roehe, R., Arnott, G., Boyle, L., Turner, S. (2018). Social network properties predict chronic aggression in commercial pig systems. *PLoS ONE*, 13, doi.org/10.1371/journal.pone.0205122.
2. Agha, S., Fàbrega, E., Quintanilla, R., Sánchez, J.P. (2020). Social Network Analysis of Agonistic Behaviour and Its Association with Economically Important Traits in Pigs. *Animals*, 10, 2123. doi.org/10.3390/genes13040561.
3. Agha, S., Foister, S., Roehe, R., Turner, S.P., Doeschl-Wilson, A. Genetic Analysis of Novel Behaviour Traits in Pigs Derived from Social Network Analysis. *Genes* 2022, 13, 561. doi.org/10.3390/genes13040561.
4. Agha S., Turner, S. P., Lewis, C. R. G., Desire, S., Roehe R., and Doeschl-Wilson A. (2022b). Genetic Associations of Novel Behaviour Traits Derived from Social Network Analysis with Growth, Feed Efficiency, and Carcass Characteristics in Pigs. *Genes*, doi.org/10.3390/genes13091616.
5. Liu D., A. Parmiggiani, E. Psota, R. Fitzgerald, T. Norton et al. (2023). Where's your head at? Detecting the orientation and position of pigs with rotated bounding boxes. *Computers and Electronics in Agriculture,* doi.org/10.1016/j.compag.2023.108099.

Ethical statement

No ethical approval is required. The data are obtained from non-invasive on-farm monitoring system and provided by the industry partner Genus – PIC.

Challenges and opportunities of working with PLF technology and AI researchers: the applied ethologist's experience

E.M. Baxter[1*], K.M.D. Rutherford[1], M.C. Reeves[1,2] and R.B. D'Eath[1]

[1]Scotland's Rural College (SRUC), Animal & Veterinary Sciences Research Group, West Mains Road, Edinburgh, EH9 3JG, UK

[2]University of Edinburgh, Royal (Dick) School of Veterinary Studies, Easter Bush, EH25 9RG, UK

***Corresponding author: Emma.Baxter@sruc.ac.uk**

Introduction

Artificial Intelligence (AI) in agriculture includes various Precision Livestock Farming (PLF) technologies, together with AI software and integration methods (machine learning, deep learning and reinforcement learning), that are intended to improve performance, efficiency, health and welfare. Whilst many of these technological advances offer such opportunities, they also present risks, including potential to threaten animal health and welfare. For example, the lack of user-driven design for wearable sensors can result in injury to the animal and changes in behaviour by both wearer and conspecifics. In addition, though many technologies are suggested only as a tool to aid the stockperson, there are risks of over- or under-reliance on PLF, potential for down-skilling of the work force, and potential to damage the human-animal relationship which is an important aspect of animal welfare. There are wider ethical debates about a loss of autonomy for the farmer [1, 2] and a general uneasiness around AI is apparent from other stakeholders, including consumers, who may already have concerns about intensive farming practices and potential for further objectification of animals [3, 4].

As applied ethologists we are increasingly involved in projects where various aspects of AI are being developed that aim to identify and track individuals, recognize their different behaviours and affective states and monitor their health and welfare. Projects can involve multiple stakeholders, with varied backgrounds, expertise and objectives. Such transdisciplinary research is commonplace in animal welfare science which is, by its very nature, a blend of various subdisciplines including ethology, zoology and veterinary sciences [5]. Adopting AI disciplines into this field presents exciting opportunities and the prospect of any or all of the aims outlined above being achieved is potentially ground-breaking. However, despite the rapid pace of technological advancements, it is very early days in this endeavour and there are many hurdles to overcome. Some of these have recently been reviewed and include some of the welfare and ethical risks introduced earlier [1,2,3]. Some involve the fine-grain details of undertaking projects combining very different disciplines that are rarely articulated in peer review articles but are introduced here to encourage discussion and develop frameworks that might help realise the full potential promised by AI.

Trusting the technology

Trusting the technology is a critical aspect of accepting AI in agriculture, but the journey from inception to market availability is not always transparent, particularly concerning validation. For example, in a recent review of pig PLF technologies, authors determined that just 5% of commercially available sensor systems had been externally validated [6]. This might explain some of the reasons why on-farm adoption has lagged behind the rapid pace of technological advancements. Other reasons cited by PLF-cautious farmers include concerns about data privacy and the influence of PLF technology on human-animal relationships and farmers' duty of care to the animals [7]. In the same study farmers highlighted practical limitations, including poor internet connectivity and the inability to use PLF data for decision-making due to needing to first complete daily on-farm tasks.

Practical limitations

Practical limitations are significant and not only experienced by farmers, but also animal scientists undertaking validation studies. Often the equipment being tested was never designed to cope with the rigours of farm life and, it is not unusal for the start of projects to be dominated by a substantial period of trial and error before experiments can begin. For example, trials capturing image data need to determine camera type, consider how to reduce glare, cope with dust, cobwebs and dirt, as well as what will be optimal placement to avoid damage from animals and/or farm equipment whilst ensuring maximum field of view. Trying to introduce technologies to extensively managed species sees similar challenges, as well as a host of other barriers to overcome. For example, data from wearable sensors will be favoured over camera-based systems but battery life can be a challenge and usually there is a trade-off between battery life and the ability to transmit data in real time. Connections to proximity beacons are sometimes lost and data can stop being collected for periods of time and/or be completely lost if real-time data transmission relies on a reliable gateway-to-network server connection [8]. Collars or tags can be lost or damaged and though these challenges are to be expected in proof-of-concept trials, they would need to be resolved before the technology is commercialised for farmers. It is worth reiterating the importance of species-specific tools that have been developed with the wearer in mind and have been properly tested across different age categories, in different management systems and during different situations for that species. Many wearable sensors have been developed for dairy cattle and been repurposed for smaller ruminants [9,10] or even tried on pigs, a species highly adept at finding the weakness in any equipment. Though the equipment often comes off second best, there are serious animal welfare concerns when sensors are too heavy or unwieldy for the wearer, cause injury or changes in behaviour. We propose that any technology should adhere to a hippocratic oath of ‘first do no harm’. In human medicine, a new drug or technology would be subject to rigorous quality assurance processes before being made readily available on the market. For example wearable digital health technologies such as FitBits are subject to quality assurance standards. The software and hardware components have to meet specific requirements that must comply with standards determined by bodies such as the IEEE and FDA [11,12]. Part of those quality assurance processes involve scrutiny of the ground-truth data upon which the technology is being compared to or the model trained – i.e. the validation process.

What constitutes proper validation for AI applications in animal behaviour and welfare?

Designing experiments to generate robust ground-truth data for AI validation studies starts with discussions about sample size. Here we encounter problems. What sample size is required to generate enough statistical power in these experiments? In any experiment, it is important to justify animal numbers and is a requirement for most funding bodies in grant applications. In machine vision work, for example, it is not clear how we determine adequate sample sizes (both animals and images) because traditional power calculations are not always suitable. We might refer to previous studies for guidance but in pig PLF, for example, most machine vision research papers use data from one pen of animals to train deep learning neural networks to find and track pigs, and perhaps also to recognise one behaviour. Its capacity to work with other groups of pigs, on different farms, or under different light conditions is never tested, and the next paper rarely builds on the last one [14, 15]. There can be different fundamental mindsets between animal scientists (thinking about the scientific method, approaches to experimental studies, hypothesis testing) and computing engineers and the lack of agreement could cause some uncertainty for those making decisions on funding. We can draw comparisons from those working in digital medicine, which like animal welfare is equally multidisciplinary. Researchers describe the diversity of their field as undoubtedly valuable but full of confusion regarding terminology and approaches, with isolated silos of knowledge producing evidentiary standards that are not aligned across the different communities, ultimately slowing advancement of digital medicine for improved health, healthcare, and health economics [13]. This mirrors what we are experiencing as AI becomes integrated into animal behaviour and welfare science. Lessons can be learned from the experience of those working in digital medicine. Fogel and Kvedar (2018) propose a three-component framework including (1) verification, (2) analytical validation, and (3) clinical validation and they stress the importance of agreeing on common terminology, proposing definitions intended to bridge disciplinary divides.

‘Rubbish in, rubbish out’ – the importance of data quality and the challenges of data size

Algorithms are only as good as the data upon which they are trained [16] and having access to properly validated (i.e. labelled) data for model training and testing is a key starting point. The 'big data' being generated from sensor-aided monitoring and other AI applications is heralded as a major advantage and selling point [17] but there are challenges when handling large datasets. Large data files require a lot of computing power, beyond what most university-issued laptops have. Powerful, specialised machines are required to process the data and to ensure proper storage for quality control purposes. During the validation process, these large data sets have still to be labelled by experts. As applied ethologists we are no strangers to the systematic and quantitative recording of animal behaviour. This is our bread-and-butter and forms the foundations of animal welfare science but it is extremely time consuming. Properly labelled datasets are therefore highly valuable and, ideally should be made available so future studies can build upon solid foundations. Such openess can be a challenge, especially if an end product is to be marketed, but it is an important part of enhancing the validation of AI applications in practice. There are other data challenges not fully discussed here, not least the problem of data interpretation, particularly in the development of tools designed to alert farmers to potential problems. For example - are there some sensor readings which are diagnostic of a specific problem on any farm (like a blood test for a specific disease)? Or does the system need to learn what is normal for that specific pen, building or farm before being able to spot anomalies? When an anomaly is detected, what does it mean, and how should the farmer respond? [14]. To answer these questions requires validation trials involving multiple stakeholders – including the end-user.

By understanding some of the common pitfalls occurring when developing AI applications in animal welfare and behaviour fields, we can develop frameworks to successfully mitigate some of the challenges and realise the opportunities more effectively.

References

1. Schillings, J., Bennett, R. and Rose, D.C., 2021. Animal welfare and other ethical implications of Precision Livestock Farming technology. *CABI Agriculture and Bioscience*, 2(1), pp.1-4.
2. Tuyttens, F.A., Molento, C.F. and Benaissa, S., 2022. Twelve threats of precision livestock farming (PLF) for animal welfare. *Frontiers in Veterinary Science*, 9, p.889623.
3. Bos, J.M., Bovenkerk, B., Feindt, P.H. and Van Dam, Y.K., 2018. The quantified animal: precision livestock farming and the ethical implications of objectification. Food Ethics, 2, pp.77-92.
4. Giersberg, M.F. and Meijboom, F.L., 2021. Smart technologies lead to smart answers? On the claim of smart sensing technologies to tackle animal related societal concerns in Europe over current pig husbandry systems. Frontiers in Veterinary Science, 7, p.588214.
5. Camerlink, I., 2021. Interdisciplinary Research to Advance Animal Welfare Science: An Introduction. In Bridging Research Disciplines to Advance Animal Welfare Science: A Practical Guide (pp. 1-16). GB: CABI.
6. Gómez, Yaneth, Anna H. Stygar, Iris JMM Boumans, Eddie AM Bokkers, Lene J. Pedersen, Jarkko K. Niemi, Matti Pastell, Xavier Manteca, and Pol Llonch. "A systematic review on validated precision livestock farming technologies for pig production and its potential to assess animal welfare." Frontiers in Veterinary Science 8 (2021): 660565.
7. Akinyemi, B.E., Vigors, B., Turner, S.P., Akaichi, F., Benjamin, M., Johnson, A.K., Pairis-Garcia, M.D., Rozeboom, D.W., Steibel, J.P., Thompson, D.P. and Zangaro, C., 2023. Precision livestock farming: a qualitative exploration of swine industry stakeholders. Frontiers in Animal Science, 4, p.1150528.
8. Waterhouse, A., Holland, J.P., McLaren, A., Arthur, R., Duthie, C.A., Kodam, S. and Wishart, H.M., 2019, August. Opportunities and challenges for real-time management (RTM) in extensive livestock systems. In The European Conference on Precision Livestock Farming (pp. 20-26).
9. Barwick, J., Lamb, D., Dobos, R., Schneider, D., Welch, M. and Trotter, M., 2018. Predicting lameness in sheep activity using tri-axial acceleration signals. Animals, 8(1), p.12.
10. Kaler, J., Mitsch, J., Vázquez-Diosdado, J.A., Bollard, N., Dottorini, T. and Ellis, K.A., 2020. Automated detection of lameness in sheep using machine learning approaches: Novel insights into behavioural differences among lame and non-lame sheep. Royal Society open science, 7(1), p.190824.

11. IEEE Computer Society. IEEE Standard for System, Software, and Hardware Verification and Validation. IEEE Std 1012-2016 (Revision of IEEE Std 1012-2012/ Incorporates IEEE Std 1012-2016/Cor1-2017) 1–260 (2017). https://doi.org/10.1109/IEEESTD.2017.8055462.
12. U.S. Department Of Health and Human Services, U.S. Food and Drug Administration, Center for Devices and Radiological Health & Center for Biologics Evaluation and Research. General Principles of Software Validation; Final Guidance for Industry and FDA Staff, 47 (2002)
13. Fogel, A.L. and Kvedar, J.C., 2018. Artificial intelligence powers digital medicine. NPJ digital medicine, 1(1), p.5.
14. D'Eath, R.B., 2022. Precision Livestock Farming Technologies for Pig Welfare-Policy Spotlight.
15. Wurtz, K., Camerlink, I., D'Eath, R.B., Fernández, A.P., Norton, T., Steibel, J. and Siegford, J., 2019. Recording behaviour of indoor-housed farm animals automatically using machine vision technology: A systematic review. PloS one, 14(12), p.e0226669.
16. Siegford, J.M., Steibel, J.P., Han, J., Benjamin, M., Brown-Brandl, T., Dórea, J.R., Morris, D., Norton, T., Psota, E. and Rosa, G.J., 2023. The quest to develop automated systems for monitoring animal behavior. Applied Animal Behaviour Science, 265, p.106000.
17. Weersink, A., Fraser, E., Pannell, D., Duncan, E. and Rotz, S., 2018. Opportunities and challenges for big data in agricultural and environmental analysis. Annual Review of Resource Economics, 10, pp.19-37.

Towards early prediction of disease in farm animals: Vision-AI for automatic analysis of face, gait and motion for sheep behavior understanding

M. Mahmoud

School of Computing Science, University of Glasgow, Glasgow, United Kingdom. marwa.mahmoud@glasgow.ac.uk

Abstract

Lameness is a critical indicator of sheep health conditions and is a serious animal welfare concern. Traditional methods of assessing lameness are slow and costly. An automated lameness detection system that can detect early signs of lameness (or other painful conditions) through continuous monitoring of the flock can save on treatment costs and improve animal welfare. With the emergence and rapid development of computer vision and deep learning technologies, the lameness detection problem has the possibility to be solved efficiently and effectively. We present our work on using inference models in a hierarchal model to detect changes in the animal's facial expressions, gait and motion from video. For facial expressions analysis, we leverage transfer learning and Convolutional Neural Networks (CNNs) for face detection, pose estimation and pose-informed landmarks detection. For gait analysis, we build on the extracted body parts from DeepLabCut and use machine learning models to predict lameness by extracting spatio-temporal features of the body and movements. We evaluate our models on a dataset of videos of sheep that suffer from lameness recorded in their natural environment. Our models achieve up to 80% accuracy in detecting pain from the sheep facial expressions and 77.5% accuracy in predicting lameness scores from automatic analysis of gait. Although our proposed methods are evaluated on a dataset specifically for lameness, but it can be utilised and extended to detect other diseases that may elicit a pain response in sheep. Our work shows the potential of using computer vision , as a non-invasive technique, that can allow for continuous monitoring of sheep – and other animals – for early prediction of diseases and to improve animal's welfare.

Introduction

Animal welfare has become a common legal and economic concern in recent years. It is an important measure of sustainable consumption and production. Disease and poor health conditions in livestock could increase the spread and severities of some zoonotic diseases. Also, injured and stressed livestock is more likely to release more pathogens during transport and slaughter [1]. Lameness is one of the most important indicators for some infectious bacterial diseases of sheep, such as footrot, bluetongue and mastitis [2]. And the lameness leads to weight loss, increased mortality and decreased wool production, which eventually causes economic loss to the sheep industry [3]. Making the correct diagnosis is key in the treatment and control of such diseases. However, sheep are stoic prey animals which rarely showing obvious signs of disease in the early stage, which makes less efficient for the human to visually inspect their health conditions.

The primary way of identifying lameness is by visually observing the sheep appearances and behaviours. The five-point gait scoring method [4] is one of the most common methods used in sheep lameness evaluation, based on assessing the definite limp and uneven gait. Although this manual assessment method is feasible to identify lame sheep, it is very time-consuming and economically costly. Besides, the judgement is strongly influenced by the inspector's experience and perception. Thus, introducing an automated lameness detection system is necessary, to increase the detection efficiency and prevent disease aggravation, which ultimately improves animal welfare and reduce the cost of farming [5].

In this work, we present our recent work on using computer vision for automatic detection of signs of pain from the facial expressions of sheep as well as our methodologies for automatic detection of lameness from automatic detection of gait and movements behavioral changes.

Open-Sheep-Face: A Comprehensive Application for Sheep Face Analysis and Pain Estimation

This section details a comprehensive, automated pipeline [6] that leverages machine learning and computer vision techniques to detect pain in sheep by analyzing their faces. We present the step-by-step construction of this pipeline, from face detection, head pose estimation, facial landmark localization, and pain estimation.

Face Detection

YOLOv5 [7] was retrained to enable sheep face detection using the same sheep dataset as [8]. This dataset contains accurate bounding box labels for sheep faces, which were utilized during the training process. Initially, the dataset was pre-processed with several steps, such as auto-orientation, resizing to 640×640, and conversion to grayscale. Subsequently, additional data is generated through data augmentations. Image augmentations involved horizontal flipping, random rotation within the range of -15° to +15°, and shearing between -15° and +15°. To further enhance the dataset, the bounding boxes were horizontally flipped, and the brightness was adjusted by -25% to +25% during augmentation. The dataset was split into training set, validation set and test set respectively.

Head Pose Estimation

Transfer learning is utilized to create a model capable of estimating sheep head poses by leveraging a deep network originally designed for human head pose estimation. The Hopenet network [10] is selected for its specific focus on head pose estimation and its superior performance among the networks trained in [10]. As the base model, a pre-trained Hopenet model trained on the 300W-LP dataset [11] is employed. To establish the ground-truth head pose for images in the sheep facial landmarks in the wild (SFLW) dataset, a 3D base landmark model is created manually, featuring a neutral head pose (0 yaw, pitch, and roll) and an average head shape [12]. A RANSAC [13] based method implemented by OpenCV [14] is then used to solve the perspective-n-point (PnP) problem, recovering the approximate head pose with the 3D points of the manually generated landmark model and the 2D annotated landmarks for each image [12]. The base model is fine-tuned on the SFLW-NCA dataset [19] with additional augmentation. Here, SFLW-NCA dataset is the SFLW dataset applied with horizontal mirroring, TPS warping and rotation augmentations, with the number of times each original image is used weighted by negatively correlated augmentation (NCA) [8]. The augmentation includes randomly flipping the input images in the x-direction and translating the image in x and y directions randomly by no more than 7% [10].

Facial Landmark Localization

The ERT [15] method was employed to identify landmarks on sheep, which required ground-truth values for the absolute yaw angle, bounding box, and facial landmark coordinates to train the model accurately. The ERT method is a decision tree used for regression tasks that integrates multiple regression tree algorithms into a single model for prediction. The cascade of regressors approach was chosen to implement ERT due to its high accuracy rate in detecting facial landmarks in humans. An existing Python implementation of the ERT algorithm for detecting human facial landmarks, created by Xiao (2019) [16], was used. The provided code creates a cascade forest by iteratively building a sequence of random forests, where each forest is trained to predict the error of the previous forest. The output of the final forest is then added to the initial predictions of the first forest to obtain the final prediction.

Pain Estimation

To assess pain levels in sheep, HOGs [17] feature descriptor was utilized. HOG generates a feature vector by calculating gradients' magnitudes and angles within an image, where an input image is partitioned into small sections and pixels are grouped into small cells. This process helps to calculate gradients, and orientation bins are formed. To identify pain indicators, the prominent facial features of sheep, such as both ears, eyes, and mouth regions, were identified, and the images were cropped to isolate these regions [18]. Then, these regions were converted into numpy representation, and HOG was applied to compute feature vectors for these regions. The feature vectors were mapped to a single sheep and flattened to create a representation vector. In addition to these features, geometric features such as the global angle of each ear and the distance between the ear root landmarks were also extracted and included in the analysis.

Experimental Evaluation and results

To evaluate the face detection step, the YOLOv5-based sheep face detector is trained using cross-validation. Evaluation metrics recommended by YOLOv5 [7], including precision, recall, F1 score, mAP@0.5, and mAP@0.5:0.95, are utilized [9]. With 5-fold cross validation, the sheep face detector achieves an average precision of 0.980, recall of 0.979, F1 score of 0.980, mAP@0.5 of 0.993, and mAP@0.5:0.95 of 0.93.

For the head pose estimation step, Training was performed with 16-batch increments over 16 epochs using an initial learning rate of 0.0001. The model is fine-tuned, and the one with the lowest validation loss is chosen as the sheep head pose estimation model. Mean Absolute Error (MAE) [10], Pearson's Correlation Coefficient (PCC), and Sign Agreement metric (SAGR) are employed for evaluation [19]. On the SFLW dataset, the fine-tuned model achieves an average MAE of 6.58, PCC of 0.75, and SAGR of 0.80 for the three angles (yaw, pitch, roll).

For the facial landmark localization step, The PI-ERT model is optimized through grid search, utilizing Mean Squared Error (MSE) and Mean Normalized Error (MNE) metrics for evaluation. Augmentation techniques, including TPS warping, are applied to the sheep dataset [19]. The PI-ERT model, with 10 forests and 500 trees each, achieves a Mean Normalized Error of 0.044 [19].

For the final step, the pain estimation, A Support Vector Machine (SVM) [20] classifier is trained for pain estimation in sheep, evaluated using accuracy, precision, recall, and F1 score. Dataset augmentation techniques, including image and landmark mirroring, are applied. The SVM model achieves a precision score of 83% [21], with recall, F1, and accuracy scores reaching 80% for all metrics.

Automatic detection of Gait

We extended the work of the automatic analysis of face to the whole body to investigate the potential of disease modelling based on animal gait analysis. In this section, we present a hierarchical model for lameness detection from gait analysis. First, the process of sheep body detection is explained. Then, the body joints localization pipeline is outlined. Finally, the feature construction and machine learning model detection is described.

Dataset

The data was collected from eight commercial farms in the UK when footrot was reported. A total of 73 animals suffered from footrot and the remaining 38 sheep were in the control group. Furthermore, all study targets were over one year of age when the data was collected. The study participants were separated into 3 treatments groups. Sheep in both group 1 (N=37) and group 2 (N=36) were injected with antibiotics and tulathromycin for footrot treatment, whereas group 2 also received an anti-inflammatory drug. Participants in group 3 (N=38) had no signs of footrot during the assessments from a veterinarian. In order to quantitatively evaluate the lameness of research participants, the gait score of sheep in all 3 groups are assessed using the five-point gait scoring method (score; 0: not lame, 1-3 moderately lame, 4: severely lame). All lameness records and their corresponding sheep ID were stored.

Multiple videos of sheep in three groups were taken on day 0 (day of footrot identification), day 7 and day 90 separately, the videos were captured by a handheld camera with a consistent resolution (1080 x 1920 px).

Sheep Detection Model

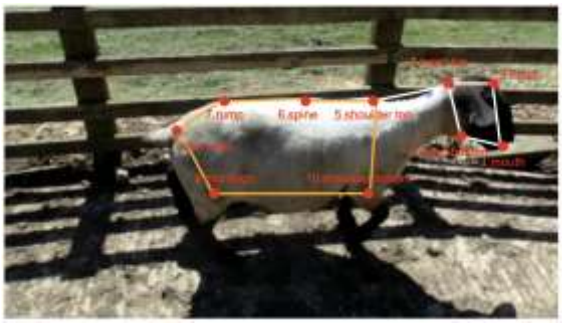
Considering the dataset is relatively small to train a deep neural network, transfer learning was applied. That is, a CNN model pre-trained on a general dataset was used as the initial state of our backbone, and fine tune the last layer for our 3 classes based on the research data. The main advantage of using transfer learning is that the training time is dramatically reduced since the network was already trained with a substantial amount of data. And the model could reuse its weights to extract abstract features in the high-level representations, which ultimately improve the accuracy and generalization ability, also circumvent the need for considerable labelled training data.

MS COCO dataset was chosen in this project since the objects in the dataset includes many animals such as sheep, dog, cow, horse and zebra, which means the pre-trained model could already abstract representation to some extent of animals including our research target - sheep. A ResNet-50 network pre-trained on MS COCO was chosen as the backbone of RPN and Fast R-CNN.

There are in total 500 annotated images and more than 600 bounding boxes from all 8 farms were used (450 images used for training and 50 images used for testing). The model was trained using an SGD optimizer for 50 epochs with 8 images per batch until the loss converged. The learning rate was initialized to 5 × 10−3, momentum was set to 0.9, and the L2 regularization parameter (weight decay) was set to 5 × 10−4. Furthermore, the input images from the training set were randomly cropped and horizontally flipped in order to improve the transformation invariance of the model.

Sheep Pose Estimation

We used the popular tool DeepLabCut for sheep pose estimation. DeepLabCut utilizes the fully-convolutional feature detector in the DeeperCut algorithm [22], which is one of the state-of-art human pose estimation methods. To train the model, 1352 annotated images cropped from the videos in 8 farms were used. And 95%(N=1284) data were used for training and 5%(N=68) were used for testing. A ResNet-50 based neural network was used with default parameters and trained end-end with SGD optimizer for 20000 training iterations. Since the size of the input image varies, the batch size was set to 1. The implementation was used the DeepLabCut Python package with version 2.2.



Lameness detection

A total number of 50 short videos were manually edited from all 184 videos across the 8 farms. The video duration varies from two seconds to ten seconds. The 50 videos contain 38 individuals (21 cases, 17 controls) of different combination of sheep ID and shooting parameters. Although some videos contain the duplicate participants, but the experimental environments and shooting parameters are very different from video to video. We assigned all videos (n=27) with sheep gait score higher than zero to the case group, and all videos (n=23) with sheep gait score equals to zero were split into the control group. The edited videos contain a sheep walking or running sideways in front of the camera, and the sheep ID on their back are visible and matches with the lameness records.

Geometrical features were extracted from the extracted body parts that represents the head movements/ nodding as well as the back arch, as two indicators of the sheep behaviours.

Three popular machine learning algorithms: logistic regression, support vector machine (SVM), and random forest were built using the back posture features (body contour area, circle radius, curvature). The logistic regression model is used as the baseline model for comparison. To train the machine learning classifiers on the sheep back posture features, 500 image frames were randomly sampled from the 50 short videos (10 frames from each video), the outcomes are either 0 (sound) or 1 (lame) based on their gait score. Then, the Faster-RCNN and DeepLabCut models were jointly used to localized the sheep body key points, we dropped the bounding boxes that have confidence lower than 0.8 were dropped, and the image with low confidence (~0.5) of any key point(n=18). Back posture features were constructed based on 482 images (268 cases, 214 controls). 90% (n=433) images were used for training, and 10%(n=49) images were used for testing. Grid search and random search were used to fine-tune the models. Also, stratified 5-fold cross-validations were used to assess the model performances.

Motion trajectories were also used for lameness detection. 50 videos (cases=27, controls=23) were decomposed into consequent frames, and temporal features were generated by the head-spine relative motions. The logistic regression, SVM and random forest classifiers were also fine-tuned and evaluated, same as before. All models were implemented with the open-sourced python library Scikit-learn 0.24.2.

Experimental Evaluation

To properly evaluate the binary classification results, we calculated accuracy, precision, recall and F1 score for lameness detection. All the three models managed to predict the lameness score in still images with accuracy 0.67, 0.69, and 0.73 with random forest having the best performance. When we applied the models on video segments with temporal features of the head and back, the performance has improved significantly, with a lameness detection accuracy of 0.84 with the random forest classifiers employed. This implies that temporal information is important in lameness detection.

Ethical statement

There is not any new training data created as part of this work and all the datasets used have been collected in previous experiments following appropriate ethical approvals that are discussed in the papers describing the datasets.

References

1. Cox, J. and Bridgers, J., 2021. Why is Animal Welfare Important for Sustainable Consumption and Production?. [ebook] UN Environment. Available at: https://apo.org.au/sites/default/files/resource-files/2019-03/aponid225921.pdf [Accessed 24 August 2021].

2. Veterinary Manual. 2021. Overview of Lameness in Sheep - Musculoskeletal System - Veterinary Manual. [online] Available at: https://www.merckvetmanual.com/musculoskeletal-system/lameness-insheep/overview-of-lameness-in-sheep [Accessed 24 August 2021].

3. Rebelo, C. J. R., McLennan, K. M., Thompson, C., Corke, M. J., ConstantinoCasas, F. 2015. Lameness scoring in sheep. AWIN. University of Cambridge, UK

4. Deeming, L., Beausoleil, N., Stafford, K., Webster, J. and Zobel, G., 2018. Technical note: The development of a reliable 5-point gait scoring system for use in dairy goats. Journal of Dairy Science, 101(5), pp.4491-4497.

5. Wu, D., Wu, Q., Yin, X., Jiang, B., Wang, H., He, D. and Song, H., 2020. Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. Biosystems Engineering, 189, pp.150-163.

6. Feng, Z., Karaskova, M. and Mahmoud, M., 2023, September. Open-Sheep-Face: A Comprehensive Application for Sheep Face Analysis and Pain Estimation. In 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-3). IEEE.

7. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., Poznanski, J., Yu, L., Rai, P., Ferriday, R. and Sullivan, T., 2020. ultralytics/yolov5: v3. 0. Zenodo.

8. H. Yang, R. Zhang, and P. Robinson, 2016. Human and sheep facial landmarks localisation by triplet interpolated features. In Proc. of WACV, pages1–8. IEEE.

9. Chen Chen, Guowu Yuan, Hao Zhou, Yi Ma, 2023 Improved YOLOv5s model for key components detection of power transmission lines[J]. Mathematical Biosciences and Engineering.

10. N. Ruiz, E. Chong, and J. M. Rehg, 2017. Fine-grained head pose estimation without keypoints. CoRR, abs/1710.00925.

11. X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, 2016. Face alignment across large poses: A 3d solution. In Proc. of CVPR, pages 146–155.

12. C. Hewitt and M. Mahmoud, 2019. Pose-Informed Face Alignment for Extreme Head Pose Variations in Animals. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1-6, doi: 10.1109/ACII.2019.8925472.

13. M. A. Fischler and R. C. Bolles, 1987. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In Readings in CV, pages 726–740. Elsevier.

14. Bradski, G. (2000), ’The OpenCV Library’, Dr. Dobb’s Journal of Software Tools.

15. Kazemi, Vahid, and Josephine Sullivan, 2014. One millisecond face alignment with an ensemble of regression trees.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1867-1874.

16. Xiao, JiaShun, 2019. Face Alignment ERT 2D. GitHub repository, Last accessed: 2023-05-03. URL: https://github.com/JiaShun-Xiao/facealignment-ert-2D.

17. Tomasi, Carlo, 2012 Histograms of oriented gradients. Computer Vision Sampler (2012): 1-6.

18. X. Ge et al., 2021. Local Global Relational Network for Facial Action Units Recognition.16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 2021, pp. 01-

08.

19. Hewitt, Charlie, and Marwa Mahmoud, 2019. Pose-informed face alignment for extreme head pose variations in animals.” In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1-6. IEEE.

20. Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, no. 4 (1998): 18-28.

21. Pessanha, Francisca, Krista McLennan, and Marwa Mahmoud, 2020. Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. In 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 387-393. IEEE.

22. The Mathis Lab of Adaptive Motor Control. 2021. DeepLabCut — The Mathis Lab of Adaptive Motor Control. [online] Available at: http://www.mackenziemathislab.org/deeplabcut

Can Artificial Intelligence Contribute to our Conceptual Understanding of Animal Welfare?

Lucy Asher

School of Natural and Environmental Sciences, Newcastle University, Newcastle, UK. **lucy.asher@ncl.ac.uk**

Introduction

Many animal welfare indicators can now be monitored remotely in detail and duration which were previously impossible. A newly arising challenge is how we best use and interpret broader and longer datasets. Artificial intelligence (AI) is a disruptive tool (one that will significantly alter the status quo) that could support meaningful representation and understanding of animal welfare data. Computer systems which can learn, reason, perceive or make decisions would be considered artificially intelligent, particularly if they can adapt and improve performance over time. Subfields of AI include machine learning (systems that learn from data), computer vision (systems which learn from images) and discovery informatics (systems capable of generating new discoveries). AI can find patterns and represent data which is measured in different modalities, over different time scales, and that are time dependent or sequenced. Some of the key advantages of AI based approaches which are already being utilised include, identifying welfare compromise in real time, measuring dynamics of behaviour (onset, offset, duration), and integration across multiple data streams. At present, animal welfare could make huge progress by adopting AI methods which already exist. However, there also exists significant potential for welfare scientists to actively participate and pioneer innovative approaches tailored to the unique challenges of their field. This presents a challenge to the community since animal welfare is not traditionally a computationally intense field.

At the leading edge of artificial intelligence research are systems which can acquire new scientific knowledge and make scientific discoveries autonomously [1]. This means that AI can support more just than speeding up current scientific approaches but also contribute to our conceptual understanding of science. Here, I argue that AI could improve our understanding of animal welfare as a concept and outline some ideas as to how the field of animal welfare could valorise AI. By conceptual understanding I mean knowledge which can be applied in a transferable way, and I consider animal welfare to be a concept, an abstract idea rather than a tangible physical object which can be directly measured or observed. I propose four areas where AI could contribute to our conceptual understanding of animal welfare: simulation of animals; identifying new welfare indicators; discovering new welfare problems or affective states; and defining animal welfare.

AI Based Simulations of Animals

Simulation or agent-based modelling is already a tool which has been widely used to understand complex and emerging behaviour of groups of animals. Such modelling has been used in animal welfare to understand group behaviour e.g. [2–4]. AI can be used to make more accurate and complex representations of animals at group and individual level and therefore provides improved functionality for existing simulations of behaviour. For example, Han and colleagues [5] have applied AI deep learning models to create individualised simulations of cattle based on behaviour which accurately predicted resting state of a cow. Simulated replicas of animals are sometimes referred to as digital twins [6], although the term has more commonly been applied to non-living objects such as buildings. Powered by AI, simulations of animal models of human disease, and of the humans directly, are already being used to replace animals in research [7]. Models can be created at an individual level incorporating genomic and physiological information, symptoms, clinical history and lifestyle to provide personalised medicine, with diagnosis and solutions tailor made to suit that person [8]. For most purposes these models will not necessitate complete simulation of an individual's physiological functioning but instead focus on aspects of relevance for the body system or disease under investigation [9]. However, for scientific purposes it may be useful to create a simulation of the body in increasingly intricate detail. For non-human animals, simulated replicas of an animal's physiology could support an understanding of complex biological interactions which can give rise to apparently contradictory findings in welfare indicators or individual differences in welfare indicators. Digital twins can be made of animal holding facilities and incorporating the animals that are housed there into the model. For example, digital twins have been created of fish farms to enhance decision support for farmers [10].

In parallel to developments in digital twins in other fields, computational neuroscience has emerged as a growing field which utilises AI to model brain processes, cognition and behaviour. Since many welfare indicators are based on behaviour or cognition, there is scope to use AI based computational neuroscience to improve our understanding of these measures. However, there may be one major difference in machine simulations and real animals or humans, emotion. Let's first assume that humans and animals are driven by emotional processes and machines are not, then we need to be cautious in assuming the extent to which machines can replicate real brains and behaviour. In this case that difference between machines and animals might prove a useful tool for animal welfare research; the difference between animals and machine providing insight into the functionality of emotional processes. Such knowledge could help us understand the extent to which different species experience affective states by understanding if behaviour and cognitive outcomes are more similar to non-emotional machine predictions. What happens if we flip the assumption, and instead consider the case that machines can be designed to replicate emotions, to a greater or lesser extent? This possibility has been considered since the advent of artificial intelligence, with section dedicated to this topic in Turing's seminal 1950 [11] paper on Computing Machinery and Intelligence. Interestingly there has recently been a suggestion that a field of AI welfare should arise based on animal welfare science principles [12]. Whilst the capability of machine to experience emotions is open to debate, it is a debate which is entangled with many similar challenges to animal welfare, specifically the extent to which we believe animals are capable of suffering or 'experiencing' and how this can be scientifically studied. Doubtless there are opportunities for the two fields to enrich each other, and much to be gained from comparisons of simulations of animals with outcomes from real animals.

AI Identifying New Welfare Indicators

Let's switch now to perhaps a more solid basis of AI application to animal welfare and focus on animal welfare indicators. One of the most prolific areas for AI application in animal welfare is behaviour recognition (see review by [13]). AI is typically trained to recognise a behaviour, or various behaviours from images or sensor data, by providing the machine with labelled examples of the behaviour. This provides for high throughput of behaviour data and real time recognition of animal welfare problems. This application of AI is solving a practical problem rather than contributing to our conceptual understanding of animal welfare and therefore for the purpose of this paper we need to consider what further insights AI based welfare monitoring could provide. As I have outlined AI is very good at finding patterns and connections and can generate new understanding. Thus, AI could also be used inductively to identify behaviour which occurs in positive or negative welfare contexts which have not previously been noticed by welfare scientists. This suggestion could be considered akin to an AI powered version of classic ethology, collecting data in different situations on indicators at their most primitive level (e.g. in behaviour, movements of each body part of facial feature) to decrease the potential for human bias. Welfare scientists are likely to have missed many potential welfare indicators because of anthropocentric biases, insufficient data quality or quantity, some aspect of time dependency, complexity, or combination of aspects of behaviour which would be difficult for a human observer to identify. It could also be that any changes in behaviour are very subtle or fast which make it difficult to observe. Subtle indicators of affective state are known to exist, for example micro expressions, which are coordinated movements of facial features which last less than 500 milliseconds in duration [14]. These are well characterised in humans and have recently been described in horses [15]. AI has been able to detect differences in micro-expressions between human infants later diagnosed with Autism Spectrum Disorder and those without this diagnosis [16], and although this is not an animal welfare study it demonstrates the potential for AI to detect subtle differences in behaviour between conditions. Inductive methods or data mining methods are very commonly applied to some areas of relevance to animal welfare and AI might be adopted fastest within these areas. Thus, it is likely there will be a proliferation of potential welfare indicators, particularly in the area of omics (e.g. metabolomics, proteomics, genomics) where data mining techniques are already embedded. In my opinion AI is not a silver bullet for identify welfare indicators but could be an exceptional tool for generating hypotheses of novel welfare indicators which can then be carefully validated. As this area of research grows it will be important to synthesise knowledge to identify patterns that are not unique to one species, or differences that occur consistently according to phenotype (e.g. coping style). It is at the synthesis stage of the research that will likely yield the greatest advances to our conceptual understanding of animal welfare as the extent to which knowledge is transferable across contexts, species and individuals, will be characterised.

AI Discovering New Welfare Problems or Affective States

If we accept that AI could support identification of novel welfare indicators, then it follows that it might also be able to identify new welfare problems. By welfare problem, I mean a specific cause of welfare compromise such as feather pecking in laying hens or separation anxiety in dogs. There may be specific causes of animal welfare compromise which are not known within the scientific community, and which can be identified using AI. This could be achieved by identifying consistent changes in animal welfare indicators and then investigating the specific contexts or situations in which these changes are observed. AI would facilitate this through provision of real-time datasets and enhanced ability to identify patterns. More speculatively AI could also support discoveries about affective or emotional states. In animal welfare science, emotional states have been conceptualised as either discrete emotional states (SEEKING, FEAR, RAGE, LUST, CARE, PANIC, and PLAY, [17]), or a continuum about two axes of valence and arousal [18]. If emotions are discrete with distinct neurobiological pathways, as argued by Panksepp [17], then it is possible that there are emotional states which have not yet been identified in animals. Humans are likely to only conceive of affective states humans feel and we assume that more complex emotional states are uniquely human. However, it is feasible that animals could experience emotions that are outside the scope of human experience, or at least are experienced in contexts not conceivable to people. Georgia Mason has previously presented this argument using the thought experiment of the satisfaction that a male lion experiences as it kills the young of a rival male. This is not a context humans experience, yet as a biologically important event for male lions, it is feasible that there might be an affective experience unique to that event. Carrying forward this thought experiment, AI could be a useful tool to detect changes in behaviour and physiology which could indicate an experience, or even emotional state, which humans have not yet perceived. Certainly, detection of emotions which are already characterised will be supported by AI, through integration of multiple data streams [19] and increasing focus on detection of hidden states [20].

AI Defining Animal Welfare

Animal welfare is a concept which means it only exists because we think about it. There are many different definitions of animal welfare, but most are concerned with an animal's suffering or at that point in time that it has a 'life worth living', in the words of Christopher Wathes [21]. These different definitions may reflect different values of what people consider is important for an animal to have a 'good life'. To some extent different definitions can be considered useful as they reflect the values that a group of people have for animals, which may be specific to the context of animal use and reflect cultural or societal norms about the ethical treatment of animals which can evolve over time. However, animal welfare is an established scientific field and therefore a commonality, or at least clarity, in what is being measured is important. If the motivation for animal welfare science is care for the animal, then animal welfare science definitions of animal welfare should be centred around the animal. Since we do not know what it is to 'be' that animal experiencing that state of welfare, we infer this from streams of data on welfare indicators either directly from the animal, or inferred by provisions the animal receives. As welfare scientists we decide which indicators to measure and agonise about how to reconcile and interpret these measures when they do not 'agree', i.e. some indicate good welfare and some bad welfare. The definition of animal welfare used will impact this interpretation, for example higher value of telos (the nature of the animal), or of health would lead to different conclusions about whether free-range or barn systems were better for hen welfare. Attempts to place the animal in the centre of welfare definitions and measurement, have been integrated within welfare science, for example Dawkins' pioneering approaches to let animals choose with their feet [22] and later Nicol, allowing animals to choose preferred environments and measuring welfare indicators which correlate with their chosen environment [23]. Most recently Paul, Nicol, Mendl and others attempted to identify welfare indicators which correlate with both the choices animals make about their preferred environment and their cognitive biases [24]. Unfortunately, these studies did not find clear relationships between welfare indicators, cognitive and decision-based indicators. The point here is that despite being an established field there is no agreed welfare definition or gold standard validation of welfare indicators.

I propose a role for AI within defining animal welfare. By integrating data across multiple studies with multiple welfare indicators, it could be possible for AI to identify commonalities in groupings of welfare indicators, perhaps groupings of individuals within this, which would help us define welfare. AI could thus support welfare definitions which put the animal at the centre because it creates the definition from animal data, rather than from our values-based judgements.

Conclusions

Here I propose that AI could support conceptual advances in animal welfare, primarily by taking an inductive approach allowing AI to make discoveries. Whilst this is a promising area, it is important to recognise the limitations or weaknesses of such an approach. Since the areas outlined in this paper are interlinked, circularity is a key risk. For example, new welfare problems could not be identified using newly identified (unvalidated) welfare indicators. Just with other forms of data analysis, results from AI can only ever be as only as good as data that goes into it, thus all the same principles of good science used to avoid bias need to be considered. Although AI can reduce anthropocentric biases, it is not free from them, since humans programme the machines and select the data which is available for them to learn from and frequently also provide a ground truth which will contain human bias. In the case of AI used in modelling or replication, it is perhaps worthwhile pointing out that the more complex the system being replicated, the more likely it is for a model to accrue errors. Perhaps the biggest concern with AI applied to animal welfare is its application without domain specific knowledge. As this area proliferates there is potential for much of this work to be completed without the knowledge and understanding of animal welfare, and therefore potential for misinterpretation or poor application.

Within AI research, the capacity of machines to make scientific discoveries is a key area of focus [1,25]. In order to build machines to make discoveries autonomously, the principles of how discoveries are made need to be identified and used to programme machines [1]. I doubt that animal welfare discoveries will be top of the list for AI researchers building these discovery machines, but it is interesting to reflect on how ideas and discoveries in our field have arisen and ponder whether these could have been made by a machine.

References

1. Wang H., Fu T., Du Y., Gao W., Huang K., Liu Z., Chandak P., Liu S., Katwyk P. Van, Deac A., Anandkumar A., Bergen K., Gomes C.P., Ho S., Kohli P., Lasenby J., Leskovec J., Liu T.Y., Manrai A., Marks D., Ramsundar B., Song L., Sun J., Tang J., Veličković P., Welling M., Zhang L., Coley C.W., Bengio Y. & Zitnik M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620 (7972), 47–60. doi:10.1038/s41586-023-06221-2.
2. Collins L.M. & Sumpter D.J.T. (2007). The feeding dynamics of broiler chickens. *Journal of the Royal Society Interface*, 4 (12), 65–72. doi:10.1098/rsif.2006.0157.
3. Asher L., Collins L.M., Pfeiffer D.U. & Nicol C.J. (2013). Flocking for food or flockmates? *Applied Animal Behaviour Science,* 147 (1–2), 94–103. doi:10.1016/j.applanim.2013.05.012.
4. Boumans I.J.M.M., Hofstede G.J., Bolhuis J.E., Boer I.J.M. de & Bokkers E.A.M. (2016). Agent-based modelling in applied ethology: An exploratory case study of behavioural dynamics in tail biting in pigs. *Applied Animal Behaviour Science*, 183, 10–18. doi:10.1016/j.applanim.2016.07.011.
5. Han X., Lin Z., Clark C., Vucetic B., Lomax S.L. (2022). AI Based Digital Twin Model for Cattle Caring. doi:10.3390/s22197118.
6. Neethirajan S. & Kemp B. (2021). Digital Twins in Livestock Farming. *Animals*, 11 (4), 1008. doi:10.3390/ani11041008.
7. Chen X., Roberts R., Tong W. & Liu Z. (2022). Tox-GAN: An Artificial Intelligence Approach Alternative to Animal Studies—A Case Study With Toxicogenomics. *Toxicological Sciences*, 186 (2), 242–259. doi:10.1093/toxsci/kfab157.
8. Johnson K.B., Wei W., Weeraratne D., Frisse M.E., Misulis K., Rhee K., Zhao J. & Snowdon J.L. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical Translational Science*, 14 (1), 86–93. doi:10.1111/cts.12884.

9. Kamel Boulos M.N. & Zhang P. (2021). Digital Twins: From Personalised Medicine to Precision Public Health. *Journal of Personalised Medicine*, 11 (8), 745. doi:10.3390/jpm11080745.
10. Ubina N.A., Lan H.Y., Cheng S.C., Chang C.C., Lin S.S., Zhang K.X., Lu H.Y., Cheng C.Y. & Hsieh Y.Z. (2023). Digital twin-based intelligent fish farming with Artificial Intelligence Internet of Things (AIoT). *Smart Agricultural Technology*, 5, 100285. doi:10.1016/j.atech.2023.100285.
11. Turing A.M. (1950). Computing Machinery and Intelligence. In (2009) Parsing the Turing Test, Springer Netherlands, Dordrecht. pp 23–65. doi:10.1007/978-1-4020-6710-5_3.
12. Ziesche S. & Yampolskiy R. (2018). Towards AI Welfare Science and Policies. Big Data and Cognitive Computing, 3 (1), 2. doi:10.3390/bdcc3010002.
13. Jukan A., Masip-Bruin X. & Amla N. (2018). Smart Computing and Sensing Technologies for Animal Welfare. *ACM Computing Surveys*, 50 (1), 1–27. doi:10.1145/3041960.
14. Yan W.J., Wu Q., Liang J., Chen Y.H. & Fu X. (2013). How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions. *Journal of Nonverbal Behaviour*, 37 (4), 217–230. doi:10.1007/s10919-013-0159-8.
15. Tomberg C., Petagna M. & Selliers de Moranville L.A. de (2023). Horses (Equus caballus) facial micro-expressions: insight into discreet social information. *Scientific Reports*, 13 (1), 8625. doi:10.1038/s41598-023-35807-z.
16. Alvari G., Furlanello C. & Venuti P. (2021). Is smiling the key? Machine learning analytics detect subtle patterns in micro-expressions of infants with ASD. *Journal of Clinical Medicine*, 10 (8), 1776. doi:10.3390/JCM10081776/S1.
17. Panksepp J. (2017). Affective Consciousness. In The Blackwell Companion to Consciousness, Wiley. pp 141–156 doi:10.1002/9781119132363.ch10.
18. Mendl M., Burman O.H.P. & Paul E.S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences*, 277 (1696), 2895–2904. doi:10.1098/rspb.2010.0303.
19. Neethirajan S. (2022). Affective State Recognition in Livestock Artificial Intelligence Approaches. *Animals*, 12 (6), 759. doi:10.3390/ani12060759.
20. Chen B., Huang K., Raghupathi S., Chandratreya I., Du Q. & Lipson H. (2021). Discovering State Variables Hidden in Experimental Data. Available at: http://arxiv.org/abs/2112.10755 (accessed on 11 January 2024).
21. Wathes C. (2010). Lives worth living? *Veterinary Record*, 166 (15), 468–469. doi:10.1136/vr.c849.
22. Dawkins M. (1977). Do hens suffer in battery cages? environmental preferences and welfare. *Animal Behaviour*, 25, 1034–1046. doi:10.1016/0003-3472(77)90054-9.
23. Nicol C.J., Caplen G., Edgar J. & Browne W.J. (2009). Associations between welfare indicators and environmental choice in laying hens. *Animal Behaviour*, 78 (2), 413–424. doi:10.1016/j.anbehav.2009.05.016.
24. Paul E.S., Browne W., Mendl M.T., Caplen G., Trevarthen A., Held S. & Nicol C.J. (2022). Assessing animal welfare: a triangulation of preference, judgement bias and other candidate welfare indicators. *Animal Behaviour*, 186, 151–177. doi:10.1016/j.anbehav.2022.02.003.
25. Kitano H. (2016). Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery. *AI Magazine*, 37 (1), 39–49. doi:10.1609/aimag.v37i1.2642.

Automatic pain estimation in equines and canines

Albert Ali Salah

Department of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands; a.a.salah@uu.nl

Abstract

Computer based automation creates new opportunities and challenges for animal wellbeing. While innovative applications are being implemented to assist pet owners, veterinarians, and for monitoring of animals, we also see challenges similar to those in automatic human behavior analysis, where automation raises new ethical issues and risks. We present here the problem of automatic pain estimation of equines and canines using computer vision and machine learning, and discuss the technical challenges of both.

Introduction

While we do not know the subjective experiences of animals, there is sufficient evidence that they occasionally feel unpleasant sensory and emotional experience that correspond to pain, and this sensation is (often) linked to physical damage or problems that need to be addressed for the wellbeing of the animal. Since verbal communication is limited with animals, observation of behaviour related to pain is an important way to estimate whether an animal is in pain or not, and if so, how much pain is there.

While automatic pain assessment in animals is much less explored than pain assessment in humans, in recent years we see that new solutions are enabled by progress in computer vision and machine learning [2]. Most importantly, clinical research into facial expressions of some animals allowed the development of grimace scales and facial action unit systems that allow quantification of pain [1, 6, 11]. Such a quantification is the first step in automated measurements, as machine learning methods require supervision for training. Other behavioural cues, such as the ones recorded in ethograms, pose individual computational problems, and can be attempted with computer vision based approaches.

In this work, we summarize two different approaches for assessing pain expressions that illustrate the challenges and opportunities in this area. The first one deals with equine pain expressions, and works on static images. The facial area of the animal is analysed to arrive at a series of pain scores per area, which are then combined. The most important challenges in this application are the appearance differences between different breeds, the difficulty of transferring knowledge between related species (such as horses and donkeys), the lack of sufficient data with good ground truth labels and annotations. The second approach is used for detecting pain in dogs, and uses dynamic information. A short video clip of the dog is used for assessment. The challenges are once more related to the appearance differences due to different species, but also the large variance of behavioural cues, issues of combining and integrating visual indicators reliably, and lack of data and annotations. We also briefly discuss the ethical issues raised by automatic pain estimation in animals.

Estimation of facial pain in equines from images

The automatic analysis of equine faces for pain estimation is possible. The visible indicators of pain can be extracted from different parts of an equine face. The ears may be turned backwards, there may be an obvious tightening of eyelids, upper eyelids can be visibly angulated, the sclera becomes more prominently visible, the mouth corners and lips are strained, and nostrils are opened. Based on these six indicators (and some additional ones), van Loon et al. proposed the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) for horses suffering from acute colic [11].

Acute pain is manifested less ambiguously compared to chronic pain, and is a good starting point for automatic assessment. In our work, we have used data from 1855 images of horses and 531 images of donkeys, annotated for pain indicators per area of face. We have not collected these images ourselves, and in each case, the data collection was performed after ethical approvals were obtained.

Our analysis approach relies on an accurate pose estimation of the equine face. For this purpose, a multi-loss convolutional neural network approach is used with transfer learning. We use a previously trained model for pose estimation in sheep and adapt it to horses and donkeys. Since frontal, tilted, and profile faces are quite different in terms of feature visibility, we train three different facial landmark estimators, one for each pose. Once the facial landmarks are located, we build area-specific classifiers for each of the six indicators mentioned earlier, using support vector machine classifiers. The limitations of the dataset prevent us from using end-to-end deep learning solutions.

We attempt to overcome the important challenge of data limitation by increasing the amount of labeled data we have synthetically. For this purpose, we take a generic 3D mesh model of a horse, map our 2D images using the located landmarks and a thin-plate spline based deformable model, and obtain a textured mesh. This is then used to create more horse faces, imagined from different angles, using the approach described in [5]. Adding these images to the training set improves the result a bit, but not significantly. We further note that horse and donkey facial morphology is different, and models trained with horses do not perform well on donkeys. These need to be treated as two separate problems, or a significantly larger dataset with both horse and donkey images should be used in training [3]. While there is potentially more information in videos, currently we do not have access to videos of horses or donkeys in pain, and this remains a future work.

Estimation of canine pain from videos

Compared to detecting pain from images, pain detecting from videos can be technically more challenging. The pain indicators may exist in some of the frames, or may require holistic analysis. For instance, the reluctance of a dog to rotate its head may be an indicator of neck pain. A single image will not reveal whether the animal is reluctant or not, but a few minutes of observation of the animal moving about can provide some cues.

To attempt pain estimation from videos, we use a small database of dogs brought to a veterinary clinic, recorded naturally in the observation room by the veterinary clinician, who also provides the pain annotations. Ethical approvals were obtained for the data collection and analysis.

In this case, we do not have area-specific facial pain indicators, but a more general label of "pain" vs. "no-pain". A total of 61 videos, each from an individual, are included. Furthermore, the resolution of the static camera and the distance to the dog, combined with the movements of the dog, make it very difficult to observe the facial details of the animal. Subsequently, we build a system to estimate pain from the analysis of the body motions.

For this task, we start by detecting the dog in the video frames. We use a YOLOv5 deep neural network model [7], pre-trained on the MSCOCO dataset [4], which includes a “dog” class. The advantage of using a pre-trained model is that a much larger training set with great variability is used for that model, which we are able to use instantly. This model gives us a bounding box that roughly locates the dog in the image. We enlarge the bounding box and supply it to a more detailed HRNet model [10], which enables to locate 17 keypoints to represent the body of the dog. Missing keypoints are estimated using symmetry and a few simple rules [12].

Now we have two representations for the behaviour of the dog, one based on cropped images, and one on keypoints. We use a convolutional long-short-term-memory (C-LSTM) for the former, and a vanilla LSTM for the latter. The outputs of these two recognition streams are then combined. Our proposed system thus takes an 8-frame video clip (sampled at a 2 frame per second rate) as input, estimates the dog’s pose, and produces a pain score in the 0-1 range, which can be treated as a probability that the video contains a dog in pain.

Finally, our model is also provides a justification of its score by indicating what precisely in the input resulted in that particular score. For this, we use the Grad-CAM method [8], which creates saliency maps over the input video, showing the important parts that played a role in the decision. When the model gives poor results, we can sometimes spot the reason directly in these saliency maps. For example, sometimes the feet of the dog are occluded by the rest of the body in a particular camera view, and the saliency map just points to a part of the body that stays relatively stable during the video. The resulting label is "no-pain", and it is incorrect, but we understand that the automatic analysis system was not able to capture the feet area, which would have revealed pain symptoms.

The challenges of this application are related to the previous case. We have very limited data with labels; for many deep learning solutions, typically, millions of samples are used. We are able to partly address this by using transfer learning and generic object detection solutions that include the animal class we work with, but these are not sufficient. For the classification task itself, there needs to be more data available for proper training.

Ethical considerations and further directions

Our proposed approached cannot be used as clinical assessment tools for pain estimation without further tests and validation. No animals were harmed in this research, there was no induced pain for the subjects, and ethics committee approvals were obtained for the preparation of the datasets. We discuss a few further points here.

The automatic observation of behavioural cues for animal pain assessment is error-prone, as the animal may not show certain indicative behaviours under certain circumstances. The behaviour is often different depending on the presence of familiar/unfamiliar people or animals around. The owner or the veterinarian may have deeper insights into the behaviour. What then is the main purpose of automatic analysis?

First of all computer analysis sometimes can surpass human analysis in its detail and accuracy. A high-rate camera can acquire images at a greater frame rate than the human eye, an algorithm can simultaneously track the movement, velocity, and even acceleration of many points on the animal. Subsequently, an automatic algorithm can provide very good feedback for the pet owner or the veterinary expert, including an explanation of what seems to be problematic. It can also provide early warning, and can help in training professionals.

Under excellent recording conditions, computer analysis may work very well. On the other hand, such algorithms do depend on illumination and other recording conditions. Their expected accuracy may vary greatly under poorer conditions. The tools we develop should be rigorously tested before any serious application is built on them. Similarly, proper annotations should come from animal ethologists, and the automatic systems should be built on rigorously collected datasets, with both recording and environmental conditions clearly depicted [9].

The limited amount of data available for analysis is the biggest challenge. In human data collection, one approach was to introduce many illumination and recording elements and to acquire thousands of images with varying conditions under a very short time. This helped model appearance variances, while minimizing the labeling effort. Another promising solution is to convert these applications into mobile applications and deploy them to get more data from users, thereby establishing a positive feedback loop. More "in-the-wild" data from the users will provide rich variance. Yet, the limitations of such applications need to be very clear to the users. What is the cost of a false positive (i.e. an animal incorrectly labeled as "in pain") and what is the cost of a false negative (i.e. missing that the animal is in pain)? Obviously, the latter is a much more important problem for the animal, but for usability of any application, the false positives cannot be ignored. Just like the child who cried wolf too many times, many false positives may also lead to the users ignoring pain signs when there is eventually a true positive. The usefulness of these applications will be maximized if they can provide explanations for their automatic decisions, telling the users the presence of this and that indicator that hints at the possibility of pain, and prompting professional consultation if necessary.

A more general problem with automatic, AI-based analysis of animals is that it may prompt human owners to rely too much on such applications and to neglect the wellfare of the animals. They can be left alone for longer periods, under AI surveillance, or crammed more densely into smaller areas. We should always ask the broader questions related to animal wellfare, when we work on these technologies. Finally, the AI Act and many AI regulations introduced by different countries have a distinct focus on human wellfare, but there is almost no mention of animal wellfare. We should also make sure that AI regulations encompass the welfare of animals affected by these technologies.

References

1. Andersen, P.H., Broomé, S., Rashid, M., Lundblad, J., Ask, K., Li, Z., Hernlund, E., Rhodin, M., Kjellström, H. (2021). Towards machine recognition of facial expressions of pain in horses. Animals 11(6), 1643

2. Broome, S., Feighelstein, M., Zamansky, A., Carreira Lencioni, G., Haubro Andersen, P., Pessanha, F., ... & Salah, A. A. (2023). Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. *International Journal of Computer Vision*, 131(2), 572-590.

3. Hummel, H. I., Pessanha, F., Salah, A. A., van Loon, T. J., & Veltkamp, R. C. (2020, November). Automatic pain detection on horse and donkey faces. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 793-800). IEEE.

4. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context," in *Proc. ECCV*. Springer, 740–755.

5. Pessanha, F., Salah, A. A., van Loon, T., & Veltkamp, R. (2022). Facial image-based automatic assessment of equine pain. IEEE Transactions on Affective Computing, 14(3), 2064-2076.

6. Rashid, M., Silventoinen, A., Gleerup, K. B., & Andersen, P. H. (2019). Analyzing horse facial expressions of pain with EquiFACS. *Pain in Animals Workshop*, Bethesda Maryland United States.

7. Redmon, J. & Farhadi, A. (2017). Yolo9000: better, faster, stronger, in *Proc. CVPR*, 7263–7271.

8. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & D. Batra, (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. CVPR*, 618–626.

9. Siegford, J. M., Steibel, J. P., Han, J., Benjamin, M., Brown-Brandl, T., Dórea, J. R., ... & Rosa, G. J. (2023). The quest to develop automated systems for monitoring animal behavior. *Applied Animal Behaviour Science,* 265, 106000.

10. Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation, in *Proc. CVPR,* 5693–5703.

11. van Loon, J. P., & Van Dierendonck, M. C. (2015). Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): a scale-construction study. *The Veterinary Journal*, *206*(3), 356-364.

12. Zhu, H., Salgırlı, Y., Can, P., Atılgan, D., & Salah, A. A. (2023). Video-based estimation of pain indicators in dogs. *Proc. Affective Computing and Intelligent Interaction (ACII).*

Symposium: Methods for the study of olfactory learning and memory

An Overview of Methods for Measuring Olfaction in Rodents

Kyle M Roddick[1] Richard E Brown[1]

[1]Department of Psychology and Neuroscience, Dalhousie University, Halifax, Canada. kyle.roddick@dal.ca

Abstract

Rodents are heavily reliant on olfactory stimuli, showing a remarkable ability to learn olfactory tasks. With a preference for olfactory stimuli, and the oversized role olfaction plays in their behaviour, it is important to consider olfaction when working with rodents. There are many ways to measure olfaction in rodents, such as tests of odour detection, odour discrimination, working memory, reversal learning, complex olfactory based tasks, and pheromonal effects on reproductive behaviour.

Why olfaction?

Rodents rely on olfactory cues for many behaviours, from sexual behaviours [1] to predator avoidance [2], and show a remarkable ability to learn olfactory based tasks. Rodents are able to detect odours at very low concentrations [3,4] and are able to learn complex olfactory based tasks, with both rats and mice showing learning to learn [5,6], and rats showing a preference for olfactory stimuli over auditory or visual stimuli [7].

Deficits in both odour detection and identification are observed in age-related neurodegenerative disorders [8,9], with olfactory dysfunction a better predictor of Alzheimer's Disease (AD) onset than loss of verbal episodic memory [10]. A meta-analysis of studies examining olfactory deficits in AD and Parkinson's Disease (PD) patients found that olfactory deficits were more severe in AD than PD, with the deficits detected in AD being primarily deficits of odour identification and recognition [11]. However, although olfactory deficits were overall less apparent in PD, deficits in olfactory sensitivity were more apparent in PD than AD, suggesting that while olfactory deficits may be a symptom common to neurodegenerative diseases, it may be possible to differentiate between the different diseases based on the nature of the olfactory deficits.

The rodent olfactory system is a complex network consisting of the main and accessory olfactory pathways. The traditional view has been that the main olfactory system was responsible for detecting and responding to environmental odours, with the accessory olfactory system responsible for responses to pheromones. However, evidence shows that this strict division of the olfactory system is misleading, with much overlap between the systems, and both playing roles in the detection of odourants and pheromones [12].

The main olfactory system starts with the olfactory epithelium in which olfactory receptor neurons express one of the approximately 1500 olfactory receptors encoded in the mouse genome, compared to the approximately 900 found in the human genome [13]. The axons of the olfactory receptor neurons project to the olfactory bulb, where neurons expressing the same olfactory receptors converge on the same glomeruli [14].

The projections of the olfactory bulb form the lateral olfactory tract, which enervates the olfactory cortex, which includes the anterior olfactory nucleus, the olfactory tubercle, the piriform cortex, the amygdala, and the entorhinal cortex [15]. There are intrinsic connections within and between the structures of the olfactory cortex, and extrinsic outputs to neocortical and subcortical regions including the insular, orbital, and perirhinal cortices, and the hippocampus, thalamus, and hypothalamus [16].

The accessory olfactory system starts with the vomeronasal organ, located within the nasal septum at the base of the olfactory cavity. The receptor neurons of the vomeronasal organ project to the accessory olfactory bulb, which is embedded in the dorsal section of the olfactory bulb [17]. The main projections of the accessory olfactory bulb are the amygdala, hypothalamus, and the anterior olfactory nucleus [15,16].

The many transgenic rodent models of neurological conditions, including nearly 200 different transgenic mouse models just of AD alone [18], combined with rodents' affinity for performing olfactory based tasks, makes the use of olfactory tasks an ideal way to study the neurobehavioral effects of these genetic manipulations, and highlights the need to understand the function of the rodent olfactory system. Tasks used to analyze olfactory abilities of these mice include tests of odour detection, odour discrimination, reversal learning and working memory, complex odour learning, and learning to discriminate between conspecific odours.

Tests of odour detection

Some of the simplest tasks are tests of odour detection. The buried food test relies on the animals natural foraging behaviour. Animals are fasted overnight and then introduced to a cage containing buried pieces of a familiar food that has an odour. The number of pieces of food the animal finds within a given time is scored [19].

Olfactory detection can be demonstrated at an early age in rodents. Olfactory cues are essential for pups to find their mother's nipples and suckle [20]. Preweaning, 18 day old rat pups show an aversion to peppermint scent and an attraction to their maternal odour [21]. As early as two days of age, mouse pups show a decrease in isolation induced ultrasonic vocalizations in response to the odour of adult male mice [22].

Olfactory sensitivity can be measured using operant olfactometers. Using a modification of a two odour discrimination task, the rewarded stimulus is the odour for which sensitivity is being assessed, while the unrewarded stimulus is an odourless substance or clean air. Organic compounds are typically used in sensitivity tasks, such as ethyl acetate [3,4,23], octyl aldehyde [24], or n-hexanal [25]. These tasks assess the detection threshold of the animal by initially training the animals to perform with an easily detectable concentration of the odourant, and then decreasing the concentration until the animal is unable to detect it and performance drops to chance levels.

Tests for simple odour discrimination learning and memory

Discrimination learning involves presenting rodents with two or more olfactory stimuli and testing their ability to discriminate between the stimuli. One common method of assessing olfactory discrimination is the habituation/dishabituation test. This test involves presenting the animal with one odour, typically on a cotton applicator, and recording the time the animal spends investigating the odour. Across subsequent presentations of the same odour the time spent investigating the odour should decrease (habituation). If a novel odour is then presented, the time the animal investigates the new odour should increase (dishabituation). This dishabituation demonstrates that the animal can discriminate between the first and second odours [26].

A simple test for odour discrimination learning and memory involves the animals learning to associate an odourant in a pot of bedding with a sugar reward buried in the bedding, while a second odourant is presented in a pot of bedding without a sugar reward across a number of training trials [27]. The ability of the animals to discriminate between the odours is assessed in a memory test by presenting them with both odourants simultaneously and measuring which odourant pot they dig in. While the original task involved four days of training, consisting of eight training trials per day [27], subsequent studies have shown that one day of eight training trials is sufficient to train the mice, but one day of four training trials is not [28].

Odour discrimination ability can also be assessed using operant olfactometers [24]. Go/no-go two odour discrimination tasks involve presenting the animal with either a rewarded odour (S+) or unrewarded odour (S-) when it inserts its nose into the odour sampling port. If the animal can perform the task, it will learn to lick the reward spout in response to the S+ and to inhibit licking to the S-. Animals are typically tested on such a task until they reach an accuracy criterion, e.g. 85%, or for a set number of trials, e.g. 200 trials.

Tests for working memory

Working memory tasks involve requiring the animal to store a memory of one or more stimuli and use that memory to determine how to respond at a later timepoint. The three main classifications of working memory tasks used with animals are i) goal maintenance, ii) memory capacity, and iii) interference control [29]. Goal maintenance tasks include delayed matching and non-matching to sample tasks, as well as delayed win-shift tasks. In these types of tasks, the animal is presented with a stimulus and must remember it over a delay period before making a response that depends upon that stimulus.

Memory capacity tests involve the animal retaining a memory of multiple stimuli and using this memory to determine how to respond. For example, in the olfactory working memory capacity task [30,31] the animal is presented with a variable number of odourant pots before being presented with the same odourant pots plus a novel odourant pot and must dig in the novel pot to get a reward. This nonmatching to sample task controls the demand on the animals' working memory by varying the number of odours initially presented.

Similarly, the odour span task involves first presenting the animal with one odourant pot containing a reward. After the animal has found and consumed the reward it is removed from the testing arena, the now unrewarded odourant pot is moved to a random position in the arena and a novel, rewarded, odourant pot is added. The animal is then returned to the arena, and should it successfully first dig in the novel odourant pot this process of removing the animal, adding a novel, rewarded, odourant pot, and randomly positioning the pots is continued until the animal first goes to an incorrect pot. The number of odours added before the animal makes an error and digs in a non-novel odourant pot gives a measure of their working memory capacity [32].

In the operant olfactometer both rats [33] and mice [34] are able to perform a go/no-go delayed matching to sample task which involves presenting the animal with one odour and following an inter-stimulus delay, a second odour. If the odours are the same the animal will be rewarded for licking, and if the two odours are different the animal will not be rewarded for licking. The demand on the working memory of the animals is controlled by varying the delay between the two odours.

Tests for reversal learning

Reversal learning can be assesed in a variety of olfactory based learning and memory tasks [35] and have been used as a measure of cognitive flexibility [36]. Olfactory reversal tasks typically involve training animals to discriminate between a rewarded stimulus (S+) and an unrewarded stimulus (S-), and once they have learned the task, the stimuli are reversed such that the rewarded stimulus becomes the S- and the unrewarded stimulus becomes the S+ [37].

Serial olfactory reversal learning was shown in mice that were trained on four pairs of odor discriminations, each followed by a reversal of that pair [38]. These mice showed a decrease in errors across both discrimination and reversal pairs, with more errors in reversal learning than in the original odor discriminations. When mice were presented with 18 odour pairs each followed by a reversal [6], and rats were presented with 16 odour pairs followed by reversals [5], both were able to achieve near errorless performance, showing evidence of learning to learn, a behaviour previously thought to be limited to primates.

Tests for complex olfactory learning

In a test of sequential order olfactory memory, rats were presented with a series of five rewarded odourant pots, and then given a two choice test where the odourant that appeared earlier in the series was rewarded [39]. The rats were able to successfully perform this task. In a What-Where-When task, rats were presented with a series of four rewarded odour pots one at a time, each in a unique position within the test arena. Memory was tested by presenting them with a random pair of the odours, in their original positions, with the odour that had been presented earlier in the series rewarded [40]. In this hippocampal-dependent task, probe trials which require the rat to rely solely on olfactory or spatial information to make their decision reveal that the rats appear to initially go to the position they think was earliest and confirm with the odour once they get there.

In a transitive inference task, rats were presented with a hierarchical series of two odour problem (A > B, B > C, C > D, D > E) and tested if they could infer indirect relationships (B > D) [41]. Rats were able to perform this task well above chance, but not if they had hippocampal lesions. In the olfactory tubing maze [42], mice move through the apparatus and are presented with junctions where they must choose to go left or right. The tubes for both choices have odourised air pumped into them, one associated with a water reward, and the other with a buzzer. Mice can learn this task, showing increasing accuracy and speed.

Pheromones

Pheromones, acting through the accessory olfactory system, can have profound effects on the reproductive behaviour of rodents. Exposing juvenile female mice to the odour of a male causes them to undergo early puberty [43]. The accessory olfactory system is involved in this acceleration of puberty, as juvenile female mice with their vomeronasal organ removed do not exhibit advanced onset of first oestrus or increased uterine weights [44]. Exposing juvenile female mice to odours from novel males causes the activation of immediate early genes in their accessory olfactory bulb [45]. Estrus can also be induced in female mice exposed to novel male odours [46,47]. In each case the female mouse must be exposed to the odour from novel males, so the neuroendocrine system distinguishes between the odours of novel and familiar males [48].

Recently mated female mice exposure to a novel male will terminate their pregnancy, importantly, this pregnancy block does not occur with re-exposure to the original sire male [49]. Exposure to just the urine of a novel male is sufficient to cause the pregnancy block [50], however, removal of the vomeronasal organ prevents this effect [51]. Importantly, this indicates that there is a mechanism by which the female is able to differentiate the odours of the novel male from that of her mate.

Summary

Rodents show a remarkable ability to learn olfactory based tasks, including odour detection, odour discrimination, working memory, reversal learning, and complex tasks such as sequential odour and transitive inference tests. Olfaction also has important effects on their more naturalistic behaviours, such as reproduction. Together, these highlight the importance of considering the role of olfaction in experiments with rodents.

References

1. Brown, R.E. (1979) Mammalian Social Odors: A Critical Review. In: Rosenblatt JS, Hinde RA, Beer C, and Busnel M-C, editors. *Advances in the Study of Behavior*, Academic Press. p. 103–62. https://doi.org/10.1016/S0065-3454(08)60094-7

2. Takahashi, L. (2014) Olfactory systems and neural circuits that modulate predator odor fear. *Frontiers in Behavioral Neuroscience*, **8**.

3. Roddick, K.M., Fertan, E., Schellinck, H.M., Brown, R.E. (2022) A Signal Detection Analysis of Olfactory Learning in 12-Month-Old 5xFAD Mice. *Journal of Alzheimer's Disease*, IOS Press. **88**, 37–44. https://doi.org/10.3233/JAD-220049

4. Roddick, K.M., Roberts, A.D., Schellinck, H.M., Brown, R.E. (2016). Sex and Genotype Differences in Odor Detection in the 3×Tg-AD and 5XFAD Mouse Models of Alzheimer's Disease at 6 Months of Age. *Chemical Senses*, **41**, 433–40. https://doi.org/10.1093/chemse/bjw018

5. Slotnick, B.M., Katz, H.M. (1974). Olfactory learning-set formation in rats. *Science*, **185**, 796–8. https://doi.org/10.1126/science.185.4153.796

6. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2023). Serial reversal learning in an olfactory discrimination task in 3xTg-AD mice. *Learning & Memory*, **30**, 310–9. https://doi.org/10.1101/lm.053840.123

7. Nigrosh, B.J., Slotnick, B.M., Nevin, J.A. (1975). Olfactory discrimination, reversal learning, and stimulus control in rats. *Journal of Comparative and Physiological Psychology*, American Psychological Association, US. **89**, 285–94. https://doi.org/10.1037/h0076821

8. Alves, J., Petrosyan, A., Magalhães, R. (2014). Olfactory dysfunction in dementia. *World Journal of Clinical Cases : WJCC*, **2**, 661–7. https://doi.org/10.12998/wjcc.v2.i11.661

9. Doty, R.L. (2017). Olfactory dysfunction in neurodegenerative diseases: is there a common pathological substrate? *The Lancet Neurology*, **16**, 478–88. https://doi.org/10.1016/S1474-4422(17)30123-0

10. Devanand, D.P., Lee, S., Manly, J., Andrews, H., Schupf, N., Doty, R.L. et al. (2015). Olfactory deficits predict cognitive decline and Alzheimer dementia in an urban community. *Neurology*, **84**, 182–9. https://doi.org/10.1212/WNL.0000000000001132
11. Rahayel, S., Frasnelli, J., Joubert, S. (2012). The effect of Alzheimer's disease and Parkinson's disease on olfaction: A meta-analysis. *Behavioural Brain Research*, **231**, 60–74. https://doi.org/10.1016/j.bbr.2012.02.047
12. Restrepo, D., Arellano, J., Oliva, A.M., Schaefer, M.L., Lin, W. (2004). Emerging views on the distinct but related roles of the main and accessory olfactory systems in responsiveness to chemosensory signals in mice. *Hormones and Behavior*, **46**, 247–56. https://doi.org/10.1016/j.yhbeh.2004.02.009
13. Young, J.M., Friedman, C., Williams, E.M., Ross, J.A., Tonnes-Priddy, L., Trask, B.J. (2002). Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Human Molecular Genetics*, **11**, 535–46. https://doi.org/10.1093/hmg/11.5.535
14. Mombaerts, P. (2006). Axonal wiring in the mouse olfactory system. *Annual Review of Cell and Developmental Biology*, **22**, 713–37. https://doi.org/10.1146/annurev.cellbio.21.012804.093915
15. de Castro, F. (2009). Wiring olfaction: the cellular and molecular mechanisms that guide the development of synaptic connections from the nose to the cortex. *Frontiers in Neuroscience*, **3**.
16. Ennis, M., Puche, A.C., Holy, T., Shipley, M.T. (2015). Chapter 27 - The Olfactory System. In: Paxinos G, editor. *The Rat Nervous System (Fourth Edition)*, Academic Press, San Diego. p. 761–803. https://doi.org/10.1016/B978-0-12-374245-2.00027-9
17. Mucignat-Caretta, C. (2010). The rodent accessory olfactory system. *Journal of Comparative Physiology A*, **196**, 767–77. https://doi.org/10.1007/s00359-010-0555-z
18. Myers, A., McGonigle, P. (2019). Overview of Transgenic Mouse Models for Alzheimer's Disease. *Current Protocols in Neuroscience*, **89**, e81. https://doi.org/10.1002/cpns.81
19. Yang, M., Crawley, J.N. (2009). Simple Behavioral Assessment of Mouse Olfaction. *Current Protocols in Neuroscience*, **48**, 8.24.1-8.24.12. https://doi.org/10.1002/0471142301.ns0824s48
20. Hongo, T., Hakuba, A., Shiota, K., Naruse, I. (2000). Suckling Dysfunction Caused by Defects in the Olfactory System in Genetic Arhinencephaly Mice. *Biology of the Neonate*, **78**, 293–9. https://doi.org/10.1159/000014282
21. Brown, R.E., Willner, J.A. (1983). Establishing an "affective scale" for odor preferences of infant rats. *Behavioral and Neural Biology*, **38**, 251–60. https://doi.org/10.1016/S0163-1047(83)90254-6
22. Lemasson, M., Delbé, C., Gheusi, G., Vincent, J.-D., Lledo, P.-M. (2005). Use of ultrasonic vocalizations to assess olfactory detection in mouse pups treated with 3-methylindole. *Behavioural Processes*, **68**, 13–23. https://doi.org/10.1016/j.beproc.2004.09.001
23. Bodyak, N., Slotnick, B.M. (1999). Performance of Mice in an Automated Olfactometer: Odor Detection, Discrimination and Odor Memory. *Chemical Senses*, **24**, 637–45. https://doi.org/10.1093/chemse/24.6.637
24. Slotnick, B.M., Restrepo, D. (2005). Olfactometry with Mice. *Current Protocols in Neuroscience*, **33**, 8.20.1-8.20.24. https://doi.org/10.1002/0471142301.ns0820s33
25. Phillips, M., Boman, E., Österman, H., Willhite, D., Laska, M. (2011). Olfactory and Visuospatial Learning and Memory Performance in Two Strains of Alzheimer's Disease Model Mice—A Longitudinal Study. *PLOS ONE*, Public Library of Science. **6**, e19567. https://doi.org/10.1371/journal.pone.0019567
26. Brown, R.E., Singh, P.B., Roser, B. (1987). The Major Histocompatibility Complex and the chemosensory recognition of individuality in rats. *Physiology & Behavior*, **40**, 65–73. https://doi.org/10.1016/0031-9384(87)90186-7
27. Schellinck, H.M., Forestell, C.A., LoLordo, V.M. (2001). A Simple and Reliable Test of Olfactory Learning and Memory in Mice. *Chemical Senses*, Oxford Academic. **26**, 663–72. https://doi.org/10.1093/chemse/26.6.663
28. Brown, R.E., Schnare, O.K., Habib, E.B., Roddick, K.M. (2023). Development of a One-Day Test of Olfactory Learning and Memory in Mice. In: Schaal B, Keller M, Rekow D, and Damon F, editors. *Chemical Signals in Vertebrates 15*, Springer International Publishing, Cham. p. 39–53. https://doi.org/10.1007/978-3-031-35159-4_3
29. Dudchenko, P.A., Talpos, J., Young, J., Baxter, M.G. (2013). Animal models of working memory: A review of tasks that might be used in screening drug treatments for the memory impairments found in schizophrenia. *Neuroscience & Biobehavioral Reviews*, **37**, 2111–24. https://doi.org/10.1016/j.neubiorev.2012.03.003

30. Huang, G.-D., Jiang, L.-X., Su, F., Wang, H.-L., Zhang, C., Yu, X. (2020). A novel paradigm for assessing olfactory working memory capacity in mice. *Translational Psychiatry*, Nature Publishing Group. **10**, 1–16. https://doi.org/10.1038/s41398-020-01120-w

31. Jiang, L.-X., Huang, G.-D., Wang, H.-L., Zhang, C., Yu, X. (2022). The protocol for assessing olfactory working memory capacity in mice. *Brain and Behavior*, **n/a**, e2703. https://doi.org/10.1002/brb3.2703

32. Young, J.W., Kerr, L.E., Kelly, J.S., Marston, H.M., Spratt, C., Finlayson, K. et al. (2007). The odour span task: A novel paradigm for assessing working memory in mice. *Neuropharmacology*, **52**, 634–45. https://doi.org/10.1016/j.neuropharm.2006.09.006

33. Lu, X.-C.M., Slotnick, B.M., Silberberg, A.M. (1993). Odor matching and odor memory in the rat. *Physiology & Behavior*, **53**, 795–804. https://doi.org/10.1016/0031-9384(93)90191-H

34. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2014). Olfactory delayed matching to sample performance in mice: Sex differences in the 5XFAD mouse model of Alzheimer's disease. *Behavioural Brain Research*, **270**, 165–70. https://doi.org/10.1016/j.bbr.2014.04.038

35. Mihalick, S.M., Langlois, J.C., Krienke, J.D., Dube, W.V. (2000). An Olfactory Discrimination Procedure for Mice. *Journal of the Experimental Analysis of Behavior*, **73**, 305–18. https://doi.org/10.1901/jeab.2000.73-305

36. Kesner, R.P., Churchwell, J.C. (2011). An analysis of rat prefrontal cortex in mediating executive function. *Neurobiology of Learning and Memory*, **96**, 417–31. https://doi.org/10.1016/j.nlm.2011.07.002

37. Johnson, C., Wilbrecht, L. (2011). Juvenile mice show greater flexibility in multiple choice reversal learning than adults. *Developmental Cognitive Neuroscience*, **1**, 540–51. https://doi.org/10.1016/j.dcn.2011.05.008

38. Caglayan, A., Stumpenhorst, K., Winter, Y. (2021). Learning Set Formation and Reversal Learning in Mice During High-Throughput Home-Cage-Based Olfactory Discrimination. *Frontiers in Behavioral Neuroscience*, **15**. https://doi.org/10.3389/fnbeh.2021.684936

39. Fortin, N.J., Agster, K.L., Eichenbaum, H.B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, Nature Publishing Group. **5**, 458–62. https://doi.org/10.1038/nn834

40. Ergorul, C., Eichenbaum, H. (2004). The Hippocampus and Memory for "What," "Where," and "When." *Learning & Memory*, **11**, 397–405. https://doi.org/10.1101/lm.73304

41. Dusek, J.A., Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences. **94**, 7109–14. https://doi.org/10.1073/pnas.94.13.7109

42. Roman, F.S., Marchetti, E., Bouquerel, A., Soumireu-Mourat, B. (2002). The olfactory tubing maze: a new apparatus for studying learning and memory processes in mice. *Journal of Neuroscience Methods*, **117**, 173–81. https://doi.org/10.1016/S0165-0270(02)00094-8

43. Vandenbergh, J.G. (1969). Male Odor Accelerates Female Sexual Maturation in Mice. *Endocrinology*, **84**, 658–60. https://doi.org/10.1210/endo-84-3-658

44. Lomas, D.E., Keverne, E.B. (1982). Role of the vomeronasal organ and prolactin in the acceleration of puberty in female mice. *Reproduction*, Bioscientifica Ltd. **66**, 101–7. https://doi.org/10.1530/jrf.0.0660101

45. Schellinck, H.M., Smyth, C., Brown, R., Wilkinson, M. (1993). Odor-induced sexual maturation and expression of c-fos in the olfactory system of juvenile female mice. *Developmental Brain Research*, **74**, 138–41. https://doi.org/10.1016/0165-3806(93)90094-Q

46. Whitten, W.K. (1956). Modification of the oestrous cycle of the mouse by external stimuli associated with the male. *Journal of Endocrinology*, Bioscientifica Ltd. **13**, 399–404. https://doi.org/10.1677/joe.0.0130399

47. Wölfl, S., Zala, S.M., Penn, D.J. (2023). Male scent but not courtship vocalizations induce estrus in wild female house mice. *Physiology & Behavior*, **259**, 114053. https://doi.org/10.1016/j.physbeh.2022.114053

48. Lendrem, D.W. (1985). Kinship Affects Puberty Acceleration in Mice (Mus musculus). *Behavioral Ecology and Sociobiology*, Springer. **17**, 397–9.

49. Bruce, H.M. (1959). An Exteroceptive Block to Pregnancy in the Mouse. *Nature*, Nature Publishing Group. **184**, 105–105. https://doi.org/10.1038/184105a0

50. Rosser, A.E., Remfry, C.J., Keverne, E.B. (1989). Restricted exposure of mice to primer pheromones coincident with prolactin surges blocks pregnancy by changing hypothalamic dopamine release. *Reproduction*, Bioscientifica Ltd. **87**, 553–9. https://doi.org/10.1530/jrf.0.0870553

51. Bellringer, J.F., Pratt, H.P.M., Keverne, E.B. (1980). Involvement of the vomeronasal organ and prolactin in pheromonal induction of delayed implantation in mice. *Reproduction*, Bioscientifica Ltd. **59**, 223–8. https://doi.org/10.1530/jrf.0.0590223

Pavlovian Conditioned Odour Preferences in Mice

Richard E. Brown, Elias B. Habib, Oliver K. Schnare, and Kyle M. Roddick

Department of Psychology and Neuroscience, Dalhousie University, Halifax, Canada, rebrown@dal.ca

Abstract

A series of experiments used Pavlovian conditioning to train mice to associate a CS+ odour with sugar and a CS- odour with no sugar. Memory strength was examined by varying the number of training trials and by increasing the time between training and memory tests from 1 to 30 days and longer. The results from four different training paradigms are compared. Studies of transfer of training and neural changes following conditioning are also reported.

The Pavlovian conditioned odour preference task

Rodents rely heavily on their sense of olfaction [1], and are capable of preforming surprisingly complex olfactory learning and memory tasks [2,3] at high levels of performance [4,5]. Thus, the use of olfactory tasks provides ethologically relevant methods for examining the cognitive processes of rodents. Schellinck et al [6] developed a simple Pavlovian conditioned odour preference task to evaluate learning and memory in mice that relies on their natural foraging behaviour. During training, mice were presented with odourant pots containing sawdust with one odour (CS+) paired with sugar, and odour pots containing another odour (CS-) without sugar, in separate cages, over four days. The mice were then tested 24 hours later for their memory by placing them in a three-chamber apparatus containing a CS+ odour pot at one end and a CS- odour pot at the other, but without sugar in either. Learning and memory were determined by measuring the amount of time spent digging in the CS+ odourant pot relative to the time digging in the CS- odourant pot. Mice did not need to be food restricted to learn the association between the CS+ odour during the training phase, but did need to be food restricted to express this association during memory testing [7].

Short, long, and very long term memory

Mice demonstrated their ability to remember the odour pairings when tested 1, 14, 30, or 60 days after the end of training [6]. Wong and Brown [8] showed that this memory persisted for 3, 6, and 9 months after training. Thus, the conditioned odour preference task can be used to investigate short, long, and very long-term memory in mice.

Sex and Genotype

Mice of various genotypes can complete this Pavlovian conditioned odour task. Schellinck et al [6] tested CD1, DBA, and C57 mice, all of whom were able to learn the task. Brown and Wong [9] tested 12 different strains of mice of various visual abilities, and all learned the task. Male and female mice learn the task equally well.

Age

Mice at 12, 18, and 24 months of age are able to learn the Pavlovian conditioned odour task, and this was not affected by visual ability [10]. An Alzheimer's disease model, the 5xFAD mouse, showed no impairment in this task when tested from 3 to 15 months of age [11].

A one-day test

Odour preferences could be conditioned with either eight, four, two, or one training trials per day over four days (distributed practice) [6]. Recently, we developed methods for training odour preference over a single day (massed practice) [1]. We showed that eight training trials with each of the rewarded (+) and non-rewarded (-) conditioned stimulus (CS) odours (CS+ and CS-) over one day is sufficient to create a memory that lasts 1, and 7 days, but not 30 days, but four training trials with each odour over one day did not create a significant memory even 1 day later.

Memory strength

The results of these one-day conditioned odour preference tests suggests that the strength of the memory for the association of sugar with the CS+ odour can be modified by altering the number of training trials the mice receive. This is important as even Alzheimer's disease model mice up to 15 months old are able to perform the task at the same level as their wildtype controls [11]. Memory strength is measured by the duration of preference for the CS+, from one to 30 days. In order to examine procedural factors affecting memory strength we examined procedural variations in the Pavlovian conditioned odour preference task.

Procedural variations

We have tested four different procedures of this task in our lab, the procedure originally described with eight 10 minute exposures to each odour per day for four days [6], a one day training procedure using eight five minute exposures to each odour in one day [1], a one day procedure using only a CS+ odour (and no CS- odour), with eight five minute exposures to the CS+ over one day [12], and a one day procedure that exposed the mice to both the CS+ and CS- simultaneously during training, with eight ten minute exposures to each odour in one day [13]. We have also tested extinction and spontaneous recovery by presenting the mice with four more days of testing following 24 hour and seven day memory tests. During all the memory tests the CS+ was presented without sugar. Seven days after the last extinction test the mice were again given a memory test to determine if they would show spontaneous recovery. All animal procedures were approved by Dalhousie's University Committee on Laboratory Animals, in accordance with Canadian Council on Animal Care guidelines.

Results

Mice given four CS+ and four CS- trials per day for four days learn to associate the CS+ odour with sugar reward, as do mice given three, two, or one trial of each stimuli per day for four days [6]. While food restricting the mice is not necessary for them to acquire the association during training, it is needed for them to express this learning during the memory tests [7]. Using two CS+ and two CS- trials per day for four days results in memories that can be assessed at least nine months later [8], and mice from 3 to 15 months old can learn the conditioned odour preference task [11]. One day of training, consisting of eight CS+ and eight CS- trials, is sufficient for mice to learn the CS+ sugar association and remember this association for 24 hours and 7 days, but not for 30 days, while four CS+ and four CS- trials over one day is insufficient to establish a 24 hr memory [1]. When given four days of extinction tests following 24 hour and seven day memory tests, the mice showed less time digging in the CS+, but did show spontaneous recovery when tested seven days after extinction [14]. When trained with simultaneous exposure to the CS+ and CS- odours, the mice showed a preference for the banana odour over the lemon odour, even when banana was the CS- [13].

The neurobiology of memory

When a mouse learns to discriminate between the olfactory CS+ and CS-, it learns two things: to approach and dig in the pot with the CS+ odour in the expectation of obtaining a sugar reward, AND to inhibit digging in the CS- odour pot as no reward is expected. Thus, there must be two memory traces in the mouse brain, one for the CS+ and one for the CS-. In order to study the neural basis of olfactory memory we trained mice with the CS+ only and then looked for epigenetic changes in dopamine (DA) systems in the brain. Male and female WT (C57BL/6J) and *Nrxn1+/-* Transgenic mice were trained at 90-130 days of age to associate a CS+ odour with a sugar reward over 8 trials in a 1-day Pavlovian conditioning protocol. Other mice received no CS+ training. We then examined *DAT*, *DRD1*, and *DRD2* gene expression in the olfactory bulb, olfactory tubercle, and hippocampus of conditioned and unconditioned mice. The Pavlovian conditioned mice showed more digging in the odour pot than naïve mice during both the learning and memory trials, and olfactory learning and memory was not impaired in the *Nrxn1+/-* mice. Olfactory bulb *DRD1* and hippocampal *DAT* and *DRD2* expression was lower in the 7-day than the 24-hour memory test, while olfactory tubercle *DAT* expression was higher following the 7-day than the 24-hour test, indicating tissue-specific gene expression was dependent on time of memory test [12]. These results show that the one day Pavlovian conditioning odour paradigm can be used to study the neurobiological processes underlying memory formation and the results suggest that the tissue-specific changes in DA receptor expression appeared to be associated with DA activity in response to odour presentation and episodic olfactory memories, which we are now examining in a new set of experiments.

Transfer of training

We have also examined whether mice trained on this conditioned odour preference task can apply what they learn in one setting to another in a transfer of training task. Mice were first trained in the one day conditioned odour preference task and given 1 and 7 day memory tests, after which they were presented with the same CS+ and CS- odour pots in either a T maze or a Plus maze. Mice were given four training trials per day for three days in these mazes. Single pieces of sugar were placed in the CS+ odour pots to ensure that the lack of a reward did not result in extinction of the learned CS+ association. Naive mice, which did not have prior odour preference conditioning and memory tests, were also tested for maze learning. The results showed that the experienced mice performed better (Learned the maze faster) than the naïve mice in both the T and Plus maze, often showing high performance from the very first trial of transfer of training [15].

Summary

The Pavlovian conditioned odour preference task is an easy-to-use test which results in short, long and very-long term memories in mice. The one-day training paradigm is useful as it still produces memories in the mice, and results in a shorter window during which learning and memory formation occur, making it a more appealing task for the molecular and cellular study of learning and memory. Mice are also able to transfer what they learn on this task to other environments with ease.

References

1. Brown, R.E., Schnare, O.K., Habib, E.B., Roddick, K.M. (2023). Development of a One-Day Test of Olfactory Learning and Memory in Mice. In: Schaal B, Keller M, Rekow D, and Damon F, editors. *Chemical Signals in Vertebrates 15*, Springer International Publishing, Cham. p. 39–53. https://doi.org/10.1007/978-3-031-35159-4_3

2. Lu, X.-C.M., Slotnick, B.M., Silberberg, A.M. (1993). Odor matching and odor memory in the rat. *Physiology & Behavior*, **53**, 795–804. https://doi.org/10.1016/0031-9384(93)90191-H

3. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2014). Olfactory delayed matching to sample performance in mice: Sex differences in the 5XFAD mouse model of Alzheimer's disease. *Behavioural Brain Research*, **270**, 165–70. https://doi.org/10.1016/j.bbr.2014.04.038

4. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2023). Serial reversal learning in an olfactory discrimination task in 3xTg-AD mice. *Learning & Memory*, **30**, 310–9. https://doi.org/10.1101/lm.053840.123

5. Slotnick, B.M., Katz, H.M. (1974). Olfactory learning-set formation in rats. *Science*, **185**, 796–8. https://doi.org/10.1126/science.185.4153.796

6. Schellinck, H.M., Forestell, C.A., LoLordo, V.M. (2001). A Simple and Reliable Test of Olfactory Learning and Memory in Mice. *Chemical Senses*, Oxford Academic. **26**, 663–72. https://doi.org/10.1093/chemse/26.6.663

7. Forestell, C.A., Schellinck, H.M., Boudreau, S.E., LoLordo, V.M. (2001). Effect of food restriction on acquisition and expression of a conditioned odor discrimination in mice. *Physiology & Behavior*, **72**, 559–66. https://doi.org/10.1016/S0031-9384(00)00439-X

8. Wong, A.A., Brown, R.E. (2013). Prevention of vision loss protects against age-related impairment in learning and memory performance in DBA/2J mice. *Frontiers in Aging Neuroscience*, **5**, 52. https://doi.org/10.3389/fnagi.2013.00052

9. Brown, R.E., Wong, A.A. (2007). The influence of visual ability on learning and memory performance in 13 strains of mice. *Learning & Memory*, **14**, 134–44. https://doi.org/10.1101/lm.473907

10. Wong, A.A., Brown, R.E. (2007). Age-related changes in visual acuity, learning and memory in C57BL/6J and DBA/2J mice. *Neurobiology of Aging*, Elsevier Science, Netherlands. **28**, 1577–93. https://doi.org/10.1016/j.neurobiolaging.2006.07.023

11. O'Leary, T.P., Stover, K.R., Mantolino, H.M., Darvesh, S., Brown, R.E. (2020). Intact olfactory memory in the 5xFAD mouse model of Alzheimer's disease from 3 to 15 months of age. *Behavioural Brain Research*, **393**, 112731. https://doi.org/10.1016/j.bbr.2020.112731

12. Garden, J.F. (2023). Pavlovian Conditioned Odour Memory in the Nrxn1+/- Mouse Model of Autism Spectrum Disorder and Modulation on Dopamine Circuitry. *MSc Thesis*.

13. Croney, P., Brown, R.E. (2022). Performance of mice on a Pavlovian conditioned odour preference task using a Simultaneous odour presentation procedure during training. Unpublished manuscript.

14 Habib, E.B., Brown, R.E. (2020). The Optimization of the Training Parameters for the Conditioned Odour Preference Task. Unpublished manuscript.

15. Burdeyny, V., Brown, R.E. (2021). Transfer of training: From Pavlovian to Instrumental in the T-maze and Plus maze. Unpublished manuscript.

Olfactometer Methodologies in Rodent Olfaction Research: *"What can we test about learning and memory?"*

Wyatt M Ortibus[1] Kyle M Roddick[1] Richard E Brown[1]

[1]Department of Psychology and Neuroscience, Dalhousie University, Halifax, Canada, wy631036@dal.ca

Abstract

The olfactory system of mice involves the prefrontal cortex, amygdala, entorhinal cortex, and hippocampus which encode the olfactory stimuli that guide behaviour. The operant olfactometer allows the evaluation of different aspects of olfactory learning and memory. This presentation examines operant methods for studying olfactory perception, discrimination learning and memory, working memory and reversal learning. Olfactometers provide measures of learning strategies and cognitive flexibility and can detect genotype, sex, and age differences.

Why study olfaction?

The olfactory system of rodents projects to the prefrontal cortex, amygdala, entorhinal cortex and hippocampus which guide instrumental tasks and memory related to odour [1-3]. In rodents, olfactory stimuli are learned and remembered so that they can predict future outcomes based on their current environmental situations. The rodent nervous system builds a representation of the environment which then guides their behaviour [3-6]. Evaluating responses in tests of odour-based incentive learning can provide useful insights into the cognitive processes of these animals.

Animal models, olfactory stimuli, and working memory

Mammals use olfactory cues to learn about their social and physical environment [1-3]. In rodents, the olfactory bulbs are large in comparison to the rest of their forebrain, and their olfactory epithelium is densely packed with olfactory sensory neurons [2-3]. Rats and mice can learn simple and complex odour discrimination tasks [3-5] and the use of olfactometry provides a way to analyze the neural and genetic mechanisms underlying olfactory learning and memory [2-8].

Olfactory working memory involves the temporary storage of information that is used in the performance of cognitive tasks [8]. Short-term olfactory memory in rodents can be analogous to primate visual working memory in terms of how they attempt to solve problems [2-8]. Rats perform better using olfactory cues than visual or auditory cues in discrimination problems [3]. This demonstrates that animals may respond differently to stimuli from different sensory systems and researchers need to be cognizant of the stimuli used in studies with animal models. In rodents, for example, measures of cognitive performance may be better with odour cues then with visual or auditory cues [3, 8].

Simple non-olfactometer olfaction methodologies (habituation & preference tests)

One of the simplest methods for the study odour discrimination in rodents is to use habituation-discrimination tests (HDT) or habituation-dishabituation tests (HDHT) in which a series of odour stimuli are presented, and the investigatory responses of the rats/mice are recorded. In HDT the subject is, in the initial phase of testing, repeatedly presented with an olfactory odour "A" and allowed time to perform investigatory behaviours [2]. During the next phase of testing a novel odour "B" is then presented simultaneously with odour "A", if the subject discriminates between the odours "A" and "B" there will be an increase (dishabituation of odour "A") in the investigatory behaviour to novel odour "B". This provides a simple assessment of memory to odours and discrimination between them. In HDHT the methodologies are similar except that only one odour at a time is presented during both phases and behaviour is scored for investigations or habituations [2]. The main problem with these tests is that when the animal fails to show dishabituation when presented with odour "B" it is not clear whether the animal can not discriminate odour B from A or whether it is simply not motivated to respond [2].

Another simple olfaction test is the Unconditioned Preference Test (UPT), where two or more odours (after a habituation period) are presented and the frequency, duration, or number of approaches to each of the odours are measured [2]. This test indicates if an odour is preferred by the test subject but there is still uncertainty as to whether an increase in investigation score (time, frequency, score, etc.) is indicative of an actual preference. The major drawback to habituation tests is the difficulty in interpreting negative results or "non-responses". The methodology has difficulty in determining if non-responses in a failure to dishabituate are a result of lack of motivation, task difficulty or other unknown variables [9].

The advantages of using operant olfactometers over simple habituation test to study olfactory learning and memory include reducing the uncertainty surrounding non-responding. As the subjects are training to associate odours with unconditioned rewards, non-responding can be assumed to be a lack of motivation. Additionally, rather than measuring the time investigating the odours, which can be subjective, specific responses in the olfactometer, such as nose pokes and licking, are recorded by a computer.

Standard olfactometer methods used for rodent models

Olfactometers allow us to control the air flow and temporal patterns of odours presented to test animals in operant conditioning paradigms which record how model animals respond to odour stimuli [2-7]. Rodents tested in olfactometers are typically water deprived before experimentation to increase their motivation to work for a liquid reward such as water or sucrose solution. Operant olfactometer "*shaping*" programs train rodents to insert their nose into an odour delivery port/tube and receive a reward in response to licking when presented with a conditioned stimulus odour (CS+). The rodents are trained not to lick for reward in the presence of a second odour that indicates a non-reward (CS-) [2-8]. Odours are typically diluted in an odourless solvent in odour saturator bottles which are connected to the olfactometer by closed solenoid valves controlled by a computer. Opening these valves sends a 50 cc/min stream of odourized air into a 1.95 L/min flow of clean air to the odour ports for the rodent. The odour delivery tubes provide a constant air flow which clears previous odour trials and removes any combination of CS+ or CS- odour mixing [2]. Nose pokes into the sample port are detected via an infrared beam over the access hole to the port [2-7].

Using this methodology, the animal's responses to the CS+ and CS- odours are recorded. The latency of each response, inter-response intervals, short samples and non-responses can be recorded for each odour stimulus presentation. There is no punishment for incorrect responses, but they can be added in the form of short timeouts before the next trial. We use a learning criterion of 85% correct responses on blocks of 20 trials, and motivated rats can complete a trial every 12-15 seconds [2-7]. A set number of trials (e.g. 200 trials per session) or a set duration (typically one hour) can be determined for the training period each day. Most rodents complete shaping and learn to discriminate their first operant pair of odours to criterion in one or two daily sessions of testing. Procedures in all of our experiments were approved by the Dalhousie University Committee on Laboratory Animals and followed Canadian Council on Animal Care research guidelines.

The responses of rodents in the olfactometer are not confounded by vision, hearing, or motor deficits, which occur in many mouse models [6,10]. The benefits of using olfactometers in learning and memory research is that olfactory discrimination learning and working memory can be coded and examined based on the test subject's responses to the CS+ and CS- odours [10]. The effects of genotype, age and sex differences on these cognitive processes can also be measured. It is important to measure sex differences as transgenic males and females can produce different discrimination results in the olfactometer due to their ability to detect odours at different concentrations, thus producing significant sex differences in performance to criterion [11].

The downsides to using the olfactometer involve experimenter errors and maintenance issues. Due to the sensitive nature of odour detection, the machines must be kept clean so as to eliminate odour contamination as odours can linger on tubes, sample ports or chambers that the test subject will be in contact with, adding unwanted variability to the experiment. In addition, the discriminations presented in the olfactometer may not represent those found in natural situations [12]. The most common olfactometer problem is odour sample chambers not being fully cleaned between experiments as they could contain unwanted scents from prior experimentation. A test for contamination is to determine if rodents can discriminate between two "clean" odour channels [2-7]. To reduce experimenter errors, we train researchers in the experimental protocols under the supervision of experienced researchers and conduct pilot studies to practice methodology.

It is also possible that some rodents will not learn the shaping procedure. While uncommon, this issue becomes problematic in the olfactometer as the mice cannot progress to the discrimination learning phase of the experiment until they have been shaped to respond to the odours. The most common problem is that some mice have numerous errors (responding to the CS- or failing to respond to the CS+) or produce short samples (removing their nose from the sample port before the odour stimulus is presented in the sample port). Typically, these rodents are removed from the study and documented in case of genotype-sex- or age-related effects. For example, in our current experiment in reversal learning, one out of thirty mice (3.33%) failed to learn the shaping program due to impulsivity or individual differences in performance.

Working memory and the olfactometer

Using an operant olfactometer with four different delays between odour presentation and responses (3, 30, 60 and 120 seconds) to a criterion of 80% correct on two consecutive test sessions results in a significant delay-related performance decrease in rodents' olfactory discrimination [8]. This decrease in working memory performance by rats in the olfactometer was exacerbated by anti-cholinergic treatments using scopolamine, a muscarinic antagonist that produces delay-independent effects in rodent memory paradigms [8]. This type of experiment provides a way to measure short-term recognition memory (working memory) in rodents in a two-choice olfactory discrimination task with a short retention interval, while performance declines when increasing the duration of retention [8,10].

Urine discrimination tasks within olfactometers

Mice can discriminate between their conspecifics using odours from scent glands and urine [1] and the operant olfactometer can be used to examine the variables influencing conspecific urine odour discrimination learning and memory. Using a two odour discrimination task with urine odours, the subject must choose to respond to the rewarded go cue (CS+) and not to respond to the unrewarded no-go cue (CS-), the discriminability of odours from rodents of different genotypes and different dietary conditions was studied [13-15]. An example of this methodology involves testing the ability of rats to discriminate between urine odours of animals maintained on different diets or the influence of gut bacteria on the discriminability of urine odours [14,15]. Using an olfactometer, it was shown that rats were able to discriminate between the urine odours from two genetically identical mice on different diets [15]. Rats also made fewer errors to criteria in learning to discriminate the urine odours of familiar rats where the bacteria had been removed in comparison to the urine odours of unknown rats, the findings suggest there is a "*savings effect*" as individual identifiers in urinary odours are retained after the removal of gut bacteria [14]. The results suggest there is both a genetic component and environmental factors that create identifiable individual odours within the urine [15].

Using signal detection to examine response patterns in the olfactometer

Signal detection theory (SDT) allows us to examine the details of response patterns in the olfactometer [16-18]. SDT enables the discrimination of states between signal and noise and can help conceptualize task performance [17]. The signal and noise conditions create four possible response patterns, described as hits, misses, false alarms, and correct rejections. Hits are a response to a CS+, misses are the non-response to a CS+, false alarms are responses to a CS- and correct rejections are non-responses to a CS-. These four measures can help us to determine how genotype, sex, age, and individual differences affect learning and memory performance [16-18] by observing cognitive and olfactory sensory deficits in mouse models [16]. Lastly, we can use SDT to identify individual behaviour or patterns of responding in tests of cognitive flexibility such as reversal learning or serial reversal learning [10,11,13-16,19,20].

Testing reversal learning & cognitive flexibility in the olfactometer

Tests of reversal learning in the olfactometer allow us to examine cognitive flexibility in genetically modified mouse models of human neurological conditions [21]. Reversal learning tasks require a subject to adjust their behaviour to a reversal of previous CS+ and CS- odours such that the CS- becomes the CS+ and vice-versa [8].

We can examine the effects of genetic manipulation on learning and memory performance by testing mouse models in the olfactometer. We have been testing Neurexin-1 knockdown (Nrxn1 +/-) mouse models of autism spectrum disorder (ASD) for cognitive flexibility using reversal learning. The Neurexin (NRXN) family of membrane protein synaptic organizers are implicated in neurodevelopmental disorders such as ASD and schizophrenia due to their unique functions at the synapse [9] and have been implicated in Alzheimer's Disease [22]

Using the olfactometer, mice are trained on odour pair one (A vs B) until they reach the criterion of 85% correct, and then they are trained on a second odour pair (C vs D) to the criterion of 85% correct. Following this second discrimination learning task, the mice are given a reversal learning task with odours C and D, where the CS+ becomes CS- and vice-versa. This allows us to measure the total number of errors to the 85% criterion, the average time per trial, and total trial blocks to reach the 85% criteria, and to observe any deficits in cognitive flexibility during the reversal learning task.

We are currently examining differences in reversal learning and memory performance between Nrxn1 +/- and wildtype (C57BL/6J) mice to determine the effects of alterations (e.g. knockdowns) in the Nrxn1 genes, which have been shown to alter the development of the prefrontal cortex resulting in impaired cognition and memory performance [23,24]. We have been using signal detection theory to determine the differences in the pattern of responding (hits, misses, false alarms, or correct rejections) in reversal learning between the Nrxn1 +/- and wildtype mice.

Serial reversal learning strategies in the olfactometer

Serial reversal learning is when subjects learn to respond to repeated reversal of the CS+ and CS- in the olfactometer. The performance of mice during these serial reversals indicates behavioural flexibility and cognitive adaptation to a changing association paradigm [19-21]. Subjects could be exposed to concurrent reversals of only one pair of CS+ and CS- (two stimuli) or multiple pairs of CS+ and CS- where only some of the pairs are reversed [21]. During serial reversal tasks, some mice demonstrate near errorless learning as it becomes easier to identify concurrent reversals in the development of a *"learning to learn"* strategy [19,21] . This type of experiment can be used to explore genotype, age, and sex differences in cognitive flexibility [19,21].

Individual rodents' differences in olfactory learning performance

The olfactometer can be used to investigate individual differences in learning and memory in mouse models and to determine differences in response patterns in the reversal learning task. We have seen significant variability of responses within the same genotypes in the learning curves of performance accuracy during reversal learning which we have analyzed using signal detection theory [16]. Using the four measurements of signal detection (*Hits, Misses, False Alarms, Correct Rejections*) and learning performance (Learning curves, Accuracy, Short samples, Trials per min, etc) we can record how each subject responds uniquely in the operant discrimination learning and reversal tasks. For example, while some mice learn the reversal task with 100 total errors in 15 blocks, some mice generate four times more errors before achieving 85% accuracy and have a learning curve that stretch over 40 blocks totaling 400 total errors. By studying the differences between similar genotype, age and sex cohorts in our experiments we can determine if there are different patterns of responses in our operant learning experiments and how the individuality of each rodent's memory performance is demonstrated.

Future applications: human studies and olfactometer measurements

The use of olfactometers allows for the characterization of brain structures in humans related to olfaction by identifying brain activation from the application of olfactory stimuli [25]. Using bold functional MRI to investigate the response in the human brain to odour stimuli it is possible to examine the different brain metabolic patterns between patients with psychotic disorders to determine deficits in olfactory cue identification [25,26]. The results indicate that the odour presentation devices can be used in fMRI sequences and that the odour stimulation activates orbitofrontal cortex, piriform cortex, and cerebellum [25]. It is also possible to investigate changes to the brain-related olfactory systems for olfactory dysfunction (identification/detection) as an early symptom of Alzheimer's disease (AD) or other neurodegenerative diseases [11]. Deficits in olfactory learning and memory have been shown in Tg2576, APP/PS1, 3xTg-AD and 5xFAD mouse models of AD [11]. If significant changes in the human brain using olfactometer measurements can be related to the genetic deficits underlying AD, we can create improved early detection systems for AD.

References

1. Brown, R.E. (1979). Mammalian social odours: A critical review. *Advances in the Study of Behaviour,* ***10***, 103-162. https://doi.org/10.1016/s0065-3454(08)60094-7

2. Slotnick, B.M., Schellinck, H., and Brown, R.E. (2005). Olfaction. In I. Q. Whishaw & B. Kolb (Eds.) *The behaviour of the laboratory rat: a handbook with tests*. Pages 90-104. New York, Oxford University Press.

3. Slotnick, B.M. (2001). Animal cognition and the rat olfactory system. *Trends in Cognitive Sciences*, ***5***(5), 216-222. https://doi.org/10.1016/s1364-6613(00)01625-9

4. Bodyak, N., Slotnick, B.M. (1999), Performance of mice in an automated olfactometer: odour detection, discrimination and odour memory. *Chem. Senses*, **24**(6), 637-45. https://doi: 10.1093/chemse/24.6.637

5. Slotnick, B.M., Restrepo, D. 2005. Olfactometry with mice. *Curr. Protoc. Neurosci,* **33**: 8.20.1–8.20.24. doi:10.1002/0471142301.ns0820s33

6. Schellinck, H.M., Cyr, D.P., Brown, R.E. (2010). How many ways can mouse behavioural experiments go wrong? Confounding variables in mouse models of neurodegenerative diseases and how to control them. *Advances in the Study of Behaviour*, **41**, 255-366.

7. Slotnick, B.M. (1984). Olfactory stimulus control in the rat. *Chemical Senses*, ***9***(2) 157-8. https://doi.org/10.1093/chemse/9.2.157

8. Winters, B., Matheson, W.R., McGregor, I.S., Brown, R.E. (2000). An automated two-choice test of olfactory working memory in the rat: Effect of scopolamine. *Psychobiology*, ***28***(1), 21-31. https://doi.org/10.3758/bf03330626

9. Südhof, T.C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. *Nature*, ***455***(7215), 903-911. https://doi.org/10.1038/nature07456

10. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2014). Olfactory delayed matching to sample performance in mice: Sex differences in the 5xFAD mouse model of Alzheimer's disease. *Behavioural Brain Research*, ***270***, 165-170. https://doi.org/10.1016/j.bbr.2014.04.038

11. Roddick, K. M., Roberts, A.D., Schellinck, H.M., Brown, R.E. (2016). Sex and genotype differences in odour detection in the 3×Tg-AD and 5xFAD mouse models of Alzheimer's disease at 6 months of age. *Chemical Senses*, ***41***(5),433-440. https://doi.org/10.1093/chemse/bjw018

12. Schellinck, H.M., Brown, R.E., Slotnick, B.M. (1991). Training rats to discriminate between the odours of individual conspecifics. *Animal Learning & Behavior*, ***19***(3), 223-233. https://doi.org/10.3758/bf03197880

13. Schellinck, H.M., Monahan, E., Brown, R.E., Maxson, S.C. (1993). A comparison of the contribution of the major histocompatibility complex (MHC) and Y chromosomes to the discriminability of individual urine odours of mice by Long-Evans rats. *Behaviour Genetics*, ***23***(3), 257-263. https://doi.org/10.1007/bf01082464

14. Schellinck, H.M., Brown, R.E. (2000). Selective depletion of bacteria alters but does not eliminate odours of individuality in *Rattus norvegicus*. *Physiology & Behavior*, ***70***(3-4), 261-270. https://doi.org/10.1016/s0031-9384(00)00277-8

15. Schellinck, H.M., West, A.M., Brown, R.E. (1992). Rats can discriminate between the urine odours of genetically identical mice maintained on different diets. *Physiology & Behavior*, ***51***(5), 1079-1082. https://doi.org/10.1016/0031-9384(92)90096-k

16. Roddick, K.M., Fertan, E., Schellinck, H.M., Brown, R.E. (2022). A signal detection analysis of olfactory learning in 12-Month-Old 5xFAD mice. *Journal of Alzheimer's Disease*, **88**(1), 37-44.

17. Wixted, J.T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, ***46***(2), 201-233. https://doi.org/10.1037/xlm0000732

18. Steckler, T. (2001). Using signal detection methods for analysis of operant performance in mice. *Behavioural Brain Research*, ***125***(1-2), 237-248. https://doi.org/10.1016/s0166-4328(01)00305-9

19. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2023). Serial reversal learning in an olfactory discrimination task in 3xTg-AD mice. *Learning & Memory*, ***30***(12), 310-319. https://doi.org/10.1101/lm.053840.123

20. Bond, A.B., Kamil, A.C., Balda, R.P. (2007). Serial reversal learning and the evolution of behavioural flexibility in three species of North American corvids (Gymnorhinus cyanocephalus, Nucifraga Columbiana, Aphelocoma californica). *Journal of Comparative Psychology*, ***121***(4), 372-379. https://doi.org/10.1037/0735-7036.121.4.372

21. Izquierdo, A., Brigman, J., Radke, A., Rudebeck, P., Holmes, A. (2017). The neural basis of reversal: An updated perspective. *Neuroscience, **345,** 12-26. https://doi.org/10.1016/j.neuroscience.2016.03.02*

22. Medina- Samamé, A., Paller, É., Bril, M.R., Archvadze, A., Simões-Abade, M.B., Estañol-Cayuela, P., Lemoult, C. (2023). Role of Neurexins in Alzheimer's disease. *The Journal of Neuroscience,* **43**(23), 4194-4196. https://doi.org/10.1523/jneurosci.0169-23.2023

23. Hughes, R.B., Whittingham-Dowd, J., Simmons, R.E., Clapcote, S.J., Broughton, S.J., Dawson, N. (2019). Ketamine restores thalamic-prefrontal cortex functional connectivity in a mouse model of neurodevelopmental disorder-associated 2p16.3 deletion. *Cerebral Cortex*, ***30***(4), 2358-2371. https://doi.org/10.1093/cercor/bhz244

24. Hughes, R.B., Whittingham-Dowd, J., Clapcote, S. J., Broughton, S. J., Dawson, N. (2022). Altered medial prefrontal cortex and dorsal raphe activity predict genotype and correlate with abnormal learning behaviour in a mouse model of autism-associated 2p16.3 deletion. *Autism Research*, **15**(4), 614-627 https://doi.org/10.1002/aur.2685

25. Sommer, J. U., Maboshe, W., Griebe, M., Heiser, C., Hörmann, K., Stuck, B.A., Hummel, T. (2012). A mobile olfactometer for fMRI-Studies. *Journal of Neuroscience Methods,* 209(1), 189-194. https://doi.org/10.1016/j.jneumeth.2012.05.026

26. Gates, L., Good, K., Schellinck, H.M., Brown, R.E., Kopala, L. (2001). Olfactometric methods to examine fMRI brain activation in healthy subjects: Application to psychotic disorders. *NeuroImage*, *13*(6), 882. https://doi.org/10.1016/s1053-8119(01)92224-4

Olfactory Cues in Zebrafish Behavioural Studies: From Anxiety to Learning

R Gerlai

Department of Psychology, University of Toronto Mississauga, Ontario, CANADA, robert.gerlai@utoronto.ca

Zebrafish have become popular in biomedical research including behavioural neuroscience. They represent a compromise between system complexity (vertebrates with evolutionarily conserved features) and practical simplicity (small, cheap to keep and test in large numbers) [1]. They share 3.3 billion years of biological evolution with mammals, and thus are translationally relevant for modeling human diseases. Most behavioral paradigms developed for zebrafish utilize visual stimuli. Nevertheless, other modalities may also be exploited. One of which is olfaction. Here, I will briefly review two domains of behavioral studies where olfactory cues have been or may be successfully employed.

The first domain is anxiety. Anxiety is a major unmet medical need. Olfactory stimuli have been successfully employed to induce anxiety-like responses. When a predator catches fish, the prey fish's skin is damaged and releases alarm substance from club-cells. This mixture of several molecules alerts neighboring fish of imminent danger. In the lab, the skin of freshly euthanized zebrafish is cut and the wound is washed with water. The collected liquid is then aliquoted, and a dilution sequence is created for dose response analysis. We found zebrafish to respond to freshly extracted alarm substance with a variety of antipredatory responses: elevated freezing, erratic movement (zig-zagging), leaps, reduction of swimming activity, tightening of shoal cohesion[2]. These responses are also elicited by other aversive stimuli and diminished by anxiolytic drugs. However, the use of alarm substance is not without problems. The concentration and compound constitution of the extracted substance are impossible to standardize between experiments. A better solution was offered after the chemical structures of specific constituents of the alarm substance have been identified. For example, one of them was found similar to H3NO (hypoxanthine 3-N-oxide), which could be chemically synthesized. We found this compound to induce anxiety-like (antipredatory) responses in zebrafish [2]. The molecular weight of a synthetic compound is known, thus its concentration is precisely determined allowing dose response analyses across experiments. The downside of H3NO is that it is unstable, looses its potency during long-term storage. Given the varied chemical composition of the alarm substance, it is possible that stable synthetic compounds will be identified. Use of such compounds will facilitate anxiolytic screens with zebrafish.

The second domain I discuss is learning & memory. Learning & memory related human disorders represent major unmet medical needs. The zebrafish is perfect for large scale mutagenesis or drugs screens for learning & memory indications, as it is prolific and can be kept in large numbers[3]. But, similarly to anxiety research, most learning tasks use visual cues. However, olfactory cues may be employed both as unconditioned stimuli (US) and as conditioned stimuli (CS). US's are stimuli that elicit innate (spontaneous) responses without the need for training [3]. The smell of alarm substance, e.g., induces fear/anxiety responses without prior experience. This aversive US may be employed in classical (CS/US) conditioning tasks where punishment is planned as a reinforcement. It would be better as the traditional electric shock because the latter may affect neuronal communication (depolarization) directly. Appetitive olfactory cues may also be available. In appetitive CS-US conditioning, the US is a reward, i.e., a stimulus that is instinctively interpreted as pleasurable. The smell of food may be such an olfactory stimulus. However, it has not been employed in classical conditioning with zebrafish. Similarly, males of several fish species can detect the presence of receptive females. This has not been shown in zebrafish, but if it applies to this species too such an olfactory cue may be utilized in appetitive conditioning. Olfactory cues may also be utilized as CS. For an olfactory cue to be an appropriate CS, it should have no intrinsic value, i.e., it should not have aversive or appetitive properties. Zebrafish have been shown to be able to detect a variety of olfactory cues. Investigation into their neutral nature thus may enable one to use such cues as CS, research that has not been conducted.

The last question we discuss is why one should consider using olfactory cues instead of visual stimuli in anxiety or learning and memory research with zebrafish at all, given how much easier it is to control the on-set and offset of the latter. Also, numerous cheap consumer grade devices (computer monitors, cameras) are available. On the other hand, delivery of olfactory cues requires ingenious hardware designs (e.g., injection methods without presence of the experimenter). Also, olfactory cues, once delivered, take time to diffuse and reach the test subject, and they are difficult to localize to particular areas of the test environment. Their removal from the test tank is also not an easy task (problems that may be solved using flow-through or filtering methods). Why would anyone want to deal with these practical complications? The reason is two-fold. One, olfactory stimulus induced responses may allow one to tap into functional aspects of the brain that differ from those engaged by visual cues. Two, in large scale mutagenesis and small molecule screens one will never know the precise reason for a positive hit. That is, a mutation or a compound may alter behavioral performance not because it diminished anxiety, or because it impaired learning, but because it affected perception. If one uses cues of only one modality, such as vision, one may obtain false positives, e.g. a blind fish in a learning task. If olfactory cue-based paradigms are available, they would allow the investigator to generalize across modalities. In sum, using multiple paradigms with distinct performance (motor, motivational and perception-related) requirements is important for behavioral screening. The use of olfactory cues thus would be a welcome development in zebrafish research [4].

Ethics statement

Research reviewed in this abstract has been peer reviewed and published, and adhere to Governmental regulations and the law concerning ethical and humane use of animals in research.

References

1. Gerlai R (2020). The zebrafish in Behavioral and Neural Genetics. In: Gerlai R (Ed). Behavioral and Neural Genetics of Zebrafish, Elsevier, Academic Press, Amsterdam, The Netherlands. ISBN: 9780128175286 pp xix

2. Gerlai R (2020). Fear responses and antipredatory behavior of zebrafish: A translational perspective. In: Gerlai R (Ed). Behavioral and Neural Genetics of Zebrafish, Elsevier, Academic Press, Amsterdam, The Netherlands ISBN: 9780128175286 pp155-173

3. Gerlai R (2016) Learning and memory in zebrafish (*Danio rerio*). In Detrich HW, Westerfield M, & Zon LI (Eds.), The Zebrafish: Cellular and Developmental Biology Part B 4th Edition. Elsevier, Amsterdam, p. 551–586.

4. Gerlai R (2020). The design of behavioural screening in zebrafish. In: Gerlai R (Ed). Behavioral and Neural Genetics of Zebrafish, Elsevier, Academic Press, Amsterdam, The Netherlands ISBN: 9780128175286 pp513-526.

Symposium: Using behavioural approaches to measure apathy-like behaviour in rodents

Development of a Behavioural Test Battery to Assess Apathy-like Behaviours in Mouse Models of Neurodegeneration

L. Robinson and G. Riedel

School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, Scotland, United Kingdom.
lianne.strachan@abdn.ac.uk

Background

Apathy is the most common behavioural and psychological symptom in Alzheimer's disease (AD) and other neurodegenerative diseases including frontotemporal dementia (FTD) and Parkinson's disease (PD). In patients, apathy can include symptoms of loss of motivation, initiative and interest, listlessness, and indifference, flattening of emotions, absence of drive and passion. Modelling of apathy-like behaviour in animals has proven difficult with many studies focusing on specific symptoms of apathy-like behaviour including reduced goal directed, motivation and reward-based activity, decreased exploration/spontaneous activity, or a reduction in species specific behaviours including nest building and burrowing. These behaviours are seen as indicators of impaired motivation to perform daily life activities that are reminiscent of the reduced activities of daily living (ADLs) observed in dementia patients with apathy.

Objective

Development of a pre-clinical behavioural test battery to assess apathy-like behaviour in transgenic mouse lines of different neurodegenerative diseases.

Materials and Methods

Subjects

Male and female transgenic mouse lines aged 6 – 8 months (Charles River Laboratories; Margate; UK) were used in the study including Line 1 and Line 66 Tau mouse models for AD and FTD [1]; PLB4 BACE1 model of AD [2] and the Line 62 alpha- synuclein model of PD [3] along with relevant Wild-type's. All N's = 10 – 17 mice. They were group housed in controlled open housing conditions (Makrolon type III cages, corncob bedding, ambient temperature 21±1°C, relative humidity 50-65%, with 17–20 air changes per hour) within the Medical Research Facility at the University of Aberdeen. Food (Special Diet Services, Witham, UK) and water was available *ad libitum*, and a circadian rhythm was maintained on a 12hr light/dark cycle (lights on at 7am) with simulated sunrise and sunset (30 min). All behavioural tests were performed during the light cycle. Sample sizes were based on power calculations and all experiments were performed blind, counterbalanced and in accordance with the European Communities Council Directive (63/2010/EU) and a project license with local ethical approval under the UK Animals (Scientific Procedures) Act (1986) and its Amended Regulations (2012), and following ARRIVE 2.0 guidelines.

Behavioural testing

Animals were randomly allocated to cohorts and tested in a series of sequential behavioural tests to assess apathy-like symptoms (See Figure 1 for timeline of experiments).

Home cage activity: Animals were initially individually housed in PhenoTyper cages (Noldus, NL). Each cage was made of clear Perspex (30 x 30 x 35 cm) and contained a food hopper and water bottle offering access to food and water ad libitum. The cages were filled with sawdust and the locomotor activity of the mice were recorded 24hrs a day for a period of 7 days via built-in digital infrared sensitive video cameras.

Nest Building: To assess nest building behaviour, animals were single housed in Makrolon Type III cages (Tecniplast, Milan, Italy) containing corn cob, saw dust bedding, a cardboard tube (DBM Scotland Ltd, UK) and

1 nestlet (50 mm x 50mm square pressed cotton, DBM Scotland Ltd, UK) prior to the start of the dark cycle with access to food and water *ad libitum*. Scoring of nests was performed after a period of 16 hours (Day 1) and 48 hrs (Day 2). The scoring of the nests was performed by three independent researchers two of whom were blind to the genotype of the mouse. The score of the three researchers was averaged and used for analysis. A score of 1 - 5 was allocated depending on how much of the nestlet had been shredded and a nest formed (see Figure 1 for representative photos). A score of 1 was assigned if the nestlet remained pre-dominantly untouched. If it was partially torn up a score of 2 was given and when the nestlet had been almost entirely shredded although there was no clear nest area a score of 3 was assigned. Only when the nestlet was entirely shredded and a nest area established was a score of either 4 (flattened nest) or 5 (perfect crater shaped nest) assigned.

Sucrose preference test: For sucrose preference the mice were individually housed in activity cages containing corn cob bedding and equipped with two drinking bottles (54 x 50 x 37cm) (Ugo Basile, Italy) and food available ad libitum. One of the bottles contained water and the other a 1% sucrose solution with the position of the sucrose bottle (left or right) counterbalanced for genotype. After 24 hrs, both bottles were weighed, and the position of the sucrose bottle alternated to avoid any spatial preference. The weights of the two bottles were recorded again after a further 24 hrs. Water and sucrose consumption for each animal was averaged for the two days and sucrose preference determined as sucrose consumption divided by the total intake of sucrose and water.

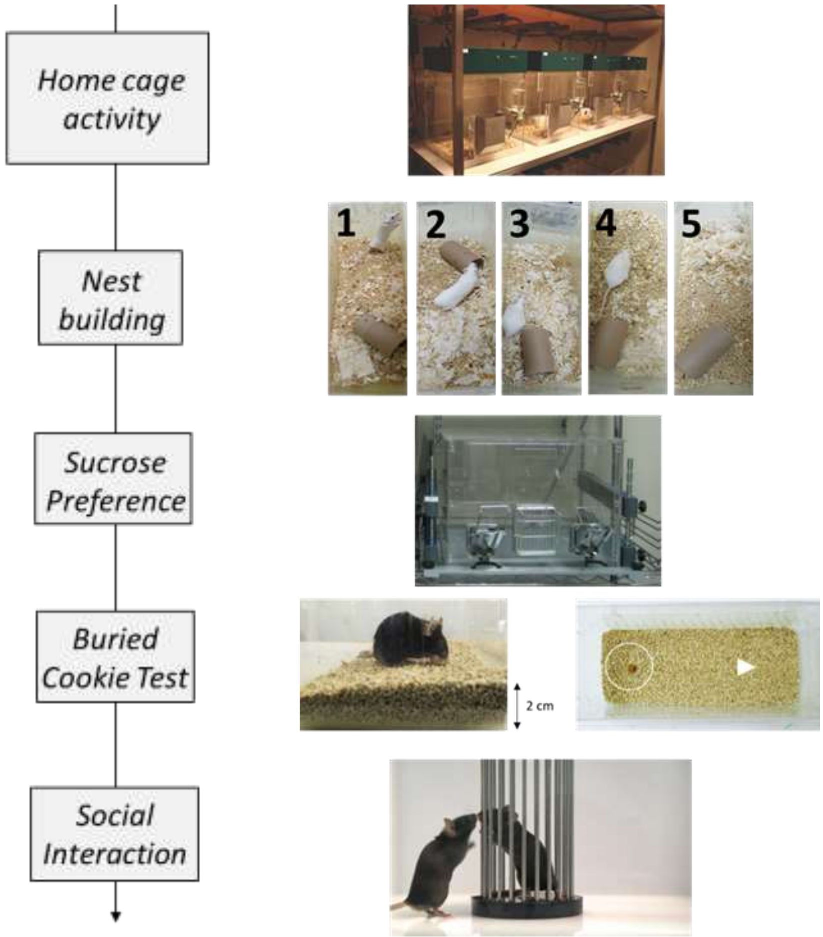


Figure 1. Experimental timeline of behavioural tests. Home cage activity of the mice was initially tested in the Phenotyper for a period of 7 days, this was followed by nest-building behaviour in their home cages and then assessment of sucrose preference. The buried cookie test was the penultimate behavioural task performed by the animals with assessment of social activity being the final task.

Buried Cookie Test: Testing was performed in Makrolon Type III cages (Tecniplast, Milan, Italy) filled with corn cob bedding to a depth of 2cm. In the week prior to the test animals were habituated to the cookie in their home cage. All animals were food restricted overnight (~16 hours) prior to testing to ensure that they were hungry and motivated to perform the task. During the test animals performed a series of four trials with each trial having a maximum duration of 5 minutes. The first trial was a habituation trial in which the animal was placed in the cage and behavioural activity recorded by an overhead camera. The habituation trial was followed by two buried cookie trials in which a piece of cookie was hidden ~1cm below the surface of the bedding. On each trial the cookie was hidden in a different location (left or right) and the animal released in the centre of the cage. Latency to locate the cookie was recorded by the experimenter using a stopwatch. During the fourth trial the cookie was visible and positioned on the surface of the corn cob. After the mouse had located the cookie on each trial it was returned to its' home cage and an inter trial interval (ITI) of ~1 minute implemented.

Social Interaction: Social activity of the mice was assessed in a three-chamber arena (each chamber: 20 x 42 x 22 cm) with interaction cylinders located in each of the outer chambers. Animals were initially habituated to the arena containing two empty cylinders for 10 minutes before the presentation of an unfamiliar stranger mouse in one of the cylinders during the sociability trial (maximum duration 10 minutes). The ITI between the trials was ~5 minutes during which time the mouse was returned to its' home cage and the arena cleaned with 70% ethanol. The position of the stranger mouse was counterbalanced for genotype and animals were always released in the centre chamber. Overall activity and time spent in the vicinity of the stranger mouse were recorded using an overhead camera.

Data Analysis

Behavioural parameters were recorded using the behavioural video-tracking software of Ethovision (Noldus, NL) or AnyMaze (Ugo Basilie, Italy). Statistical analyses were performed using both parametric and non-parametric tests followed by appropriate post hoc tests using GraphPad Prism (version 10; GraphPad Software Inc., USA) with a 95% confidence level assumed and alpha set to 5%.

Results

L66 homozygous mice displayed a reduction in ambulatory activity compared to WTs in the phenotyper although no further differences in activity were observed between transgenic lines. Similarly, only L66 homozygous mice presented with a significant impairment in nest building compared to WT's (p=0.002). In the sucrose preference test a deficit in sucrose preference was observed for both L66 homozygous mice (p=0.0081) and PLB4 (p=0.0008) mice when compared to respective WT's with L62 mice displaying intact preference for the sucrose solution. The performance of transgenic mice in the buried cookie test was also comparable to that of WTs, although sex differences were apparent with the L62 mice. Finally, no differences in social interaction were observed for any of the transgenic lines although L66 mice continued to present with an overall reduction in activity levels (p=0.02).

Conclusions

Utilising a battery of behavioural assays allowed us to model behavioural traits indicative of apathy-like symptoms in mouse models of neurodegenerative diseases. The FTD mouse model L66 displayed reduced nest building behaviour and sucrose preference which are consistent with previous findings from our group. Furthermore, the reduced activity levels observed with L66 mice could also indicate motivational deficits. The BACE1 mouse model PLB4 presented with impaired sucrose preference although no further apathetic-like behaviours were observed in these mice. The presence of an apathy-like phenotype in the FTD mouse model could be accounted for by the genetic species of human tau expressed in these mice (P301S) and is consistent with previous studies using mice that express htau mutations.

References

1. Melis, V., Zabke, C., Stamer, K., Magbagbeolu, M., Schwab, K., Marschall, P., Veh, R.W., Bachmann, S., Deiana, S., Moreau, P.H., Davidson, K., Harrington, K.A., Rickard, J.E., Horsley, D., Garman, R., Mazurkiewicz, M., Niewiadomska, G., Wischik, C.M., Harrington, C.R., Riedel, G., Theuring, F. (2015).

Different pathways of molecular pathophysiology underlie cognitive and motor tauopathy phenotypes in transgenic models for Alzheimer's disease and frontotemporal lobar degeneration. *Cell Mol Life Sci.* **72** (11): 2199-222.

2. Plucińska, K., Crouch, B., Koss, D., Robinson, L., Siebrecht, M., Riedel, G., Platt, B. (2014). Knock-in of human BACE1 cleaves murine APP and reiterates Alzheimer-like phenotypes. *J Neurosci.* **34**(32):10710-28.

3. Frahm, S., Melis, V., Horsley, D., Rickard, J.E., Riedel, G., Fadda, P., Scherma, M., Harrington, C.R., Wischik, C.M., Theuring, F., Schwab, K. (2018). Alpha-Synuclein transgenic mice, h-α-SynL62, display α-Syn aggregation and a dopaminergic phenotype reminiscent of Parkinson's disease. *Behav Brain Res.* **26**; 339:153-168.

The splash test as a model of rodent apathy

Nicole Edwards, Shuzo Sakata, Trevor J. Bushell

Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow.

Introduction

Major depressive disorder (MDD) is one of the leading causes of disability worldwide, affecting 5% of the population, with the lifetime prevalence for MDD as high as 15-20%. MDD is diagnosed when the three prominent features; low mood, anhedonia, and apathy, are present for a minimum period of two weeks[1–3]. The duration and severity of MDD can range from a moderate singular event to a recurrent lifelong condition. Depression is a common co-morbidity of many chronic diseases including cancer, cardiovascular, inflammatory, and neurodegenerative disorders, and can contribute to the psychological burden associated with these diseases[4,5]. In addition, MDD has been shown to cause cognitive decline, particularly in elderly patients, including executive function impairment and memory deficits. Furthermore, when MDD is concomitant with a neurodegenerative disease such as Alzheimer's disease (AD), evidence shows that this can exacerbate disease progression leading to a worse prognosis[6]. MDD has been identified as the most common neuropsychiatric disorder in patients with diagnosed AD, and evidence also supports a 2-3 fold increase in risk of developing AD in patients with a history of MDD[7,8]. Furthermore, studies have demonstrated an association between the number of depressive episodes and risk of developing dementia, with each recurrent episode of depression increasing the risk of dementia by 14%[9,10]. MDD has also been suggested to play a role in the progression from normal cognition to MCI and from MCI to dementia(11,12). Reactive glial cells, including astrocytes and microglia, are proposed to contribute to the AD development, with the theory that neuroinflammation drives disease progression[13,14]. Neuroinflammation has also been indicated in MDD, with patients with MDD reporting upregulated inflammatory cytokines which may underlie this accelerated progression in AD patients[15,16]. However, whether depression-like behaviour induced in AD mouse models leads to accelerated AD pathology is unknown.

Hypothesis

Recently, we revealed that protease-activated receptor 2 (PAR2) activation induces depression-like behaviour and cytokine release *in vivo*. Hence, we hypothesise that depression-like behaviour induced in a mouse model of AD will increase the neuroinflammatory response and exacerbate amyloid pathology. Using the 5xFAD mouse model of amyloid pathology in AD, a model that rapidly develops amyloid plaque pathology at approximately 2 months of age, a variety of behavioural tests, immunohistochemistry, ELISAs and qPCRs will be performed to test our hypothesis and provide answers to the following research questions:

7. Do 5xFAD$^{+/-}$ mice exhibit depression-like behaviour when compared to 5xFAD$^{-/-}$ littermate controls?
8. Are 5xFAD$^{+/-}$ mice more susceptible to pharmacologically-induced depression-like behaviour than 5xFAD$^{-/-}$ littermate controls?
9. Does single intervention or prolonged multiple dosing of pharmacologically induced depression-like behaviour exacerbate behavioural, inflammatory, and molecular pathology in 5xFAD$^{+/-}$ mice?

Behavioural methods

Modelling depression in rodents has commonly involved techniques including chronic stress exposure, genetic manipulation, and pharmacological intervention[17,18]. A depressed-like phenotype is then tested and measured in the form of behavioural despair, usually by means of escape-oriented behavioural tests situation, such as the forced swim test and tail suspension test and learned helplessness tests, such as shock avoidance[19–21]. Despair behaviour is often extrapolated as being depression-like, but chronic stress to rodents also produces anxiety-like behaviour[22]. These tests can often be unpleasant and invasive and take weeks of repeated stressors to develop. Our project is focussing on testing PAR2-pharmacologically-induced depression-like behaviour, rather than stress-inducing depression-like behaviour[23,24]. Therefore, we are aiming to test behaviours that recapitulate the three core symptoms in human depression: low mood, anhedonia, and apathy.

Low mood in humans can be represented in rodents as reduced locomotor activity and willingness to explore which can be measured as distance travelled in the open field test (OFT) [19,25]. Anhedonia, the inability to experience pleasure from normally pleasurable activities is commonly measured in rodents using the sucrose preference test (SPT), whereby mice experiencing depression-like behaviour show a decreased preference to a 1% sucrose solution in a two-bottle choice paradigm[26,27]. Apathy, defined as a loss of interest or motivation, is associated with impairment in goal-directed behaviours such as lack of self-care in MDD. In rodents, common apathetic features include unkempt fur due to lack of self-care. The splash test can be used to evaluate self-care in rodents, whereby a 10% sucrose solution is sprayed on the rodents dorsal coat and grooming behaviour is recorded. Duration of grooming and latency to first groom is usually measured with a decreased period of grooming considered as reduced motivation and self-care behaviour associated with depression-like behaviour[28].

To test the research questions proposed, we are using the PAR2 activator, AC264613 (AC), which we have previously been shown to induce both inflammatory changes and depression-like behaviour *in vivo*[23,24]. AC (100mg kg^{-1}) will be tested alongside vehicle (0.9% saline solution with 1% Tween 80) control and lipopolysaccharide (LPS: 0.5mg kg^{-1}) as a positive control well documented to induce depression-like behaviour[29,30]. Experiments were performed using 10-12-week old transgenic 5xFAD heterozygous mice ((5xFAD$^{+/-}$ and FAD$^{-/-}$ male and female littermates, (strain: B6SJL-Tg (APPSwFILon, PSEN1*M146L*L286 V) 6799Vas/Mmjax: genetic background: CJ57BL/6J)) and CJ57BL/6J WT mice, which were group-housed under standard conditions: 21± 2°C, 45–65% humidity, 12hr dark/light cycle (7am-7pm) in MB1 cages (45 × 28 × 13 cm) lined with grade 6 wood bedding, containing a plastic house, tunnel, and nesting material. During week one animals were singly housed in MB1 cages for the purposes of the sucrose preference and regrouped after the last measurement. Animals had free access to food and water with all behavioural procedures carried out between 9am-5pm. Treatments were randomly assigned and administrated intraperitoneally (i.p.). Behavioural tests to examine locomotor activity, anhedonia and apathy were performed 2hr and 24hr post-injection using the open field test, sucrose preference test and the splash test, respectively. Behavioural testing took place over two weeks, with one injection given per week. The open field test and sucrose preference test took place in week one and the splash test was conducted in week two, allowing animals one week to fully recover from the intervention treatment (Fig 1).

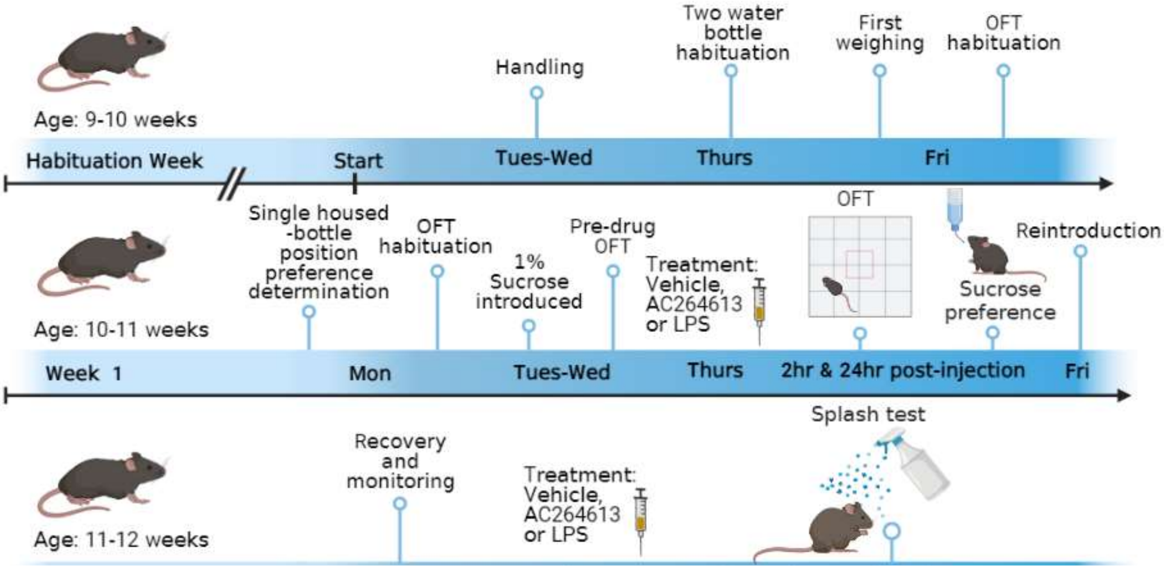


Figure 1: Schematic representation of the behavioural testing protocol starting from the first handling sessions and habituation to behavioural testing arena and apparatus. Following the relevant i.p. injection, behavioural tests were carried out 2h and 24h post injection, over a period of two weeks with OFT and sucrose preference conducted in week 1 and the splash test conducted in week 2.

For the open field test, mice were individually placed in a 40x40x40cm white Perspex box, with a webcam (Nulaxy HD 1080p) situated overhead and habituated for two, 20-40 min sessions on consecutive days. Following habituation, mice were recorded for 20 mins for two pre-injection (pre-drug 1 and 2) open field test sessions on consecutive days to determine baseline locomotor activity. Following i.p. injection, mice were recorded in the OFT arena 2h and 24h post-injection for 20mins. Locomotor activity was measured as distance travelled, determined using an artificial intelligence (AI) software python package, DeepLabCut (DLC)(31)to track animal movement. Animal pose co-ordinates generated in DLC were imported into a MatLab script (written by Dr Shuzo Sakata, University of Strathclyde) to plot body position trajectories and calculate animal speed and distance travelled.

Before the test, each mouse was singly housed in an MB1 cage with two identical bottles of water, with the amount of water consumed for each measured over 24h to determine bottle position preference. For the next 2 days, mice were habituated to sucrose by replacing the water bottle in the non-preferred position with an identical bottle containing 1% sucrose solution, made up with sucrose (Fisher Scientific, UK) and fresh tap water (Fig. 12). The weights of both bottles were recorded daily, and the sucrose and water bottle positions were switched after 24h to allow mice to habituate to the sucrose bottle in both positions. Following habituation, sucrose and water consumption was measured pre-injection, 2h post- and 24h post-injection with sucrose consumption (%) being calculated at each time point. Sucrose was deemed preferred over water at 50% sucrose consumption at the pre-injection measurement. Mice below this threshold were excluded from results. Mice were regrouped in MB1 cages after the 24h sucrose weighing and monitored.

In testing week 2, mice received a second i.p. injection of assigned treatment intervention. At 2h and 24h post injection, mice were sprayed on their dorsal coat twice with a 10% sucrose solution (made up of sucrose (Fisher scientific, UK) and fresh tap water), then immediately placed in the OFT arena and recorded for 10mins. Grooming behaviour was measured as total duration of grooming, time to first groom and number of grooming episodes using DLC to track animal movement and a MatLab script (written by Dr Shuzo Sakata, University of Strathclyde) to analyse animal pose co-ordinates. Grooming behaviour was also measured manually to validate AI-generated results.

All data is reported as mean ± S.E.M. with individual data points and n = the number of animals. Behavioural tested was analysed blind to treatment group. Sex and genotypic differences were compared to identify any significant differences in drug effect groups. The statistical analysis tests used are one-way repeated-measures ANOVA with Tukeys post hoc multiple comparison and two-way repeated-measures ANOVA with Bonferroni post hoc multiple comparison tests conducted using GraphPad Prism (v8.4.3). Significance is reported when $p<0.05$.

Results

AC- and LPS-induced depression-like behaviour was observed 2h post-injection as reduced locomotor activity (AC: ($F_{(1-23)} = 9.581$ **p=0.0025 vs pre-drug, n=19; LPS: $F_{(2-27)} = 21.21$, ***p<0.0001 vs pre-drug, n=16), reduced sucrose preference ($F_{(2-36)} = 16.18$ ***p<0.0001 vs pre-drug, n=19; LPS: $F_{(2-29)} = 6.219$, **p=0.0062 vs pre-drug, n=16) and reduced grooming behaviour (AC: **p=0.0028 AC vs vehicle, n=15; LPS: **p=0.025 LPS vs vehicle, n=13) in both 5xFAD$^{+/-}$ and 5xFAD$^{-/-}$ mice which was recovered 24h post-injection. No behavioural differences between sex and genotype in naïve 5xFAD mice were observed. The three behavioural tests performed have proven useful models for measuring depression-like behaviour and the splash test has demonstrated good validity as a measure of natural grooming behaviour and apathetic behaviour.

To continue testing our hypothesis, immunohistochemistry is currently underway to assess amyloid plaque density, as well as microglial and astrocytic reactivity following intervention. Further, AC- and LPS-induced peripheral inflammatory markers will be examined between 5xFAD$^{+/-}$ and 5xFAD$^{-/-}$ mice. In conclusion, this study reveals that pharmacologically-induced depression-like behaviour is similar in both 5xFAD$^{+/-}$ and 5xFAD$^{-/-}$ mice, with analysis underway to examine its effect on the inflammatory and molecular pathology observed in 5xFAD$^{+/-}$ mice.

References

1. Bains N, Abdijadid S. Major Depressive Disorder [Internet]. StatPearls. 2023. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30396512
2. Kamran M, Bibi F, ur. Rehman A, Morris DW. Major Depressive Disorder: Existing Hypotheses about Pathophysiological Mechanisms and New Genetic Findings. Genes (Basel) [Internet]. 2022 Apr 6;13(4):646. Available from: https://www.mdpi.com/2073-4425/13/4/646
3. Fekadu N, Shibeshi W, Engidawork E. Major Depressive Disorder: Pathophysiology and Clinical Management. J Depress Anxiety [Internet]. 2017;06(01). Available from: https://www.omicsonline.org/open-access/major-depressive-disorder-pathophysiology-and-clinical-management-2167-1044-1000255.php?aid=80829
4. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Prim [Internet]. 2020 Aug 20;6(1):69. Available from: https://www.nature.com/articles/s41572-020-0200-2
5. Berk M, Williams LJ, Jacka FN, O'Neil A, Pasco JA, Moylan S, et al. So depression is an inflammatory disease, but where does the inflammation come from? BMC Med [Internet]. 2013 Dec 12;11(1):200. Available from: http://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-11-200
6. Martín-Sánchez A, Piñero J, Nonell L, Arnal M, Ribe EM, Nevado-Holgado A, et al. Comorbidity between Alzheimer's disease and major depression: a behavioural and transcriptomic characterization study in mice. Alzheimers Res Ther [Internet]. 2021 Apr 2;13(1):73. Available from: https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00810-x
7. Sáiz-Vázquez O, Gracia-García P, Ubillos-Landa S, Puente-Martínez A, Casado-Yusta S, Olaya B, et al. Depression as a Risk Factor for Alzheimer's Disease: A Systematic Review of Longitudinal Meta-Analyses. J Clin Med [Internet]. 2021 Apr 21;10(9). Available from: http://www.ncbi.nlm.nih.gov/pubmed/33919227
8. Ownby RL, Crocco E, Acevedo A, John V, Loewenstein D. Depression and risk for Alzheimer disease: systematic review, meta-analysis, and metaregression analysis. Arch Gen Psychiatry [Internet]. 2006 May;63(5):530–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16651510
9. Kessing L V, Andersen PK. Does the risk of developing dementia increase with the number of episodes in patients with depressive disorder and in patients with bipolar disorder? J Neurol Neurosurg Psychiatry [Internet]. 2004 Dec;75(12):1662–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15548477
10. Dafsari FS, Jessen F. Depression—an underrecognized target for prevention of dementia in Alzheimer's disease. Vol. 10, Translational Psychiatry. Springer Nature; 2020.
11. Gallagher D, Kiss A, Lanctot K, Herrmann N. Depression and Risk of Alzheimer Dementia: A Longitudinal Analysis to Determine Predictors of Increased Risk among Older Adults with Depression. Am J Geriatr Psychiatry [Internet]. 2018 Aug;26(8):819–27. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1064748118303270
12. Brendel M, Pogarell O, Xiong G, Delker A, Bartenstein P, Rominger A. Depressive symptoms accelerate cognitive decline in amyloid-positive MCI patients. Eur J Nucl Med Mol Imaging [Internet]. 2015 Apr 29;42(5):716–24. Available from: http://link.springer.com/10.1007/s00259-014-2975-4
13. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nat Rev Neurol [Internet]. 2021 Mar 14;17(3):157–72. Available from: https://www.nature.com/articles/s41582-020-00435-y
14. Kwon HS, Koh SH. Neuroinflammation in neurodegenerative disorders: the roles of microglia and astrocytes. Transl Neurodegener [Internet]. 2020 Dec 26;9(1):42. Available from: https://translationalneurodegeneration.biomedcentral.com/articles/10.1186/s40035-020-00221-2
15. Dowlati Y, Herrmann N, Swardfager W, Liu H, Sham L, Reim EK, et al. A meta-analysis of cytokines in major depression. Biol Psychiatry [Internet]. 2010 Mar 1;67(5):446–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20015486
16. Felger JC, Lotrich FE. Inflammatory cytokines in depression: Neurobiological mechanisms and therapeutic implications. Neuroscience [Internet]. 2013 Aug;246:199–229. Available from: https://linkinghub.elsevier.com/retrieve/pii/S030645221300393X

17. Willner P. Validity, reliability and utility of the chronic mild stress model of depression: a 10-year review and evaluation. Psychopharmacology (Berl) [Internet]. 1997 Dec;134(4):319–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9452163
18. Willner P. The chronic mild stress (CMS) model of depression: History, evaluation and usage. Neurobiol Stress [Internet]. 2017 Feb;6:78–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28229111
19. Wang Q, Timberlake MA, Prall K, Dwivedi Y. The recent progress in animal models of depression. Prog Neuropsychopharmacol Biol Psychiatry [Internet]. 2017 Jul 3;77:99–109. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28396255
20. Steru L, Chermat R, Thierry B, Simon P. The tail suspension test: A new method for screening antidepressants in mice. Psychopharmacology (Berl) [Internet]. 1985 Mar;85(3):367–70. Available from: http://link.springer.com/10.1007/BF00428203
21. Porsolt RD, Le Pichon M, Jalfre M. Depression: a new animal model sensitive to antidepressant treatments. Nature [Internet]. 1977 Apr;266(5604):730–2. Available from: http://www.nature.com/articles/266730a0
22. Krishnan V, Nestler EJ. Animal Models of Depression: Molecular Perspectives. In 2011. p. 121–47. Available from: http://link.springer.com/10.1007/7854_2010_108
23. Moudio S, Willis A, Pytka K, Abulkassim R, Brett RR, Webster JF, et al. Protease-activated receptor 2 activation induces behavioural changes associated with depression-like behaviour through microglial-independent modulation of inflammatory cytokines. Psychopharmacology (Berl) [Internet]. 2022 Jan 9;239(1):229–42. Available from: https://link.springer.com/10.1007/s00213-021-06040-1
24. Abulkassim R, Brett R, MacKenzie SM, Bushell TJ. Proteinase-activated receptor 2 is involved in the behavioural changes associated with sickness behaviour. J Neuroimmunol [Internet]. 2016 Jun;295–296:139–47. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0165572816300935
25. Seibenhener ML, Wooten MC. Use of the Open Field Maze to measure locomotor and anxiety-like behavior in mice. J Vis Exp [Internet]. 2015 Feb 6;(96):e52434. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25742564
26. Becker M, Pinhasov A, Ornoy A. Animal Models of Depression: What Can They Teach Us about the Human Disease? Diagnostics [Internet]. 2021 Jan 14;11(1):123. Available from: https://www.mdpi.com/2075-4418/11/1/123
27. Cathomas F, Hartmann MN, Seifritz E, Pryce CR, Kaiser S. The translational study of apathy—an ecological approach. Front Behav Neurosci [Internet]. 2015 Sep 9;9. Available from: http://journal.frontiersin.org/Article/10.3389/fnbeh.2015.00241/abstract
28. Isingrini E, Camus V, Le Guisquet AM, Pingaud M, Devers S, Belzung C. Association between repeated unpredictable chronic mild stress (UCMS) procedures with a high fat diet: a model of fluoxetine resistance in mice. PLoS One [Internet]. 2010 Apr 28;5(4):e10404. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20436931
29. Yin R, Zhang K, Li Y, Tang Z, Zheng R, Ma Y, et al. Lipopolysaccharide-induced depression-like model in mice: meta-analysis and systematic evaluation. Front Immunol [Internet]. 2023 Jun 8;14. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1181973/full
30. Zhao J, Bi W, Xiao S, Lan X, Cheng X, Zhang J, et al. Neuroinflammation induced by lipopolysaccharide causes cognitive impairment in mice. Sci Rep [Internet]. 2019 Apr 8;9(1):5790. Available from: https://www.nature.com/articles/s41598-019-42286-8
31. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci [Internet]. 2018 Sep 20;21(9):1281–9. Available from: https://www.nature.com/articles/s41593-018-0209-y

Ethical statement

All *in vivo* experimental procedures were in accordance with UK legislation (Animals (Scientific Procedures) Act 1986) and with approval by the University of Strathclyde Ethics Committee. Experiments were performed using 10-12-week old transgenic 5xFAD heterozygous mice ((5xFAD$^{+/-}$ and FAD$^{-/-}$ male and female littermates, (strain: B6SJL-Tg (APPSwFILon, PSEN1*M146L*L286 V) 6799Vas/Mmjax: genetic background: CJ57BL/6J)) and

CJ57BL/6J WT mice, which were group-housed under standard conditions: 21± 2°C, 45–65% humidity, 12hr dark/light cycle (7am-7pm) in MB1 cages (45 × 28 × 13 cm) lined with grade 6 wood bedding, containing a plastic house, tunnel, and nesting material. During week one animals were singly housed in MB1 cages for the purposes of the sucrose preference and regrouped after the last measurement. Animals had free access to food and water with all behavioural procedures carried out between 9am-5pm.

PIL: I01746114, Granted: 08 Nov 22 | 5-year review: 08 Nov 27
PLL: Dr Shuzo Sakata, Strathclyde Institute of Pharmacy & Biomedical Sciences University of Strathclyde 161 Cathedral Street Glasgow G4 0RE. PPL number: PP0688944 | Granted: 24 Dec 20 | Expires: 24 Dec 25

Measuring “Apathy” in Mouse Models of AD

Richard E. Brown[1]*, Emre Fertan[1], Michaela K. Purdon[1], Wyatt Ortibus[1], and Fuat Balci[2]

[1] Department of Psychology and Neuroscience, Dalhousie University, Halifax, Canada. rebrown@dal.ca

[2] Department of Biological Sciences, University of Manitoba, Winnipeg, Canada

Abstract

Apathy, a lack of motivation for goal-directed behaviour, is a symptom of Alzheimer's disease. Mouse models of Alzheimer's disease returned to the start box without entering the goal box or failed to eat the reward in the Hebb Williams Maze, showed reduced motivation to respond on tests of peak interval timing, and stopped responding in tests of reversal learning in the operant olfactometer, suggesting Apathy-like behaviour in these tests.

This research was funded by an NSERC of Canada Research Grant to REB.

What is Apathy?

Apathy is defined as a lack of motivation, involving a reduction in goal-directed behaviour, affecting behavior, thoughts, emotions, and social interactions, which lasts for at least four weeks and causes identifiable functional impairments which are not the result of other factors such as substance abuse or environmental changes [1, 2].

The Neural Basis of Apathy.

Apathy has been attributed to altered functional connections of the neural networks that regulate motivated behaviours. These include the frontostriatal circuits and the dopaminergic projections from the ventral tegmental area (VTA) to the ventral striatum (vStr), with projections to the ventromedial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC) and the anterior cingulate cortex (ACC) and involve the basal ganglia [2, 3] and the amygdala, which has been termed the “hub in the brain network of apathy” [4]. Drugs used to treat apathy include cholinergic (anticholinesterases), glutamatergic (memantine) and dopamine agonist drugs but the 'gold standard' drug has not yet been discovered [5]. It is possible that some of the drugs used to treat different neuropsychiatric disorders may reduce the symptoms of apathy while others exacerbate the symptoms of apathy.

Apathy in Neurological Disorders.

Apathy is a common behavioural disorder in older, cognitively impaired adults, and has been described as a symptom of Parkinson's disease, frontotemporal dementia, Alzheimer's disease, Huntington’s Disease, and other neurodegenerative disorders [3, 6, 7]. Apathy, anxiety and depression are all early symptoms of Alzheimer's Disease [8] and MCI patients with apathy are more likely to develop dementia than those with depression [9]. Apathy is correlated with ABeta deposits and grey matter atrophy in the anterior cingulate and medial fontal cortex [10].

Diagnosis of Apathy.

Because apathy is so closely related to anxiety, depression, anhedonia and other disorders, specific scales have been developed to diagnose apathy and differentiate it from other disorders [11]. For example, the patterns of apathy symptoms may differ between Parkinson's disease, Huntington's disease, frontotemporal dementia, and Alzheimer's Disease. Thus, confusion may exist in trying to develop a single definition for apathy if the phenotypes of apathy depend on the type of neurodegenerative disorder in which it occurs [1]. Diagnosis of apathy thus requires clinicians to distinguish loss of motivation from loss of ability due to cognitive and motor dysfunction [12]. The ABC (Affective, Behavioural, Cognitive) model of apathy allows clinicians to associate different symptoms of apathy with different neuropsychiatric disorders and their underlying neuropathology [13]. However, although apathy is a multi-dimensional disorder, some claim that there is not much evidence for separate behavioural, cognitive and emotional domains [14]. Since apathy is defined as "a quantitative reduction of goal directed behaviour", then "persistence of effort", is the capacity to continue with a task, despite setbacks, difficulty, or fatigue [15]. Persistence of effort can be related to motivational level and been related to neurons of the dorsolateral prefrontal cortex and the frontostriatal reward circuit.

Dissociating apathy from anxiety and depression.

There is a significant overlap in the symptoms of apathy and depression [1], thus considerable effort has gone into differentiating apathy from anxiety and depression [9, 16]. Diagnosis of apathy requires test batteries which create unique diagnostic criteria for apathy [17] and include measures of depression and anxiety [18].

Dissociating apathy from anhedonia.

While apathy is conceptualized as a loss or reduction of motivation, anhedonia refers to a diminished feeling of pleasure in daily activities. Anhedonia occurs when a person becomes “unwilling to derive normal pleasure associated with reward” and is an indication of apathy [19]. Anhedonia may also reflect a loss of motivation to seek pleasure and therefore, like apathy, is a disorder of diminished motivation [3]. "Wanting a reward (anticipation of a reward) appears to be reduced in apathy (lack of motivation or drive), whereas liking (hedonic value) of a reward is reduced in anhedonia and depression [20].

Rodent models of Apathy.

In order to develop rodent (mouse) models of apathy, one must be able to translate the behavioural symptoms of apathy in humans to the homologous behaviours in mice. Since it is not possible to determine whether or not mice experience apathy, the term “apathy-like behavior” is more appropriate in studies of animal models [21]. This requires a definition of “apathy-like behaviour” in mice and Cathomas, et al. [22] define five categories of apathy in humans (self-care, social interaction, exploration, work/education and recreation) and equate these with tests for “apathy-like behaviour” in rodents.

Apathy-like behaviours in mouse models. What is measured?

Some of the most difficult problems in diagnosing apathy in human patients are in dissociating measures of apathy from those of anxiety, depression, anhedonia, and fatigue, as well as controlling for sensory-motor and cognitive deficits. Thus, it is no surprise that the same problems arise in developing measures of apathy-like behaviour in mice. In general, two categories of tests have been used: ethologically-based tests of species-typical behaviour and tests of reward seeking in instrumental conditioning tasks [22]. It is important to use a battery of tests which enable one to dissociate apathy-like behaviour from the measures listed above in order to eliminate confounds in the study of apathy [23].

Ethologically-based tests of species-typical behaviour.

If apathy is defined as a lack of motivation to perform self-generated, voluntary goal-directed behaviors, one can measure species-typical behaviours such as grooming, burrowing, and nest building, as well as spontaneous locomotor activity, exploration of a novel arena, novel object exploration, spontaneous alternation in a T or Y maze, marble burying, and social interaction with conspecifics [24]. The logic of studying species-typical behaviours as indicators of apathy in mice is that they measure motivation to engage in activities of daily living. For example, reductions in grooming [22] and nest building have been used as measures of apathy-like behaviour in L66 tau mouse models of AD [25], and 5xFAD mouse models of AD [26]. Decreased burrowing has been used as a measure of Apathy-like behaviour in 5xFAD mouse models of AD [26]. Spontaneous locomotor activity in the home cage has been used as a measure of apathy in 3xTg-AD mice [27]. Jackson et al. [28] used exploration of a novel arena to measure apathy-like behaviour in aged C57BL/6J mice and found that aged mice covered less distance and had a slower mean velocity than younger mice. However, this could be due to deficits in motor behaviour in elderly mice [29]. The 5xFAD mice showed deficits in marble burying and spontaneous alternation in a Y maze, which Keszycki et al. [26] interpreted as evidence for increased apathy-like behaviour.

Tests of reward seeking in instrumental conditioning tasks.

Evaluating operant behaviour in rodents is used for the study of apathy, because it allows an evaluation of motivational processes involved in goal directed behaviour [30]. Given the uncertainty in interpreting the results of deficits in ethologically-based tests of species-typical behaviour as "apathy-like" behaviours, it may be more efficacious to focus on the results of instrumental learning studies which examine deficits in goal-directed behaviour, reduced motivation to work for a reward and/or a disinterest in the reward once it is achieved [21, 28].

Measures of motivation in operant tasks include the latency to first response (lever press or nose-poke), the time taken to complete responses and the interval between responses. These latency measures are prolonged if motivation is impaired [21]**.**

Motivated behaviour in an instrumental learning task can be divided into appetitive and consummatory behaviour. Appetitive behaviour involves seeking a reward such as food and can be considered a goal-directed instrumental (seeking) behaviour which involves effort-based choice, persistence and vigour of response [2]**.** Consummatory behaviour involves eating the food reward. Measures of motivation (appetitive behaviour) in an instrumental task include the latency to first response and the time taken to complete the responses. Trials in which the mouse fails to earn the reward are either omission trials (no lever press) or failure trials (task started but not completed). Therefore, prolonged time to complete the task or increased number of failure/omission trials reflect apathy-like behavior in mice because these changes cause the reduction of goal-directed behavior [21, 31]**.**

In an operant chamber (Skinner box), different schedules of reinforcement can be used to evaluate the animal's motivation to work for a reward. For example, on a progressive fixed-ratio schedule for an evaporated milk reward in an operant chamber, two mouse models of HD (BAC HD and z_Q175 KI mice) had reduced response rates compared to WT mice, indicating a reduced motivation to work for a food reward or a deficit in consuming the reward [32]. In an operant lever pressing task for sucrose reward to study goal directed behaviou, the hAPP-J20 mouse model of AD showed an initial impairment in goal-directed behaviour which was reduced by extra training [33]. This procedure might be used to study apathy-like behaviour in other mouse models of AD.

Tests of effort-based decision making (EBDM).

Studies of effort-based decision-making offer rodents a choice between a preferred reinforcer (eg, high-carbohydrate pellets) that can be obtained only with a high-effort instrumental action versus a less preferred option (eg, standard lab chow) which requires less effort to obtain [21]. Effort-based decision making and the instrumental behaviours involved in appetitive behaviour can thus be used to study apathy, while consummatory behaviour might be used as a measure of anhedonia. Increased latency measures to perform in Fixed Ratio (FR) or Progressive Ratio (PR) tasks for reward in an operant chamber indicate reduced motivation, which has been interpreted as "apathy-like behaviour" [21]. Aged C57BL/6J mice showed reduced motivation for reward in progressive ratio and Effort for Reward tasks in an operant chamber, suggesting that these mice showed apathy-like behaviour [28].

When using a T-maze to study effort-based choice procedures, one arm of the maze contains a higher density of food reward that can be obtained only by climbing a barrier, whereas the alternative choice is to access an arm that contains less food, but has no or a lower barrier [2]. Thus, one can measure appetitive behaviour (how much work an animal will do to reach low versus high value rewards, how long an animal will persist in getting a reward) and consummatory behaviour (is the reward eaten). Measures can also be made on the rate of learning to reach a reward, and the memory for the reward [2].

Dissociating apathy from anxiety, depression, anhedonia, as well as sensory, motor and cognitive deficits.

Dissociating measures of apathy from anxiety and depression.

Apathy has been measured by the tail suspension test and the forced swimming test [34, 37], but these are generally considered tests for depression-like behaviour [38] and may not be relevant for apathy because they do not reflect a lack of motivation but rather resignation to a stressful situation. Jackson and Robinson [11] argue that because major depression and apathy share many symptoms, they are difficult to dissociate in mouse models and suggest that apathy and depression can best be dissociated by measuring behaviours in unpleasant or aversive situations, such as exaggerated emotional responses to aversive contexts or situations such as anxiety and fear which are related to negative valence (NV). Anxiety in mice is often measured in the open field, elevated plus maze and light-dark box [39]. Taken together, this suggests that a test battery that measures appetitive behaviour for rewards (apathy), sucrose preference (anhedonia), depression and anxiety may be required to dissociate apathy-like behaviour from anhedonia, anxiety and depression.

Dissociating measures of apathy from anhedonia.

Apathy-like behaviour has been shown in mice in the saccharin preference task [34] and the sucrose preference test [28] but these are used as tests for anhedonia. A reduction in sucrose preference has been associated with the emotional blunting domain of apathy [11]. For example, a reduced reduced sucrose preference in L66 tau mouse models of AD has been interpreted as an example of anhedonia in these mice [25]. However, the sucrose preference task may be more complex than it seems and the reduction in sucrose preference may not be a reliable measure of anhedonia [35]. Another way of measuring anhedonia in mice is by measuring facial expressions but this also has its drawbacks [36]. Husain and Roiser [3] suggest that apathy and anhedonia can be dissociated by analyzing appetitive and consummatory behaviours in instrumental learning tasks. Those tasks measuring motivation can be used to measure apathy-like behaviour, while measuring consummatory responses can be used to measure anhedonia.

Dissociating apathy from sensory, motor and cognitive deficits.

If animals perform poorly in instrumental learning tasks, it could be due to sensory, motor or motivational impairments. While impairments in motivation (apathy-like behaviour) can be dissociated from learning and memory impairments in instrumental learning tasks (see below), the issue of sensori-motor impairment is more difficult to interpret: A lack of motivation to explore may be a symptom of apathy, but a lack of ability to explore may be due to sensori-motor dysfunction [29]. Sensory deficits are very common in mice. Many strains have visual deficits [40]; auditory deficits [41]; olfactory deficits [42]; deficits in somatosensory perception [43]; or deficits in pain perception [44] and these deficits may be interpreted as apathy-like behaviours, so must be controlled [23].

Three experiments indicating apathy-like behaviour in our mice.

Apathy-like behaviour of the 3xTg AD mice in the Hebb-Wiliams maze.

Apathy-like behaviour has been documented in both the 5xFAD [26] and 3xTg-AD mouse models of AD [45]**.** With beta-amyloid plaque and tau tangle related mutations, the 3xTg-AD mice are one of the most commonly used models of AD with high face validity [46]. They have been used to study the cognitive [47], motor [48], and sensory [43] deficits in AD. All test procedures conducted in our laboratory were approved by the Dalhousie University Committee on Laboratory Animals. The Hebb–Williams maze, as described by Stanford and Brown [49], consisted of a 60x60 cm box made of black Plexiglas, with 10 cm high walls and was covered with clear, removable Plexiglas lid. The floor was divided into 36 squares (10x10 cm) with white lines that were used to place the barriers and score the errors made by the mice. The start box and the goal box (10 cm wide x 20 cm long) were located at opposite corners and were covered with clear Plexiglas lids. A set of barriers that were 10 cm tall with a 2.5 cm base and lengths of 10, 20, 25, 30, 40 and 50 cm, were used to construct the different problem designs.

The study consisted of three phases: habituation, acquisition, and test. Mice were food deprived to ~80% of their ad libitum weight and were weighed before testing each day and fed after testing. During habituation, mice were allowed to explore the maze for 20 min/day for 4 days with no barriers present and a single Froot Loop in the goal box. During acquisition, mice were given six practice mazes (A, B, C, D, E, F) shown in Fig. 1B for nine trials per day and the time to reach the goal box was recorded, as well as occurrences of returning to the start box without entering the goal box and entering the goal box without eating the food reward. If mice did not enter the goal box within 120 sec, the trial was stopped, and the mouse placed back in the start box without receiving the reward. The criterion for completing the acquisition phase was that the mice completed all nine trials (in total) in under 60 s [50]. In the test phase, mice were given 12 problems (Fig. 1C) over six days, with five trials for each problem. The latency to reach the goal box and the number of errors made were recorded on each trial. An error was scored when all 4 paws of the mouse were in an error zone, as indicated by the broken lines in Fig. 1C. Based on previous results [49], these tasks were graded in difficulty (easy, medium, and difficult; Fig. 1C).

Female 3xTg-AD and WT mice performed significantly worse than males in trials assessing long-term memory [50]. Moreover, female mice often did not consume the food reward upon finding it in the goal box. Eating the reward correlated with better learning and memory performance, indicating that the female mice had deficits in both appetitive and consummatory behaviour in responding to rewards in this test, suggesting possible apathy-like behaviour and anhedonia [50].

Apathy-like behaviour of the 3xTg AD mice in the timing study.
In an experiment designed to study interval-timing in the 3xTg-AD mice [51] the peak interval (PI) procedure was utilised to evaluate multiple components of timing ability [52]. The experiments were done using a nine-hole box (Cambridge Cognition Ltd., England) which had a grid floor with a removable reinforcement tray and a was placed in a MDF sound/light attenuating box which had a camera on the top. The reinforcement tray contained a lick tube attached to a peristaltic pump that delivered liquid reinforcement (0.025 ml/sec of 5% sucrose solution). Only three of the nine holes were used (holes 4, 5, 6) and were equipped with infrared beams to detect nose pokes. All nose poke response holes and the reinforcement tray had lights that were controlled by the Cambridge Cognition Control computer software which also recorded timestamped nose pokes.

Mice were water deprived starting five days before the experimental sessions began and were kept at 85% of their ad libitum body weights throughout the experiment. Laboratory chow was available at all times in the home cages. The experiment had three phases: magazine training, fixed interval (FI) training, and peak interval (PI) testing. There were two 20-min session of magazine training in which the reinforcement tray was lit at all times and 0.7 mL of sucrose water was delivered to the drinking spout every 40 s. Following magazine training, mice were given four sessions of FI-15 s training during which the first nose poke into the central hole after 15 s from the onset of the conditioned stimulus (i.e., active hole light) was followed by 0.03 mL sucrose water delivery to the drinking spout. The inter-trial interval was 20 s ±10 s. Peak interval (PI) testing sessions included PI probe trials intermixed with FI trials (2:1 ratio of FI to PI trials). PI probe trials were not reinforced. Data were recorded for 39 sessions in this phase. In these probe trials, mice typically cluster their responses around the target interval, namely the latency to reinforcement availability. The core timing performance is characterized by timing accuracy (e.g., peak location with respect to delay to the reward) and timing precision (e.g., the width of the response curve). Several aspects of anticipatory timed responses are sensitive to the motivational state of the subjects [53]. Hunger level of the subjects affects the rate of responding and thus the amplitude of response curve and the peak amplitude typically decreases with longer delays to reward (lower expected reward rate) [54].

The timing of response initiation and the rate of responding are best captured by analyzing data for individual trials [53]. When the performance of individual animals is examined, rather than the average response curves, differences in patterns of responding can be determined by examining start times and stop times. Considering such response patterns, differences in expected reward magnitude and reward devaluation led to earlier start times without a clear effect on the stop times [55]. In line with these observations, Gür et al. [51] found that while the 3xTg-AD mice did not show deficits in their core timing ability compared to wild-type controls, the results revealed a prominent sex difference in the measures of timing behaviour that reflect motivational deficits, as the female mice initiated their timed anticipatory responses (i.e., start times) later in the trial and emitted them at a lower rate than male mice regardless of genotype [51]. These findings emphasize the importance of analyzing the motivational aspect of timing behaviour.

Apathy-like behaviour in reversal learning in the operant olfactometer.
Mice were trained in an operant olfactometer to lick a tube for sucrose reward to odour A (CS+) and not for odour B (CS-) to a criterion of 85% correct and then given reversal training using the procedure of Roddick et al. [56] During reversal learning, mice go through four phases of responding as measured by signal detection theory: perseverance in responding to the original CS+ odour (now the CS-), responding to neither odour, responding to both odours and finally responding to the new CS+ odour [57]. In reversal learning, the formerly rewarded responses to Odour A are no longer rewarded and, as a result, mice often stop responding to both odours, behaviour which Amsel [58] labelled "frustrative non reward". In this case, the reduced motivation to respond to either stimulus may qualify as "apathy-like behaviour" as the goal-directed behaviour has been halted. This suggests that "frustrative non-reward" may be an apathy-like behaviour.

Validity and reliability of mouse models of Apathy-like behaviour.
How can we evaluate the validity and reliability of mouse models of "apathy-like behaviour"? Animal models of human neuropsychiatric disorders should have high face validity, construct validity, and predictive validity [59].

Face validity.
Face validity is the degree of descriptive similarity between the behavioural dysfunction seen in an animal model and the human neurobehavioral disorder. Thus, one constructs a list of symptoms in Human and mouse models and compares them [22]. While face validity is an important criterion for model evaluation, the strong emphasis on this criterion has been criticized [59]. One of the problems in developing animal models of apathy is to determine the face validity of the model: does the mouse behavioural phenotype match that of the symptoms of apathy in humans? In this case, the study of goal-directed behaviour may have greater face value than the study of species-typical behaviour.

Construct validity.
High construct validity occurs when the animal model shares the same mechanisms (neuropathology or environmental situation) underlying the symptoms of the disorder as human patients. Construct validity is the most important criterion for animal models because it addresses the soundness of the theory underlying the model, and because it provides the framework for interpreting data generated by the model [59]**.** The problem in determining construct validity is that we do not know the exact neuropathology in humans showing apathy. There are many indications of the neural network and neurochemical dysfunctions causing apathy (see above) but no definitive knowledge. Thus, construct validity is impossible to determine, as the genetic, neuroanatomical and environmental causes of apathy are unknown. It has been proposed that the neural networks that regulate motivated behaviour, including the frontostriatal circuits, medial prefrontal cortex, basal ganglia, and amygdala underlie apathy [2, 3, 4] but no definitive "apathy-circuit" has been identified.

Predictive validity.
An animal model with high predictive validity allows extrapolation of the effect of a particular experimental manipulation from the animal model to humans, and from the laboratory to the "real world" [59]. Predictive validity is particularly important in psycho-pharmacology where the effects of new drugs with animal models must translate to humans [60]. Predictive validity refers to the ability of a drug screening or an animal model to correctly identify the efficacy of a putative therapeutic, but there are often failures in this as has occurred in the development of new drugs for AD [61]. The predictive validity of animal models of apathy may depend on their use in the development of new drugs for its treatment [5].

Reliability and replicability of animal models of apathy.
Reliability measures the extent to the same results can be obtained in different studies using the same animal models in the same experimental conditions. Replicability or reproducibility is the degree of accordance between the results of the same experiment performed independently in the same or different laboratories [59]. There have been few studies which have attempted to measure the replicability of studies of mouse models of apathy. The symposium is a first attempt to do this.

Summary and conclusions

This paper has attempted to summarize the mouse behavioural phenotypes that have been proposed for the study of apathy. Three experiments from our lab examined apathy in mouse models in studies of instrumental learning for food reward: the Hebb-Williams maze, the operant timing task and the operant olfactometer. In each case, mice showed evidence of reduced motivation to seek the reward and there was some evidence that female mice and neurexin1+/- mouse models showed a decrease in motivation to seek the reward, an increase in apathy-like behaviour. The main problem in defining apathy in both humans and mouse models is to dissociate the behavioural phenotypes and neuro-chemical pathways underlying apathy from those responsible for anxiety, depression, and anhedonia.

The results of our studies on the Hebb-Williams maze, the timing task and the operant olfactometer suggest that these tasks could be used to dissociate motivational from cognitive factors in instrumental learning tasks. By measuring the patterns of responses of the mice in these tasks, we were able to distinguish learning and memory deficits from apathy-like behaviours. This is significant because motivation and learning are regulated by different brain mechanisms and the use of double dissociation studies might be able to separate the underlying neural networks. This problem becomes significant given the involvement of apathy in many neurological disorders such as Alzheimer’s disease [10] and other neurodegenerative disorders [3, 6] in which motivational factors might be confounded with learning and memory disorders.

References

1. Fahed, M., Steffens, D.C. (2021). Apathy: neurobiology, assessment and treatment. *Clin. Psychopharmacology Neurosci.* **19**, 181-189. doi:10.9758/cpn.2021.19.2.181.

2. Le Heron, C., Holroyd, C.B., Salamone, J., Husain, M. (2019). Brain mechanisms underlying apathy. *J. Neurol. Neurosurg. Psychiatry* **90**, 302-312. doi:10.1136/jnnp-2018-318265.

3. Husain, M., Roiser, J.P. (2018). Neuroscience of apathy and anhedonia: a transdiagnostic approach. *Nat. Rev. Neurosci*. **19**, 470-484. doi:10.1038/s41583-018-0029-9.

4. Zeng, N., Aleman, A., Liao, C., Fang, H., Xu, P., Luo, Y. (2023). Role of the amygdala in disrupted integration and effective connectivity of cortico-subcortical networks in apathy. *Cereb. Cortex* **33**, 3171-3180. doi:10.1093/cercor/bhac267.

5. Costello, H., Husain, M., Roiser, J.P. (2024). Apathy and motivation: Biological basis and drug treatment. *Ann. Rev. Pharmacol. Toxicol.* **64**, 14.1-14.26. doi:10.1146/annurev-pharmtox-022423-014645.

6. Lanctôt, K.L., Agüera-Ortiz, L., Brodaty, H., Francis, P..T, Geda, Y.E., Ismail, Z., Marshall,. GA, Mortby, M.E., et al. (2170). Apathy associated with neurocognitive disorders: Recent progress and future directions. *Alzheimer's Dement.* **13**, 84-100. doi:10.1016/j.jalz.2016.05.008.

7. Dolphin, H., Dyer, A..H, McHale, C., O'Dowd, S., Kennelly, S.P. (2023). An update on apathy in Alzheimer's disease. *Geriatrics (Basel)* **8**, 75. doi:10.3390/geriatrics8040075.

8. Johansson, M., Stomrud, E., Lindberg, O., Westman, E., Johansson, P.M., van Westen, D., Mattsson, N., Hansson, O. (2020). Apathy and anxiety are early markers of Alzheimer's disease. *Neurobiol. Aging* **85**, 74-82. doi:10.1016/j.neurobiolaging.2019.10.008.

9. Ma, L. (2020). Depression, anxiety, and apathy in mild cognitive impairment: Current perspectives. *Front. Aging Neurosci.* **12**, 9. doi:10.3389/fnagi.2020.00009.

10. Nobis, L., Husain, M. (2018) Apathy in Alzheimer’s disease. *Curr. Opin. Behav. Sci*. **22**, 7–13. doi10.1016/j.cobeha.2017.12.007.

11. Jackson, M.G., Robinson, E.S.J. (2022). The importance of a multidimensional approach to the preclinical study of major depressive disorder and apathy. *Emerg. Top. Life Sci.* **6**, 479-489. doi: 10.1042/ETLS20220004.

12. Landes, A.M., Sperry, S.D., Strauss, M.E., Geldmacher, D.S. (2001). Apathy in Alzheimer's disease. *J Am Geriatr Soc*. **49**, 1700-1707. doi10.1046/j.1532-5415.2001.49282.x.

13. Kumfor, F., Zhen, A., Hodges, J.R., Piguet, O., Irish, M. (2018). Apathy in Alzheimer's disease and frontotemporal dementia: Distinct clinical profiles and neural correlates. *Cortex* **103**, 350-359. doi: 10.1016/j.cortex.2018.03.019.

14. Dickson, S.S., Husain, M. (2022). Are there distinct dimensions of apathy? The argument for reappraisal. *Cortex* **149**, 246-256. doi:10.1016/j.cortex.2022.01.001.

15. Dalléry, R., Saleh, Y., Manohar, S., Husain, M. (2023). Persistence of effort in apathy. *Revue neurologique* **179**, 1047–1060. doi:10.1016/j.neurol.2023.03.017.

16. Ineichen, C., Baumann-Vogel, H. (2021). Deconstructing apathy in Parkinson's disease: Challenges in isolating core components of apathy from depression, anxiety, and fatigue. *Front Neurol.* **12**, 720921. doi:10.3389/fneur.2021.720921.

17. Miller, D.S., Robert, P., Ereshefsky, L., Adler, L., Bateman, D., Cummings, J., DeKosky, S.T., Fischer, C.E., Husain, M., et al. (2021). Diagnostic criteria for apathy in neurocognitive disorders. *Alzheimer's Dement.* **17**, 1892-1904. doi10.1002/alz.12358.

18. Robert, P. H., Mulin, E., Malléa, P., David, R. (2010). Review: Apathy diagnosis, assessment, and treatment in Alzheimer's disease. *CNS Neurosci. Ther.* **16**, 263-71. doi:10.1111/j.1755-5949.2009.00132.x.

19. Milton, L.K., Oldfield, B..J, Foldi, C.J. (2018). Evaluating anhedonia in the activity-based anorexia (ABA) rat model. *Physiol Behav*. **194**, 324-332. doi:10.1016/j.physbeh.2018.06.023.

20. Simon, J.J., Biller, A., Walther, S., Roesch-Ely, D., Stippich, C., Weisbrod, M., Kaiser, S. (2010). Neural correlates of reward processing in schizophrenia--relationship to apathy and depression. *Schizophr. Res.* **118**, 154-61. doi:10.1016/j.schres.2009.11.007.

21. Tanaka, K.F., Hamaguchi, T. (2019). Translational approach to apathy-like behavior in mice: From the practical point of view. *Psychiatry Clin. Neurosci.* **73**, 685-689. doi:10.1111/pcn.12915.

22. Cathomas, F., Hartmann, M.N., Seifritz, E., Pryce, C.R., Kaiser, S. (2015). The translational study of apathy-an ecological approach. *Front. Behav. Neurosci*. **9**, 241. doi:10.3389/fnbeh.2015.00241.

23. Schellinck, H.M., Cyr, D.P., Brown, R.E. (2010). How many ways can mouse behavioral experiments go wrong? Confounding variables in mouse models of neurodegenerative diseases and how to control them. *Advances in the Study of Behavior* **41**, *255-366.*

24. Si, Y., Guo, C., Xiao, F., Mei, B., Meng, B. (2022). Noncognitive species-typical and home-cage behavioral alterations in conditional presenilin 1/presenilin 2 double knockout mice. *Behav. Brain Res.* **418**, 113652. doi:10.1016/j.bbr.2021.113652.

25. Robinson, L., Dreesen, E., Mondesir, M., Harrington, C., Wischik, C., Riedel, G. (2024). Apathy-like behaviour in tau mouse models of Alzheimer's disease and frontotemporal dementia. *Behav. Brain Res.* **456**, 114707. doi:10.1016/j.bbr.2023.114707.

26. Keszycki, R., Rodriguez, G., Dunn, J.T., Locci, A., Orellana, H., Haupfear, I., Dominguez, S., Fisher, D.W., Dong, H. (2023). Characterization of apathy-like behaviors in the 5xFAD mouse model of Alzheimer's disease. *Neurobiol. Aging* **126**, 113–122. doi:10.1016/j.neurobiolaging.2023.02.012

27. Pardossi-Piquard, R., Lauritzen, I., Bauer, C., Sacco, G., Robert, P., Checler, F. (2016). Influence of genetic background on apathy-like behavior in triple transgenic AD mice. *Curr. Alzheimer Res.* **13**, 942–949. doi:10.2174/1567205013666160404120106.

28. Jackson, M.G., Lightman, S.L., Gilmour, G., Marston, H., Robinson, E.S.J. (2021). Evidence for deficits in behavioural and physiological responses in aged mice relevant to the psychiatric symptom of apathy. *Brain Neurosci. Adv.* **5**, 23982128211015110. doi:10.1177/23982128211015110

29. O'Leary, T.P., Mantolino, H.M., Stover, K., Brown, R.E. (2020). Age-related deterioration of motor function in male and female 5xFAD mice from 3-16 months of age. *Genes Brain Behav.* **19**, e12538. doi.org/10.1111/gbb.12538.

30. Magnard, R., Vachez, Y., Carcenac, C., Krack, P., David, O., Savasta, M., Boulet, S., Carnicella, S. (2016). What can rodent models tell us about apathy and associated neuropsychiatric symptoms in Parkinson's disease? *Transl. Psychiatry* **6**, e753. doi:10.1038/tp.2016.17

31. Yoshida, K., Drew, M.R., Mimura, M., Tanaka, K.F. (2019). Serotonin-mediated inhibition of ventral hippocampus is required for sustained goal-directed behaviour. *Nat. Neurosci.* **22**, 770–777. doi:10.1038/s41593-019-0376-5

32. Oakeshott, S., Port, R., Cummins-Sutphen, J., Berger, J., Watson-Johnson, J., Ramboz, S., Paterson, N. et al. (2012). A mixed fixed ratio/progressive ratio procedure reveals an apathy phenotype in the BAC HD and the z_Q175 KI mouse models of Huntington's disease. *PLoS Curr.* **4**, e4f972cffe82c0. doi:10.1371/4f972cffe82c0.

33. Dhungana, A., Becchi, S., Leake, J., Morris, G., Avgan, N., Balleine, B.W., Vissel, B., Bradfield,. LA. (2023). Goal-directed action is initially impaired in a hAPP-J20 mouse model of Alzheimer's disease. *eNeuro* **10**, ENEURO.0363-22.2023. doi:10.1523/ENEURO.0363-22.2023.

34. Baumann, A., Moreira, C.G., Morawska, M.M., Masneuf, S., Baumann, C.R., Noain, D. (2016). Preliminary evidence of apathetic-like behavior in aged vesicular monoamine transporter 2 deficient mice. *Front. Hum. Neurosci.* **10**, 587. doi:10.3389/fnhum.2016.00587

35. Verharen, J.P.H., de Jong, J.W., Zhu, Y., Lammel, S. (2023). A computational analysis of mouse behavior in the sucrose preference test. *Nat. Comm.* **14**, 2419. doi:10.1038/s41467-023-38028-0

36. Scheggi, S., De Montis, M.G., Gambarana, C. (2018). Making sense of rodent models of anhedonia. *Int. J. Neuropsychopharmacol.* **21**, 1049-1065. doi:10.1093/ijnp/pyy083.

37. Vernay, A., Sellal, F., René, F. (2016). Evaluating behavior in mouse models of the behavioral variant of frontotemporal dementia: Which test for which symptom? *Neurodegener. Dis.* **16**, 127–139. doi:10.1159/000439253.

38. Martin, A.L., Brown, R.E. (2010). The lonely mouse: verification of a separation-induced model of depression in female mice. *Behav. Brain Res.* 207, 196–207. doi:10.1016/j.bbr.2009.10.006

39. O'Leary, T.P., Gunn, R.K., Brown, R.E. (2013). What are we measuring when we test strain differences in anxiety in mice? *Behav. Gen.* **43**, 34-50.

40. Wong, A.A. Brown, R.E. (2006). Visual detection, pattern discrimination and visual acuity in 14 strains of mice. *Genes Brain Behav.* **5**, 389-403.

41. O'Leary, T.P., Shin, S., Fertan, E., Dingle, R.N., Almuklass, A., Gunn, R.K., Yu, Z., Wang, J., Brown, R.E. (2017). Reduced acoustic startle response and peripheral hearing loss in the 5XFAD mouse model of Alzheimer's disease. *Genes Brain Behav.* **16**, 554-563.

42. Roddick, K..M, Roberts, A.D., Schellinck, H.S., Brown,. RE. (2016). Sex and genotype differences in odour detection in the 3xTg-AD and 5XFAD mouse models of Alzheimer's disease at 6 months of age. *Chemical Senses* **41**, 433–440.

43. Simanaviciute, U., Ahmed, J., Brown, R.E., Connor-Robson, N., Farr, T.D., Fertan, E., Gambles,, N., Garland, H., et al. (2020). Recommendations for measuring whisker movements and locomotion in mice with sensory, motor, and cognitive deficits. *J. Neurosci. Methods* **331**, 108532.

44. Mogil, J.S., Ritchie, J., Sotocinal, S.G., Smith, S.B., Croteau, S., Levitin, D.J., Naumova, A.K. (2006). Screening for pain phenotypes: analysis of three congenic mouse strains on a battery of nine nociceptive assays. *Pain* **126**, 24-34. doi:10.1016/j.pain.2006.06.004.

45. Bourgeois, A., Lauritzen, I., Lorivel, T., Bauer, C., Checler, F., Pardossi-Piquard, R. (2018). Intraneuronal accumulation of C99 contributes to synaptic alterations, apathy-like behavior, and spatial learning deficits in 3×TgAD and 2×TgAD mice. *Neurobiol. Aging* **71**, 21–31. doi:10.1016/j.neurobiolaging.2018.06.038

46. Oddo, S., Caccamo, A., Shepherd, J.D., Murphy, M.P., Golde, T.E., Kayed, R., Metherate, R., Mattson, M.P., Akbari, Y., LaFerla,. FM. (2003). Triple-transgenic model of Alzheimer's disease with plaques and tangles: intracellular Aβ and synaptic dysfunction. *Neuron* **39**, 409–421. https://doi.org/10.1016/S0896-6273(03) 00434-3.

47. Stevens, L.M., Brown, R.E. (2015). Reference and working memory deficits in the 3xTg-AD mouse between 2 and 15 months of age: A cross-sectional study. *Behav. Brain Res.* **278**, 496-505.

48. Garvock-de Montbrun, T., Fertan, E., Stover, K., Brown, R.E. (2019). Motor deficits in 16 month-old male and female 3xTg-AD mice. *Behav. Brain Res.* **356**, 305-313. doi:10.1016/j.bbr.2018.09.006.

49. Stanford, L., Brown, R.E. (2003). MHC-congenic mice (C57BL/6J and B6-H-2-K) show differences in speed but not accuracy in learning the Hebb-Williams maze. *Behav. Brain Res.* **144**, 187-197.

50. Fertan, E., Wong, A.A., Vienneau, N.A., Brown, R.E. (2019). Age and sex differences in motivation and spatial working memory in 3xTg-AD mice in the Hebb–Williams maze. *Behav. Brain Res.* **370**, 111937. doi:10.1016/j.bbr.2019.111937.

51. Gür, E., Fertan, E., Kosel, F., Wong, A.A., Balcı, F., Brown, R.E. (2019). Sex differences in the timing behavior performance of 3xTg-AD and wild-type mice in the peak interval procedure. *Behav. Brain Res.* **360**, 235-243. doi:10.1016/j.bbr.2018.11.047.

52. Balci, F., Gallistel, C.R., Allen, B.D., Frank, K.M., Gibson, J.M., Brunner, D. (2009). Acquisition of peak responding: what is learned? *Behav. Processes* **80**, 67-75. doi:10.1016/j.beproc.2008.09.010.

53. Balcı, F. (2014). Interval timing, dopamine, and motivation. *Timing Time Percept.* **2**, 379-410. doi:10.1163/22134468-00002035.

54. Galtress, T., Kirkpatrick, K. (2010). Reward magnitude effects on temporal discrimination. *Learn. Motiv*. **41**, 108-124. doi:10.1016/j.lmot.2010.01.002.

55. Galtress, T., Marshall, A.T., Kirkpatrick, K. (2012). Motivation and timing: clues for modeling the reward system. *Behav. Processes* **90**, 142-53. doi:10.1016/j.beproc.2012.02.014.

56. Roddick, K.M., Schellinck, H.M., Brown, R.E. (2023). Serial reversal learning in an olfactory discrimination task in 3xTg-AD Mice. *Learn. Mem.* **30**, 310-319. doi:10.1101/lm.053840.123

57. Ortibus, W., Roddick, K., Brown, R.E. (2023). Olfactory discrimination tasks and reversal learning in a neurexin1 (+/-) mouse model of autism spectrum disorder. Manuscript in progress.

58. Amsel, A. (1962). Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension. *Psych. Rev.* **69**, 306-328.

59. van der Staay, F.J., Arndt, S..S, Nordquist, R.E. (2009). Evaluation of animal models of neurobehavioral disorders. *Behav. Brain Funct.* **5**, 11 doi:10.1186/1744-9081

60. Banik, A., Brown, R.E., Bamburg, J., Lahiri, D.K., Khurana, D., Friedland, R.P., et. al., (2015). Translation of pre-clinical studies into clinical trials for Alzheimer's disease: What are the roadblocks and how can they be overcome? *J. Alz. Dis.* **47**, 815-843.

61. Kim, C.K., Lee, Y.R., Ong, L., Gold, M., Kalali, A., Sarkar, J. (2022). Alzheimer's disease: Key insights from two decades of clinical trial failures. *J. Alz. Dis.* **87**, 83-100. doi:10.3233/JAD-215699.

Modeling Aspects of Apathy in Rodents using Effort-based Choice Procedures

J.D. Salamone [1], A. Ecevitoglu[1], G. Edelstein[1], M. Correa[2].

[1]**Behavioral Neuroscience Div, Psychological Sciences, University of Connecticut, Storrs, CT, USA**

[2] **Area de Psicobiologia, Universitat Jaume I, Castelló, Spain**

Introduction

Apathy is a complex symptom in humans, with multiple components including emotional flattening and a lack of goal directed behavior. Effort-based choice tasks represent one way of assessing the exertion of effort in goal-directed activity. Effort-based choice is studied using procedures that offer choices between high effort options leading to highly valued reinforcers vs. low effort/less preferred options.

Behavioural Tasks: Effort-based Choice Behavior

Our laboratory has developed several behavioral tasks in rodents that assess the effects of pharmacological manipulations on exertion of physical effort and effort-related choice. We developed a T-maze procedure [6], in which a barrier is placed in the arm with a higher density of reward to provide an effort-related challenge. This procedure can be run with multiple variants (Figure 1), including conditions in which the barrier arm contains four 45 mg food pellets and the no-barrier arm contains two pellets, as well as conditions in which there is no barrier in either arm, or there is no reinforcement in the no-barrier arm so the only way to get food is to climb the barrier. We also developed operant behavior procedures that offer rats a choice between lever pressing to obtain a relatively preferred food (high carbohydrate pellets), vs. approaching and consuming a less preferred food (lab chow) that is concurrently available [5]. One such task is the concurrent fixed ratio (FR) 5/chow feeding choice procedure. Under baseline conditions, rats typically get most of their food by FR5 lever pressing, and eat only small amounts of chow. Another task we have employed is a progressive ratio (PROG)/chow feeding concurrent choice task [2,3], which is a variant of the lever pressing/chow intake choice procedure described above, but instead uses a PROG lever pressing requirement rather than a FR5 schedule. The PROG/chow feeding procedure offers the choice of lever pressing on a PROG schedule reinforced by the preferred high carbohydrate pellets vs. approaching and consuming the less preferred chow. The PROG schedule requires that the rat repeatedly make within-session choices between lever pressing and chow intake under conditions in which the ratio requirement is gradually incrementing. Furthermore, having a choice between responding on the PROG schedule vs. approaching and consuming concurrently available chow allows for the determination of whether or not a manipulation is selectively affecting selection of lever pressing or is more broadly affecting food motivation.

Pharmacological Studies of Effort-based Choice

Using drug probes to tease apart various aspects of these tasks, we have shown that interference with dopamine (DA) transmission using DA antagonists and the DA depleting agent tetrabenazine (TBZ) shifts effort-based choice and induces a low-effort bias (i.e., decreased selection of the high-effort options of barrier climbing and lever pressing; [1, 2, 3, 5, 6, 8, 9]). TBZ has been the focus of much current research because this drug induces motivational symptoms such as apathy and fatigue in humans. In rats tested with the T-maze barrier task, TBZ reduced selection of the barrier arm but had no effect on choice when there was no barrier in either arm, or when the only way to obtain food reinforcement was to climb the barrier [9]. Drug treatments that produce the shift in choice behavior did not alter food intake or preference in free-feeding choice tests [4, 5]. Although DA antagonists reduce FR5 lever pressing and increase chow intake, reinforcer devaluation by pre-feeding [5] and appetite suppressant drugs do not increase chow intake at doses that suppress lever pressing [2, 3, 8]. Furthermore, a dose of TBZ that shifts choice behavior did not reduce binge-like eating of a highly palatable food (i.e., Cadbury's chocolate; [7]).

In a recent study using the FR5/chow feeding choice task [5], detailed timing of lever pressing was monitored with an event recording system, and the temporal characteristics of operant behavior seen after 1.0 mg/kg TBZ or vehicle injections in rats were analyzed using MatLab programs. TBZ shifted effort-based choice, decreasing lever pressing while concurrently increasing chow consumption. There was a very robust increase in the total duration of pauses in responding, although TBZ did not specifically increase the duration of post-reinforcement pauses. TBZ increased time spent feeding and the number and duration of feeding bouts, but did not affect feeding rate (grams/minute spent feeding). Importantly, the main effect of TBZ was to shift time allocation; rats showed a significant decrease in time spent lever pressing, and an increase in time spent feeding, but there was no significant effect on total time spent lever pressing for pellets and consuming chow. Thus, TBZ predominantly affected the relative allocation of lever pressing vs. chow consumption, with little alteration in consummatory motor acts involved in chow intake. TBZ is being used to model motivational symptoms in psychopathology such as anergia, fatigue and apathy, and these effects in rats could have implications for psychiatric research.

Conclusions

Rat models involving assessments of effort-based choice are potentially useful for assessing the effects of novel drug treatments for apathy and other motivational dysfunctions [10]. Reversal of the effects of TBZ in rats tested in the T-maze or FR5/chow feeding choice task can be used as a measure of the ability of a drug to reverse an impairment [7,9]. In contrast, a test such as the progressive ratio/chow feeding choice task can be useful for assessing the ability of a drug to increase selection of high-effort lever pressing when administered alone [7].

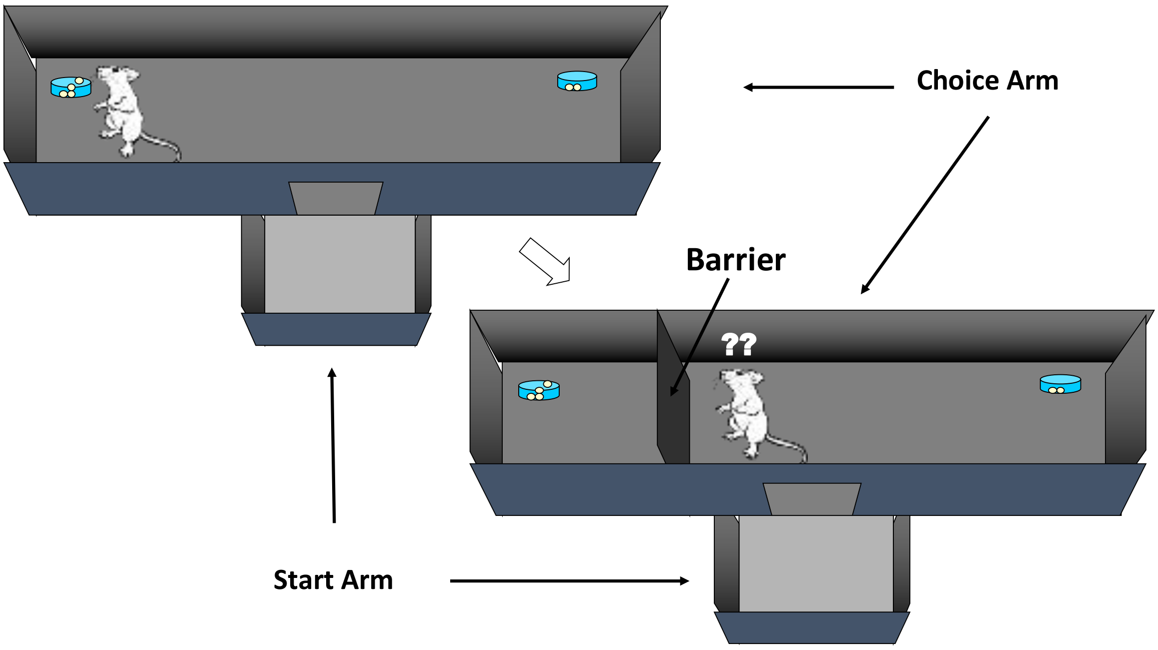


Figure 1. The T-maze barrier test of effort-based choice. Animals are placed in the start arm and then a door is lifted to start the choice trials. In some conditions, the rat simply makes a choice between four pellets in one arm vs. two in the other (top left). In order to assess effort-based choice, rats are tested under conditions in which the arm with the higher density of food reinforcement is obstructed by a vertical barrier, and the animal has the option of climbing the barrier to obtain the higher density of reinforcement vs. choosing the other arm in which there is no barrier, but the density of reinforcement is lower (bottom right). See references [6, 9] for details.

Ethical Statement

These animal experiments were conducted according to US NIH guidelines, and were approved by the Institutional Animal Care and Use Committee of the University of Connecticut.

References

1. Nunes, E. J., Randall, P. A., Hart, E. E., Freeland, C., Yohn, S. E., Baqi, Y., Müller, C. E., López-Cruz, L., Correa, M., & Salamone, J. D. (2013). Effort-related motivational effects of the VMAT-2 inhibitor tetrabenazine: implications for animal models of the motivational symptoms of depression. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *33*(49), 19120–19130.

2. Randall, P. A., Pardo, M., Nunes, E. J., López Cruz, L., Vemuri, V. K., Makriyannis, A., Baqi, Y., Müller, C. E., Correa, M., & Salamone, J. D. (2012). Dopaminergic modulation of effort-related choice behavior as assessed by a progressive ratio chow feeding choice task: pharmacological studies and the role of individual differences. *PloS one*, *7*(10), e47934.

3. Randall, P. A., Lee, C. A., Nunes, E. J., Yohn, S. E., Nowak, V., Khan, B., Shah, P., Pandit, S., Vemuri, V. K., Makriyannis, A., Baqi, Y., Müller, C. E., Correa, M., & Salamone, J. D. (2014). The VMAT-2 inhibitor tetrabenazine affects effort-related decision making in a progressive ratio/chow feeding choice task: reversal with antidepressant drugs. *PloS one*, *9*(6), e99320.

4. Ren, N., Carratala-Ros, C., Ecevitoglu, A., Rotolo, R. A., Edelstein, G. A., Presby, R. E., Stevenson, I. H., Chrobak, J. J., & Salamone, J. D. (2022). Effects of the dopamine depleting agent tetrabenazine on detailed temporal parameters of effort-related choice responding. *Journal of the experimental analysis of behavior*, *117*(3), 331–345.

5. Salamone, J. D., Steinpreis, R. E., McCullough, L. D., Smith, P., Grebel, D., & Mahan, K. (1991). Haloperidol and nucleus accumbens dopamine depletion suppress lever pressing for food but increase free food consumption in a novel food choice procedure. *Psychopharmacology*, *104*(4), 515–521.

6. Salamone, J. D., Cousins, M. S., & Bucher, S. (1994). Anhedonia or anergia? Effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a T-maze cost/benefit procedure. *Behavioural brain research*, *65*(2), 221–229.

7. Salamone, J. D., Ecevitoglu, A., Carratala-Ros, C., Presby, R. E., Edelstein, G. A., Fleeher, R., Rotolo, R. A., Meka, N., Srinath, S., Masthay, J. C., & Correa, M. (2022). Complexities and paradoxes in understanding the role of dopamine in incentive motivation and instrumental action: Exertion of effort vs. anhedonia. *Brain research bulletin*, *182*, 57–66.

8. Sink, K. S., Vemuri, V. K., Olszewska, T., Makriyannis, A., & Salamone, J. D. (2008). Cannabinoid CB1 antagonists and dopamine antagonists produce different effects on a task involving response allocation and effort-related choice in food-seeking behavior. *Psychopharmacology*, *196*(4), 565–574.

9. Yohn, S. E., Thompson, C., Randall, P. A., Lee, C. A., Müller, C. E., Baqi, Y., Correa, M., & Salamone, J. D. (2015). The VMAT-2 inhibitor tetrabenazine alters effort-related decision making as measured by the T-maze barrier choice task: reversal with the adenosine A2A antagonist MSX-3 and the catecholamine uptake blocker bupropion. *Psychopharmacology*, *232*(7), 1313–1323.

10. Salamone, J. D., & Correa, M. (2024). The Neurobiology of Activational Aspects of Motivation: Exertion of Effort, Effort-Based Decision Making, and the Role of Dopamine. *Annual review of psychology*, 75, 1–32.

The 3 choice-T-maze task with running wheel: a mice paradigm to evaluate preference for reinforcers that require vigor and the role of dopamine in anergia

M. Correa [1], P. Matas-Navarro[1], R. Olivares-García[1], A. Martinez-Verdu[1], E. Arias-Sandoval[1], C. Carratalá-Ros[1*], J. D. Salamone[2].

[1] Area de Psicobiologia, Universitat Jaume I, Castelló, Spain

[2] Behavioral Neuroscience Div, University of Connecticut, Storrs, CT, USA.

[*] Present address: Area de Psicobiologia, Universidad de Castilla-La Mancha, Albacete, Spain

Motivated behavior is characterized by a high degree of activity, effort, vigor, and persistence. The ability to produce vigorous and rapid responses and maintain them over time, is a fundamental and highly adaptive feature of motivational processes because these responses enable organisms to exert the effort necessary to reach the work-related limitations that separate them from important stimuli (Salamone and Correa 2023). Organisms should be invigorated to approach selected positive stimuli, but they also have to work to get away from negative stimuli (e.g. painful conditions, predators or stressors).

Studies of behavioral activation are important for understanding some aspects of psychopathology. Thus, symptoms such as anergia, apathy, psychomotor retardation, and fatigue refer to a lack of behavioral activation that can be seen in multiple psychiatric disorders such as depression and schizophrenia and also, in some neurological diseases such as Parkinson disease (Treadway et al, 2012). The severity of these behavioral activation impairments is highly correlated with problems in social function, employment and treatment response. The development of animal models of behavioral activation dysfunctions could enhance the understanding of the neurochemical basis of motivational symptoms in pathologies, as well as it would allow the development of efficient psychopharmacology strategies to treat these motivational symptoms (Salamone and Correa 2023).

In the context of motivation, fatigue, anergia and effort-based decision making is studied using tasks that offer choices. Thus, animals have to choose between high effort options leading to highly valued reinforcers vs. low effort options that procure a less valued reward. Among effort-based decision-making tasks, operant chambers are broadly used for the evaluation of willingness to work for more valued reinforcers. In these studies, animals have to physically work lever pressing to get access to a highly palatable food or, alternatively during the session, approach and consume the less-preferred laboratory chow (standard food) that is concurrently available. Interference with Nucleus Accumbens (Nacb) dopamine (DA) modifies the selection of high-cost high-reward alternatives and biases individuals towards less-effort but lower-reward ones (Salamone and Correa 2023).

Active reinforcers as a tool to evaluate anergia in animal models

Effort-related tasks are based upon the choice of the animal to exert an active response by pressing a lever or climbing a barrier in order to obtain a palatable reinforcer. In addition, some animal studies have developed other methods in which behavioral activation can be studied by using the possibility of engaging in vigorous activities such as wheel running, which is a highly preferred reinforcer in rodents. Mice are a very active species that shows a high level of preference for this activity. Voluntary wheel running occurs spontaneously in mice of all strains, sexes, and ages. Vigorous physical activities, such as wheel running, can have intrinsic reinforcing properties in mice. Therefore, running on a running wheel appears to be motivationally regulated like other appetite behaviors, used for the establishment of a conditioned place preference, and also as a reinforcer in operant-conditioning procedures (Belke and Pierce, 2014).

Of course, engaging in voluntary physical activity is always undertaken in relation to the possible selection of other alternatives; if a running wheel (RW) is present in a complex environment that offers other alternatives such as drugs of abuse, animals will spend a significantly amount of time engaged in the RW (Cosgrove et al, 2002), and rodents often choose running over food consumption when they have a choice (Mueller et al, 1997).

Our group has developed a T-maze task to assess spontaneous preferences between reinforcers that require vigorous activity versus sedentary ones (Carratalá-Ros et al, 2020, 2021a, 2021b; Correa et al, 2016, 2020; López-Cruz et al, 2018; Matas-Navarro et al., 2023; Presby et al., 2021). Thus, the 3-choice-T-maze task for mice establishes voluntary running in a wheel as the high effort/highly preferred option competing with palatable food and exploration of a fruit odor. This rodent task offers the opportunity to evaluate changes in preferences between these concurrently presented reinforcers, as measured by time interacting with each stimuli (Correa et al., 2016). Thus, this task could be used to evaluate the brain mechanisms involved in the decision-making processes that are involved in the spontaneous preference for sustaining voluntarily vigorous behaviors over time.

Using the 3-choice-T-maze task, we can establish a hierarchy of preferences. In this case on the top, and significantly different from the other two, is interacting with a RW that allows a high degree of voluntary physical exercise. In a 15 minutes session, adult male mice spend most of their time running (between 600-800 seconds), but they also spend some time eating (or drinking a sucrose solution in another version; Correa et al., 2020) as the second most preferred option (between 20-40 seconds), and finally they spend very little time sniffing a neutral fruit odor (1-5 seconds). Under these conditions, both sexes behave equally, and both show a strong preference for the RW compared to more sedentary reinforcers. However, older animals of both sexes spent less time running and ate more than young ones, thus showing a more sedentary profile (Matas-Navarro et al., 2023). The second most preferred option for any sex and age is the non-habitual and highly palatable food, which has been used in other rodent paradigms as the most preferred reinforcer (in choice-operant procedures for example; Presby et al., 2021). Interestingly, although food is still the second option, females, independently of age, spend more time eating and consume more pellets than males (Matas-Navarro et al., 2023). The third stimulus, the non-social odor used for these types of studies is a fruit odor (strawberry) that has demonstrated to not generate avoidance as is the one that generates more exploration among several floral or fruit odors (López-Cruz et al., 2018). This type of odor, always generates minimal attraction since it has never been associated with food or any other more powerful reinforcer. This third option only increased in preference when the fruit odor was substituted by a social odor from the same or different sex conspecific. Thus, increasing the salience, and potential value of the olfactory stimuli, increases the time dedicated to explore it, specially, when it was an odor from a different sex, however, even then, the odor was the third option (Carratalà-Ros et al., 2020).

This T-maze task, in contrast to other effort-related tasks, does not involve working with the objective to get a reinforcer of which the animal is partially deprived (i.e food), but instead allows the possibility of freely engaging with reinforcers that are not present usually in their environment. Mice do not need to be food deprived in order to consume the palatable food reinforcer that contains 50% carbohydrates, although they need to be previously exposed to this new test in order to reduce neophobia (Correa et al., 2016). In addition, operant tasks require a long period of training and neural changes due to age can be mediating long term changes in effort. The T-maze task requires only a short training period to have a stable performance (less than 10 days), and then it is able to account for age. Moreover, this test can assess anergia in a non-stressful setting, while traditional rodent models for the study of antidepressant drugs, such as the forced swim test (FST), assesses behavioral activation induced by stressful conditions.

The 3-choice-T-maze task for mice: parameters and dependent variables

The T-Maze apparatus consists of a central area that leads to three arms (25 cm L X 11 cm W X 30 cm H). In each of them, there is a different stimulus; a cotton ball soaked with the fruit odor, high carbohydrate pellets (TestDiet, 50% sucrose, 45mg each), or a RW (based on López-Cruz et al. 2018). Tests are conducted during the light cycle (the least active period for nocturnal animals like rodents) in 15-min sessions, once a day, 5 days per week. During the first week of training, and in order to avoid neophobia to the sweet-tasting pellets, mice are enclosed in the food arm, and a measured quantity of pellets is the only stimulus available during the entire session. Number of pellets consumed each of the 5 day habituation period is recorded.

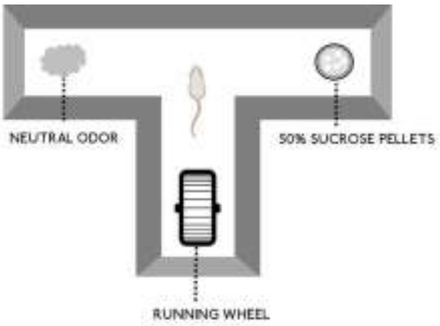


Figure 1. Schematic representation of the 3-choice T-maze task settings for mice.

During the following two weeks, mice are exposed to the T-maze with free access to the three stimuli, each one in a different arm. The location of the reinforcers is counterbalanced between animals in order to avoid an effect of arm orientation preference. Typically, data of baseline preferences correspond to the last day of the second week (the 10th day of free access to the 3 stimuli). Sessions are videotaped, and a trained observer unaware of the experimental condition manually registers all the parameters. Time interacting with the stimuli is the main dependent measure, because it allows for the evaluation of interactions with the three different stimuli with the same units (i.e., seconds). It is well known that time allocation is a useful measure of preference, relative reinforcement value, and response choice. Additionally, in some experiments we also record time spent in each compartment as an index of avoidance to the stimuli or the conditions in that compartment (Carratalá-Ros et al. 2020). Total arm entries are used as a measure of ambulation and general exploration, and number of pellets consumed during the session are also recorded. Presently, by using optical detection methods, additional measures such as average speed and the graphical depictions of heat maps and tracks can be presented, and turns counting in the RW have been added to the description of behavior. After the evaluation of spontaneous preferences, mice are trained for 4 more days, and on the fifth, behavioral or pharmacological manipulations are introduced (Correa et al. 2016, 2020; López-Cruz et al. 2018; Carratalá-Ros et al. 2020, 2021a, b; Matas-Navarro et al., 2023). All procedures were covered by a protocol approved by the Institutional Animal Care and Use Committee of the Universitat Jaume I, in compliance of directive 2010/63/EU of the European Parliament and of the Council.

Conditions that induce anergia in the T-maze task

In our studies, in addition of studying spontaneous sex, age and strain differences, we employed a model of anergia induced by DA receptor antagonists (Correa et al., 2016, 2020) or most commonly, by DA depletion, using a vesicular monoamine transport type 2 (VMAT-2) blocker (tetrabenazine, TBZ) to alter choice behavior. The administration of doses of TBZ, that deplete DA in NAcb (López-Cruz et al., 2018), reduced the time engaged in the RW, although no dose suppressed the preference for the RW as the first option (Carratalá-Ros et al., 2020, 2021a, b; López-Cruz et al., 2018; Matas-Navarro et al., 2023). Furthermore, TBZ produced a reorganization of behavior, compensating for the loss of interaction with the vigorous reinforcer, by increasing time engaged in consuming the highly palatable food. This shift suggests that the animal is not generally anhedonic, since it increases food consumption, and is not fundamentally impaired motorically, since it still spends a significant amount of time running, although mice take more pauses during running (Correa et al., 2020). Interestingly, TBZ reduced time running in middle age and older mice, but not in adolescents, emphasizing the importance of taking into account differences in age when evaluating willingness to exert effort for specific reinforcers.

Thus, anergia is commonly seen as a reorganization of time allocation compared to baseline with no change in absolute preferences: reduced time in the RW, but increased time in food consumption and no change in the third least preferred option. Anergia in older females receiving TBZ showed a different pattern; there was a decrease in RW activity, but no change in time eating, and a small increase in exploration of the most passive reinforcer (time sniffing the neutral odor), probably because the level of palatable food intake was already very high in the older females, thus making difficult a further increase in this reinforcer and the "compensation" was shown as an increase in the third stimulus (Matas-Navarro et al., 2023).

Pharmacological, genetic and behavioral strategies to reduce DA interference-induced anergia in the 3-choice-T-maze task

Caffeine is a minor stimulant well known for its energetic effects. This drug acts as non-selective adenosine A1 and A2a receptor antagonist, acting on Nacb neurons that colocalize DA receptors, and producing the opposite effect to DA receptor antagonist drugs (Correa et al., 2016). Thus, in the T-maze although caffeine did not increase RW selection, it reversed the anergia pattern produced by DA depletion (López-Cruz et al., 2018). Moreover, genetic deletion of the adenosine A2A receptor in mice produces a "resilient" phenotype to DA antagonist-induced anergia in the T-maze (Correa et al., 2016).

Moreover, antidepressants with different mechanisms of action have also been tested. Thus, bupropion, a catecholamine transport blocker that elevates extracellular DA levels and is used as an antidepressant, was administered on its own, and also in combination with TBZ. Bupropion alone increased RW time, and produced a small reduction in time interacting with the food that was not significant. Moreover, it showed a "therapeutic" reversal of TBZ-induced anergia (Carratalá-Ros er al., 2021b). In addition, since clinical practice has focused on pharmacological strategies that increase serotonergic transmission via SERT inhibitors such as fluoxetine, the impact of fluoxetine alone or its capability to alleviate TBZ-induced anergia was assessed. In this case, fluoxetine alone behave as TBZ; reducing time in the RW but increasing time eating. Moreover fluoxetine potentiated TBZ-induced anergia (Carratalá-Ros er al., 2021a), demonstrating that only manipulations that affect the mesolimbic dopaminergic system produce changes coherent with the baseline pattern in the T-maze.

More recently, we are evaluating the protective effect of previous running wheel training (forced or voluntary) for extended periods of time and the impact of environmental enrichment with RW, on DA depletion induced anergia. In addition, the pattern of preferences in the 3-choice-T-maze task of mice that show individual differences in RW performance during long training sessions.

Behavioral manipulations that change spontaneous preferences: patterns different to DA-induced anergia

Because the anergic pattern previously described in the T-maze is dependent on the neural system that regulates decision-making based on the perceived invigoration required by the reinforcer and the perceived state of activation of the organism, we have compared the effects induced by DA depletion against behavioral manipulations that change motivation by changing the homeostatic or emotional value of the reinforcers that are available in the T-maze task. Thus, we could discard potential confounding factors such as physical force, aversion, loss or increase in appetite, etc., in the interpretation of the DA deleting results.

Thus, we started by changing factors that affect the RW value, such as increasing RW resistance to turn and, consequently, increasing the force required to run. During this manipulation, animals reduced time spent running, and were taking more pauses, as is typical after DA depletion, staying in proximity of the RW, thus not showing avoidance of the RW. However, differently to the effect of DA depletion, increasing RW resistance did not increase food consumption as a compensation. In another experiment, making the compartment of the RW aversive by placing an anxiogenic light over the RW, reduced time running and produced a shift towards time consuming food. However, under these conditions animals showed avoidance of the RW chamber and increased time in the other two compartments, something that was not produced by DA depletion (Carratalá-Ros er al., 2020). Moreover, "RW-satiation" allowing mice to run in a home cage RW during 2 hours just before test, reduced time interacting with the RW in the T-maze, and increased the time that animals did not interacted with any stimulus with no change in sucrose consumption (Correa et al., 2020).

Increases in palatable food value were performed by restricting standard food and inducing a binge-like eating profile (Carratalá-Ros er al., 2020). Food restricting mice in the home cage the night before the test, increased drastically time eating, and time spent in the food compartment in detriment of time running and in the RW compartment. DA depletion produces the same profile in terms of time interacting with the stimulus but it does not affect time in the compartment. Moreover, the number of pellets consumed for restricted animals increased five fold while in DA depleted animals this increase is not even double. A very similar type of effects in all these variables was found in animals that have been trained during weeks previous to test to have a binging pattern of intake. These animals reduced time running, increasing time eating but also time signifying the non-social odor, thus showing a stronger sedentary profile, since they compensated for both types of alternative stimuli. Inducing satiation by allowing the animals to eat as much as they wanted of the palatable pellets or making the pellets aversive by adding a bitter flavor, did not produce the same pattern than the pharmacological manipulations that increase DA, such as bupropion. These manipulations produced a significant increase in time running and a dramatic reduction in time eating, and pellet intake (9 times for the bitter food and 4 times for prefeeding), an effect that was not seen in bupropion treated animals (Carratalá-Ros er al., 2021a,b).

Comparison of the 3-choice-T-maze task with mice models based on behavioral activation by aversive reinforcers: the forced swim test (FST)

This new animal paradigm for the study of anergia after DA depletion with TBZ, has been compared with a classical animal model of anergia induced, in this case, by learned helplessness in an aversive context, the FST. TBZ produced a very similar pattern of results in both paradigms; the FST and the 3-choice-T-maze task. However, there are important differences between both paradigms. For instance, the FST uses an acute and inescapable stressor (a tank with water) to generate baseline anergia; the animal rapidly learns that it can not escape and minimizes effort staying immobile or mildly swimming to keep afloat. This is used as a model of depressive behavior. DA depletion potentiates that baseline inactivity. However, in the case of the 3-choice-T-maze task among the 3 different reinforcers, all of them appetitive, the spontaneous behavior of any mice is to select the one that allows voluntary vigorous exercise, the RW. Thus, in the T-maze, DA depletion is required in order to generate anergia. This anergia leads to a partial reorganization of preferences. While the FST tries to model a pathological state, the T-maze is intended for the study of factors that modulate normal behavior.

In addition, we compared the impact of two of the most common antidepressant drugs used in clinical practice: fluoxetine and bupropion, for their effects in both models of anergia. Both of these drugs are monoamine uptake inhibitors, but they seem to differ substantially in their improvement of motivational dysfunctions such as fatigue. Thus, bupropion alone increased time engaged in running in the T-maze task and also increased time climbing in the FST, and it was able to reverse the effects of DA depletion in both animal models. The SERT inhibitor fluoxetine was effective at increasing swimming and climbing. However, in the 3-choice-T-maze task fluoxetine produced anergia-like pattern of behaviors, very similar in fact to what was characteristic of TBZ: reduction in running and increasing time consuming sucrose pellets. Furthermore, fluoxetine failed to reverse the behavioral impairment produced by TBZ in the FST, and it was not able to reverse TBZ-induced suppression of time running in the RW.

The idea of different animal tests for the evaluation of specific symptoms, but not for a pathology is consistent with the research domain criterion (RDoC) approach that highlights the importance of describing the neural circuits that mediate specific symptoms in psychopathology, and not simply the traditional diagnostic categories.

References

1. Carratalá-Ros C, López-Cruz L, SanMiguel N, Ibáñez-Marín P, Martínez-Verdú A, Salamone JD, Correa M (2020) Preference for Exercise vs. More Sedentary Reinforcers: Validation of an Animal Model of

Tetrabenazine-Induced Anergia. Front Behav Neurosci 13:289. https://doi.org/10.3389/fnbeh.2019.00289

2. Carratalá-Ros C, López-Cruz L, Martínez-Verdú A, Olivares-García R, Salamone JD, Correa M (2021a) Impact of Fluoxetine on Behavioral Invigoration of Appetitive and Aversively Motivated Responses: Interaction With Dopamine Depletion. Front Behav Neurosci 15:700182. https://doi.org/10.3389/fnbeh.2021.700182

3. Carratalá-Ros C, Olivares-García R, Martínez-Verdú A, Arias-Sandoval E, Salamone JD, Correa M (2021b) Energizing effects of bupropion on effortful behaviors in mice under positive and negative test conditions: modulation of DARPP-32 phosphorylation patterns. Psychopharmacology (Berl) 238:3357-3373. https://doi.org/10.1007/s00213-021-05950-4

4. Correa M, Pardo M, Bayarri P, López-Cruz L, San Miguel N, Valverde O, Ledent C, Salamone JD (2016) Choosing voluntary exercise over sucrose consumption depends upon dopamine transmission: effects of haloperidol in wild type and adenosine A2AKO mice. Psychopharmacology (Berl) 233:393-404. https://doi.org/10.1007/s00213-015-4127-3

5. Correa M, Pardo M, Carratalá-Ros C, Martínez-Verdú A, Salamone JD (2020) Preference for vigorous exercise versus sedentary sucrose drinking: an animal model of anergia induced by dopamine receptor antagonism. Behav Pharmacol 3:553-564. https://doi.org/10.1097/FBP.0000000000000556

6. López-Cruz L, San Miguel N, Carratalá-Ros C, Monferrer L, Salamone JD, Correa M (2018) Dopamine depletion shifts behavior from activity based reinforcers to more sedentary ones and adenosine receptor antagonism reverses that shift: Relation to ventral striatum DARPP32 phosphorylation patterns. Neuropharmacology. 138:349-359. https://doi.org/10.1016/j.neuropharm.2018.01.034

7. Matas-Navarro, P. Carratalá-Ros C, Olivares-García R, Martínez-Verdú A, Salamone JD, Correa M (2023) Sex and age differences in mice models of effort-based decision-making and anergia in depression: the role of dopamine, and cerebral-dopamine-neurotrophic-factor. Psychopharmacology (Berl). 240(11): 2285-2302. doi: 10.1007/s00213-023-06430-7.

8. Presby RE, Rotolo RA, Hurley EM, Ferrigno SM, Murphy CE, McMullen HP, Desai PA, Zorda EM, Kuperwasser FB, Carratala-Ros C, Correa M, Salamone JD (2021) Sex differences in lever pressing and running wheel tasks of effort-based choice behavior in rats: Suppression of high effort activity by the serotonin transport inhibitor fluoxetine. Pharmacol Biochem Behav 202:173115. https://doi.org/10.1016/j.pbb.2021.173115

9. Salamone JD, Correa M. (2023). The Neurobiology of Activational Aspects of Motivation: Exertion of Effort, Effort-Based Decision Making, and the Role of Dopamine. Annu Rev Psychol.doi: 10.1146/annurev-psych-020223-012208.

10. Treadway MT, Bossaller NA, Shelton RC, Zald DH. (2012). Effort-based decision-making in major depressive disorder: a translational model of motivational anhedonia. J Abnorm Psychol. 121(3):553-8. doi: 10.1037/a0028813.

Symposium: Bioacoustics

Acoustic features of vocalisations of laying hens in positive and negative emotional states

Kriengwatana, B.P., Golfidis, A., Norton, T.

KU Leuven, Department of Biosystems, Division of Animal and Human Health Engineering, B-3000 Leuven, Belgium

Abstract

Vocalisations are a promising method for welfare monitoring but assessing the affective state of animals based solely on the type of vocalization produced has limitations. This study compared the acoustic features of specific vocalization types of laying hens in positive and negative emotional states. Vocalisations were recorded when hens received stimuli that varied in arousal and valence. It is expected that vocalisations in positive contexts will be shorter and lower pitched than negative contexts.

Introduction

Vocalisations are a promising method for welfare monitoring because they can provide information about the affective states of animals. Most animals produce vocalisations that are reserved for specific contexts (e.g. predator alarm calls), but some vocalisations can be produced in both positive and negative context, making it problematic to monitor and assess the state of animals based solely on the type of vocalization produced. For example, in laying hens, the food call is often emitted in anticipation of food or dustbathing substrate (positive) [1,2], while the gakel and whine calls are frequently emitted when hens were prevented from accessing food or nest boxes (negative; [3]). However, food calls are also produced during frustration contexts, and gakels and whines can be found during food anticipation contexts [2,3].

In mammals, it has been found that vocalization types produced in both positive and negative context have different acoustic characteristics. For instance, horse whinnies produced during social reunion (positive) were lower pitched and shorter in duration than whinnies produced during social isolation (negative) [4]. Grunts, screams, and squeals of wild boars during food anticipation and affiliative interactions (positive) were also shorter in duration, lower frequency, and had less amplitude modulation than during agonistic interactions (negative) [5]. Such comparisons between vocalization types produced in positive and negative states is scarce for poultry [6], and studies that do exist in birds do not compare acoustic features of vocalisations in positive and negative contexts (e.g. [2,7]). Nonetheless, this knowledge is needed to improve welfare monitoring of the billions of farmed poultry worldwide [8].

The study aimed to compare the acoustic features of specific vocalization types produced by laying hens in positive and negative emotional states. The vocalization data analysed were collected from hens after they had interacted with an automated device that delivered positive or negative stimuli. Hens could freely and voluntarily activate the device, which was located in their home pen. Stimuli were chosen to affect arousal and valence based on previous literature linking these stimuli with approach or avoid behaviours [9–13], as it is assumed that animals generally move towards stimuli that are positive and away from stimuli that are negative.

Ethical statement

Ethical approval was obtained prior to experiments from the Ethical Committee for Animal Experimentation at KU Leuven (project number 082/2023).

Materials & Methods

Animals and housing

Eight ISA brown laying hens were obtained from TRANSfarm, KU Leuven. Hens were tested in pairs and ranged from 19 – 42 weeks of age at the start of the experiment. Hens were housed in pens 2.3 x 2.3 x 0.8 m (L x W x H) with a net covering the top of the pen, in a climate-controlled room in the Department of Biosystems, KU Leuven animal facility. The ambient temperature was kept at 21°C and humidity between 60-70%. The daily light/dark cycle was 14h/10h light dark, with lights on between 08:00 – 22:00. Hens received ad libitum food and water in conventional poultry feeders, access to a dust-bathing substrate, pecking stone, elevated perching area for roosting, and nest boxes (2 per pair). Daily checks were conducted to ensure the birds had sufficient food and water and were in good health.

Stimuli

The experiment contained five different stimulus types: *Neutral*, *Positive valence + High arousal*, *Positive valence + Low arousal*, *Negative valence + High arousal*, and *Negative valence + Low arousal*. Positive stimuli were mealworms (high arousal) and rice (low arousal), respectively. Negative stimuli were two air puffs (low arousal), four air puffs (high arousal), rice coated with 1% quinine (a bittering agent; low arousal) and rice coated with 4% quinine (high arousal). Food colouring was used to make the rice with 1% quinine green and the rice with 4% quinine blue so that hens could differentiate between the negative low and high conditions by associating the colours with their respective conditions.

Recording Setup

Two automated devices were placed in the pen. Each device was a circular-shaped container covered by a rotating disc and elevated above the ground. Six separate containers were inside the device, hidden by the rotating disc which had a single hole to allow hens to access one of the six containers at a time. Five containers were filled with regular feed, rice, rice + 1% quinine, or rice + 4% quinine. The final container remained empty with an outlet for the air puff. An ultrasonic sensor was located on the side of the device and would register hens that were within 10 cm of the device. If no hens were detected (baseline condition), the disc would remain in the start position, above the empty container. If a hen was detected, the disc could rotate to give access to one of the food stimuli or deliver puffs of air. The order of each stimulus given was randomised.

A directional microphone (AKG C 391 B) was set up above each device, and another omnidirectional one (AKG SE300 B) was used to record ambient sounds within the enclosure. All microphones were connected to a soundcard (Focusrite Clarett+4Pre) which was then connected to a laptop that stored audio recordings. Recordings were made continuously (24h/day). To capture video data, a camera was installed above the pen (Dahua DH-SD1A203T-GN). This camera was connected to a Network Video Recorder (Dahua DHI-NVR4208-8P-4KS2) that stored the video recordings. Video recordings were made during lights on only (from 08:00 – 22:00).

Discussion of methodology

Results indicate that hens were motivated to engage with the device frequently despite receiving both positive and negative stimuli, and they only showed a gradual decline in interest after more than a week in one of the trials [14]. Hens quickly learned how to activate the device, although there was a noticeable difference between pairs. For example, pair 1 averaged 244 activations daily whereas pair 2 averaged 72 activations daily. In both pairs, however, their usage persisted over a week, demonstrating peaks and troughs despite the introduction of mild punishments such as air puffs and distasteful food. Across all trials, we obtained over 8000 times total device activations.

Expected results

When they interact with the device, hens were likely to vocalise approximately 50% of the time [14], leading to an estimate of at least 4000 trials in which a vocalisation is present. Thus, we will have a large dataset from which we can perform acoustic analyses. We hypothesise that, similar to mammals, vocalisations of laying hens in positive contexts will be shorter and lower pitched than negative contexts. This expectation is based on two previous studies, where zebra finches in a negative context (social isolation) produced vocalisations with higher pitch and longer duration than neutral contexts (social housing) [7] and laying hens produced food calls that were lower pitched when anticipating a reward [2].

References

1. Evans CS, Evans L. Representational signalling in birds. Biol Lett. 2007;3:8–11.

2. McGrath N, Dunlop R, Dwyer C, Burman O, Phillips CJC. Hens vary their vocal repertoire and structure when anticipating different types of reward. Animal Behaviour. 2017 Aug 1;130:79–96.

3. Zimmerman PH, Lundberg A, Keeling LJ, Koene P. The Effect of an Audience on the Gakel-Call and Other Frustration Behaviours in the Laying Hen (*Gallus Gallus Domesticus*). Anim welf. 2003 Aug;12(3):315–26.

4. Briefer EF, Maigrot AL, Mandel R, Freymond SB, Bachmann I, Hillmann E. Segregation of information about emotional arousal and valence in horse whinnies. Scientific Reports. 2015 May 22;5(1):9989.

5. Maigrot AL, Hillmann E, Briefer E. Encoding of Emotional Valence in Wild Boar (Sus scrofa) Calls. Animals. 2018 Jun 5;8(6):85.

6. Laurijs KA, Briefer EF, Reimert I, Webb LE. Vocalisations in farm animals: A step towards positive welfare assessment. Applied Animal Behaviour Science. 2021 Mar 1;236:105264.

7. Perez EC, Elie JE, Soulage CO, Soula HA, Mathevon N, Vignal C. The acoustic expression of stress in a songbird: Does corticosterone drive isolation-induced modifications of zebra finch calls? Hormones and Behavior. 2012 Apr 1;61(4):573–81.

8. Eurostat. Agricultural production - livestock and meat [Internet]. European Union; 2023 [cited 2024 Jan 11]. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Agricultural_production_-_livestock_and_meat#Livestock_population

9. Davies AC, Radford AN, Nicol CJ. Behavioural and physiological expression of arousal during decision-making in laying hens. Physiology & Behavior. 2014 Jan 17;123:93–9.

10. Davies AC, Radford AN, Pettersson IC, Yang FP, Nicol CJ. Elevated arousal at time of decision-making is not the arbiter of risk avoidance in chickens. Sci Rep. 2015 Feb 3;5(1):8200.

11. Paul ES, Edgar JL, Caplen G, Nicol CJ. Examining affective structure in chickens: valence, intensity, persistence and generalization measured using a Conditioned Place Preference Test. Applied Animal Behaviour Science. 2018 Oct 1;207:39–48.

12. Skelhorn J, Rowe C. Birds learn to use distastefulness as a signal of toxicity. Proceedings of the Royal Society B: Biological Sciences. 2010 Feb 3;277(1688):1729–34.

13. Ganchrow JR, Steiner JE, Bartana A. Behavioral Reactions to Gustatory Stimuli in Young Chicks (Gallus gallus domesticus). Developmental Psychobiology. 1990;23(2):103–17.

14. Golfidis, A, Kriengwatana BP, Mounir M, Norton T. An interactive feeder to induce emotions for assessing positive welfare in chickens. (Unpublished).

Primate Detection Through Passive Acoustic Monitoring Varies According to Species and the Biome

J.C. Lacerda[1,a], G.B. Lima[1,b], R. Taynor[1,2,c], M. Araújo[2,d], J.P.S. Alves[1,e], B. Bezerra[1,f]

[1]Programa de Pós-Graduação em Biologia Animal, Departamento de Zoologia, Universidade Federal de Pernambuco, Recife, Brasil. [a]juliana.lacerda@ufpe.br, [b]geovana.lima@ufpe.br, [c]rick.taynor@ufpe.br, [e]joao.alves@ufpe.br, [f]bruna.bezerra@ufpe.br.

[2]Instituto SOS Caatinga, São José da Tapera, Brasil. [d]institutososcaatinga@gmail.com

Abstract

Primate loud calls are adapted for transmission over long distances, which could facilitate passive acoustic monitoring (PAM). We tested PAM to assess the use of the acoustic space and presence of *Callithrix jacchus* and *Sapajus flavius* in sympatry in Caatinga and Atlantic Forests. We found that primate detection through PAM varies according to species and biome. Thus, we must consider life history traits, landscape attributes and call features to optimise PAM methods for primates.

Keywords

PAM; call detection; life history; landscape; call structure; acoustic signals; vocal behaviour.

Introduction

Vocal communication is essential for arboreal primates since the dense vegetation may make it difficult for other forms of communication, such as through visual signals [1]. Primates have loud vocalisations with specific characteristics adapted for transmission over long distances [2]. Therefore, they are potential models for passive acoustic monitoring (PAM) in difficult-to-access areas. Here, we aimed to test PAM to access the use of the acoustic space and the presence of two primate species (*Callithrix jacchus* and *Sapajus flavius*) living in sympatry in two biomes: the Caatinga scrub forest and the Atlantic Forest in Northeast Brazil. The Caatinga is an exclusively Brazilian biome that occupies the largest semi-arid region in the country [3]. The Atlantic Forest is Brazil's third-largest biome, and it is one of the most threatened tropical forests on the planet [4]. *Callithrix jacchus* belongs to the Callitrichidae family, lives in small territorial family groups and is categorised as Least Concern in the IUCN red list of threatened species [5; 6]. *Sapajus flavius* belongs to the Cebidae family, lives in large groups under a fission-fusion system, and is categorised as Endangered [7; 8]. Both species are known to occur in the Caatinga and Atlantic forests, but most of the studies with both species have been conducted in the Atlantic Forest [5; 8].

Methods

We conducted Passive Acoustic Monitoring (PAM) in two fragments of Caatinga forest in Alagoas State and two fragments of Atlantic Forest, one in Pernambuco State and the other in Paraíba State (see Figure 1) over six months, with 48h of monthly recordings. We conducted recordings between February and July 2022 (Atlantic Forest fragment in Paraíba State), July and December 2022 (Caatinga forest fragments in Alagoas State) and September 2022 and February 2023 (Atlantic Forest fragment in Pernambuco State). We used Audiomoth recorders (Open Acoustic Devices) under 48kHz and 16-bit sampling rate. We conducted a clustering analysis to detect and sort similar acoustic signals from all the recordings using Kaleidoscope Pro software (Wildlife Acoustics).

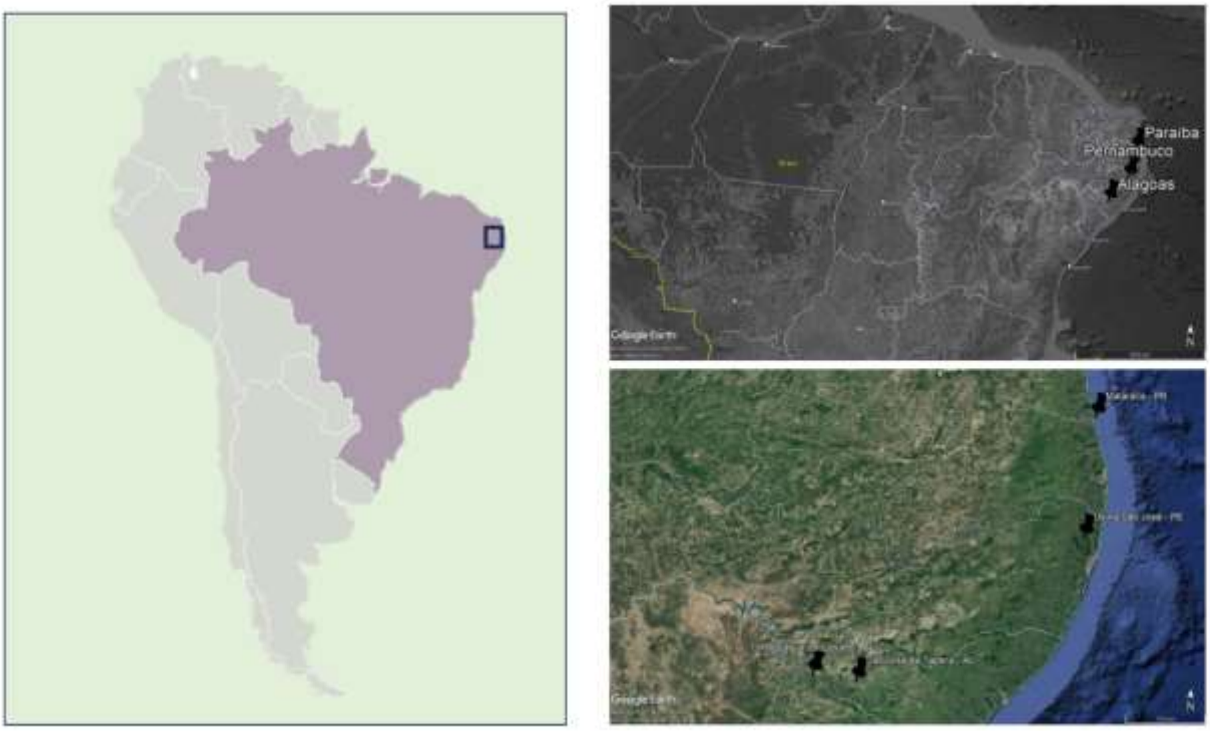


Figure 1. Map of the study sites in Paraíba, Pernambuco and Alagoas States in Northeast Brazil. Atlantic forest fragments: Mataraca-PB and Usina São José-PE; Caatinga Forest fragments: Olho D´água do Casado-AL and São José da Tapera-AL.

Results

We obtained recordings of *Callithrix jacchus* at both biomes, with seven different call types registered through PAM (see Figure 2). Call types recorded varied according to the biome. In the Atlantic Forest, we recorded the following call types for *Callithrix jacchus*: Long Phee Call, Brief Phee Call 1, Brief Phee Call 2 and Brief Phee Call 3. In the Caatinga forest, we recorded the following call types: Long Phee Call, Brief Phee Call 1, Brief Phee Call 2 and Brief Phee Call 3, Tisik Call, Scream Call and Twitter Call. Nevertheless, most of the vocalisation records for *Callithrix jacchus* were from the Atlantic Forest, representing 74% of the signals detected. Records of *Callithrix jacchus* vocalisations varied throughout the day, with records occurring between 4:00 and 17:00.

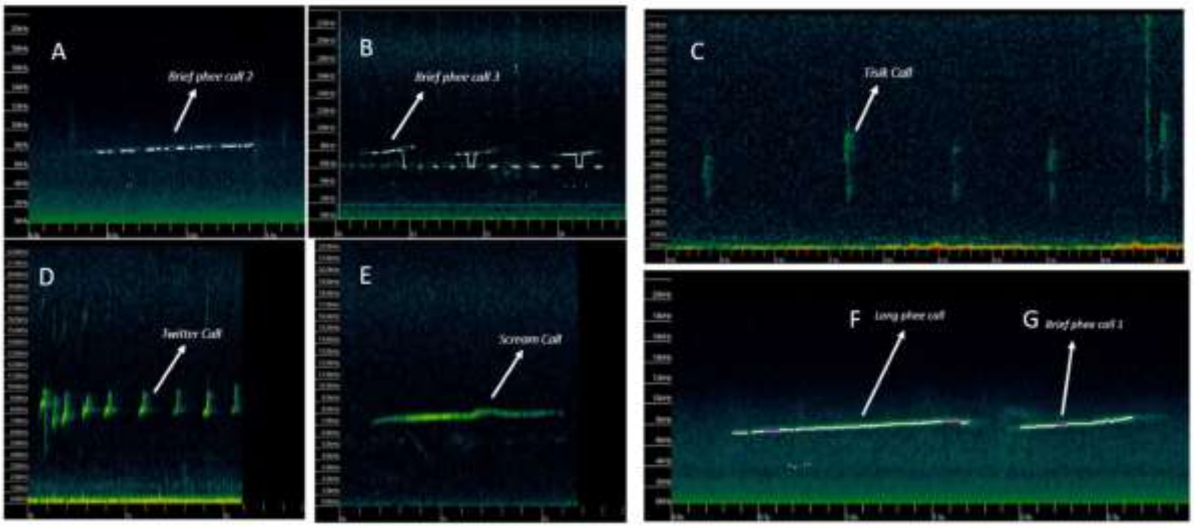


Figure 2. Call types we recorded for *Callithrix jacchus* through PAM in the Atlantic and Caatinga forests. A) Brief phee call 2; B) Brief phee call 3; C) Twitter call; D) Scream call; E) Tisik call; F) Long phee call; G) Brief phee call 1.

Even though *Sapajus flavius* occurs in all four study fragments, we acoustically detected the species only in the Atlantic Forest fragments, with 17 different call types registered (see Figure 3). We recorded the following call types for *Sapajus flavius*: Bellow, Cheep, Clack, Fear Call, Heh, Hoot, Howl, Huh, Huh-1var, Huh-2var, Huh-3var, Huh-4var, Lost Call, Shout, Trill, Whoop and Yell. Records of *Sapajus flavius* vocalisations varied throughout the day, occurring between 5:00 and 17:00.

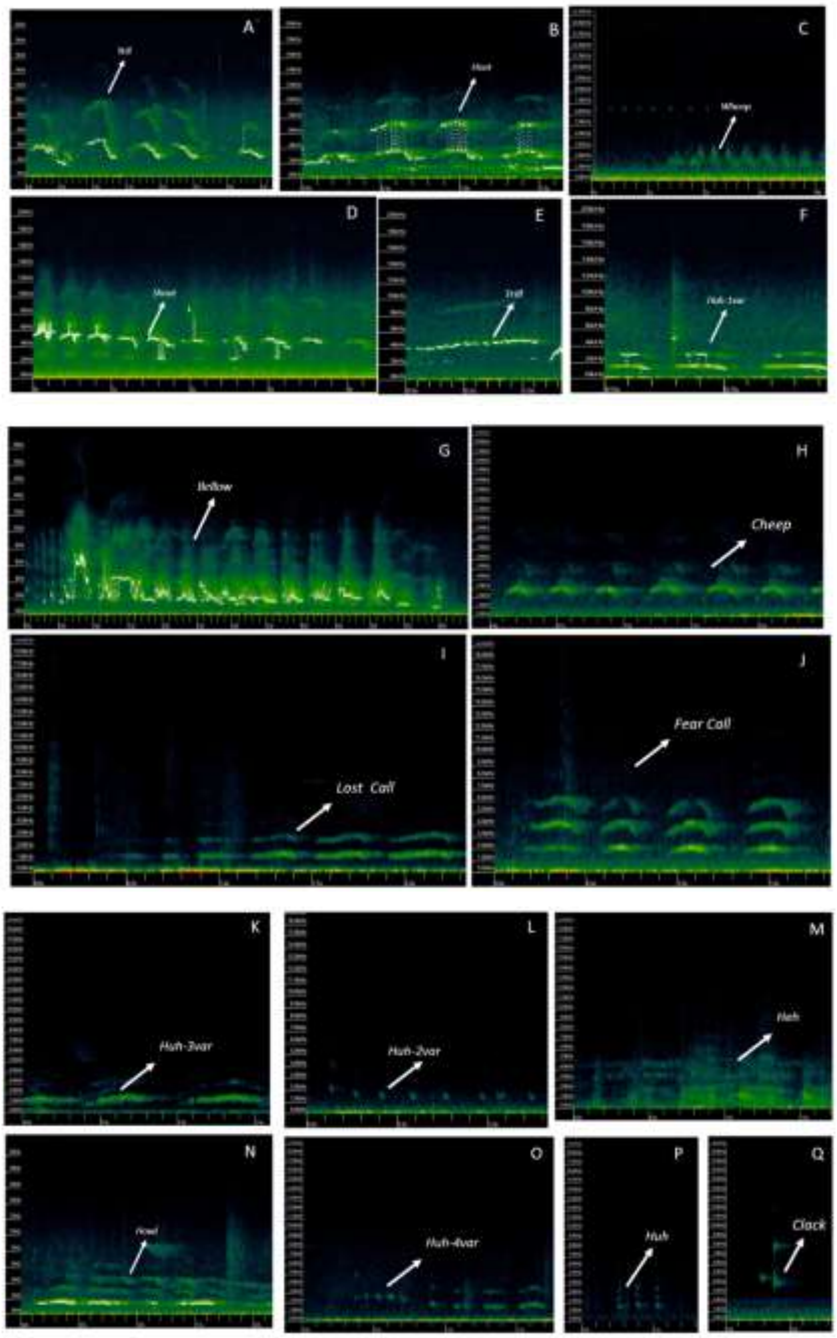


Figure 3**.** Call types we recorded for *Sapajus flavius* through PAM in the Atlantic and Caatinga forests. A) Yell; B) Hoot; C) Whoop; D) Shout; E) Trill; F) Huh-1var; G) Bellow; H) Cheep; I) Lost Call; J) Fear Call; K) Huh-3var; L) Huh-2var; M) Heh; N) Howl; O) Huh-4var; P) Huh; Q) Clack.

Discussion

We show here that primate detection through PAM varies according to the species, the fragment, and the biome in which they occur. *Callithrix jacchus* lives in relatively small family groups of up to 15 individuals, has small home ranges and is very territorial [9; 10]. Thus, for PAM aiming to detect this species, once the recorder is placed within their territory, there is a good chance of detection. On the other hand, *Sapajus flavius* can live in groups with more than 150 individuals under complex fission-fusion social dynamics, and like other capuchin monkeys, they travel long distances daily in their large home ranges [11; 12; 8]. Thus, the chances of acoustic detection can be reduced. Also, the acoustic properties of the long-distance calls and their propagation may influence acoustic detection through PAM. The frequencies of long-distance *phee* calls from *Callithrix jacchus* are around 8 kHz, whereas the frequencies of long-distance *Trill* calls from *Sapajus flavius* are about 4 kHz. We trust that one needs to consider life history traits (home range sizes and social patterns), the acoustic properties of the vocalisations and landscape attributes (at least biome type and vegetation density) to optimise PAM design for primate species and their habitats.

Acknowledgements

We thank CAPES, CNPq, and Facepe for their financial support. The study complies with Brazilian law, and data were collected under the SISBIO licence 25727.

References

1. Bezerra, B., Barnett, A., Souto, A., & Jones, G. 2013. Vocal communication in *Cacajao*, *Chiropotes* and *Pithecia*: Current knowledge and future directions. In L. Veiga, A. Barnett, S. Ferrari, & M. Norconk (Eds.), Evolutionary Biology and Conservation of Titis, Sakis and Uacaris (Cambridge Studies in Biological and Evolutionary Anthropology). Cambridge: Cambridge University Press. p. 303-308.

2. Schneider, C., Hodges, K., Fischer, J., Hammerschmidt, K. 2008. Acoustic niches of Siberut primates. *International Journal of Primatology*, 29: 601–613.

3. Tabarelli, M., Filgueiras, B K.C., Ribeiro, E. M.S., Lopes, A. V., Leal, I. R. 2024 Tropical Dry Forests. In: Scheiner Samuel M. (eds.) Encyclopedia of Biodiversity 3rd edition, vol. 1, pp. 294–312. Oxford: Elsevier.

4. Safar, N. V. H.; Magnago, L. F. S.; Schaefer, C. E. G. R. 2020. Resilience of lowland Atlantic forests in a highly fragmented landscape: insights on the temporal scale of landscape restoration. *Forest Ecology and Management, Amsterdam*, 470, 1181-83.

5. Schiel, N., Souto, A. 2016. The common marmoset: An overview of its natural history, ecology and behavior. *Developmental Neurobiology*, 77(3), 244–262.

6. Valença-Montenegro, M.M., Bezerra, B.M., Ruiz-Miranda, C.R., Pereira, D.G., Miranda, J.M.D., Bicca-Marques, J.C., Oliveira, L., Da Cruz, M.A.O.M., Valle, R.R., Mittermeier, R.A. 2021. *Callithrix jacchus* (amended version of 2018 assessment). The IUCN Red List of Threatened Species 2021: e.T41518A191705043.

7. Valença-Montenegro, M.M., Bezerra, B.M., Martins, A.B., Jerusalinsky, L., Fialho, M.S., Lynch Alfaro, J.W. 2021. *Sapajus flavius* (amended version of 2020 assessment). The IUCN Red List of Threatened Species 2021: e.T136253A192592928.

8. Medeiros, K., Bastos, M., Jones, G., Bezerra, B. 2019. Behavior, diet, and habitat use by blonde capuchin monkeys (*Sapajus flavius*) in a coastal area prone to flooding: Direct observations and camera trapping. *International Journal of Primatology* 40, 511–531.

9. Stevenson, M. F.1988. The marmosets, genus *Callithrix*. Ecology and Behavior of Neotropical Primates, v. 2, p. 131-222.

10. Kinzey, W. G. 1997. New World primates: ecology, evolution, and behavior (pp. 192-199). New York: Aldine de Gruyter.

11. Presotto, A., Izar. P. 2010. Spatial reference of black capuchin monkeys in Brazilian Atlantic Forest: egocentric or allocentric? *Animal Behaviour*, 80(1). 125-132.

12. Izar. P. 2004. Female social relationships of *Cebus apella nigritus* in a southeastern Atlantic Forest: an analysis through ecological models of primate social evolution. *Behaviour*, 141(1). 71-99.

Bioacoustic sensors to monitor farm animal welfare: why the ethology matters

M. Coutant

1 Department of Animal and Veterinary Sciences, Aarhus University, Tjele, Denmark.
mathilde.coutant@anivet.au.dk

Introduction

Bioacoustics refers to the study of animal sounds, which includes recording, analyzing and interpreting vocalizations as well as non-vocal signals to gain information about the animal producing the sound. The field of bioacoustics is growing, particularly in relation to the detection and monitoring of wild animal species, thanks to the development of complex machine learning algorithms capable of detecting patterns across large datasets [1]. Yet, the methods developed for that sake may be applied to other domains, such as animal husbandry. While species detection requires algorithms to classify vocalizations according to the emitting species of e.g., bird or whale, monitoring of animals in husbandry may include the classification of vocalizations and other sounds as belonging to a healthy vs. sick individual, eating vs. resting individual, stressed vs. relaxed individual, and so on. Hence, various indicators of health, physiology, behaviour, and affect could potentially be detected and monitored in the sounds of livestock animals [2]. While such monitoring is highly promising for farmers, as it would allow them to surveil their animals in a remote, continuous, and cost-effective manner [3], it may also become an important tool for researchers to assess animal welfare and evaluate the effects of different husbandry management practices. Together with researchers from Copenhagen University (Denmark), we therefore recently reviewed the literature to determine the potential of sounds-based sensors to monitor livestock animal welfare, and form conclusions on the future of this technology in this field [4].

The potential of bioacoustic sensors to monitor farm animal welfare

The review showed a range of promising research, including potential detection of various respiratory diseases, heat stress and other environmental factors, as well as assessment of physical characteristics such as weight, sex or age. Using both vocalizations and sounds (such as chewing and biting), studies have also showed potential to monitor feeding behaviour, sexual behaviour, as well as various maladaptive behaviours such as feather pecking in poultry or tail biting in finisher pigs. A large body of research has additionally focused on assessing affective states including stress, pain or anticipation, as well as valence (i.e. whether the animal is in a positive or negative state of mind) in multiple species. This collection of studies illustrates a strong potential for bioacoustic sensors to become a valuable tool for the management of animals. Yet, despite this array of promising research, a main conclusion of the review was that most of the studies published in the last decade are still at an experimental stage. While many managed to achieve a high accuracy of detection based on vocalizations obtained experimentally from animals held in small, controlled environments, validation of these methods in practice remains complex. In fact, only few studies presented promising results for algorithms tested in commercial herds, and only one product (SoundTalks®) is, to our knowledge, currently available commercially.

The main explanation for the difficult transition from experimental studies to detection in practice is likely the conditions of most animal husbandry systems, characterized by a large amount of animals kept together and an often high level of background noise. In comparison to wildlife monitoring, the review also highlighted a lack of replicability of studies performed on farm animals, and called for attention on the lack of large open databases for husbandry animal vocalizations. While the field of bioacoustics applied to animal husbandry is likely to continue growing, focusing on the monitoring of even more aspects of animal welfare, we therefore hypothesized that a major challenge in the future will be the validation of algorithms in practice to obtain reliable sensors usable by both farmers and researchers. Advances in machine learning and computing technology will likely play a major role in this matter, along with the development of large databases of farm animal sounds.

The importance of understanding farm animal vocal communication

While the technical and computational limits of bioacoustic research are often well understood and documented, we believe that other aspects of bioacoustics remain understudied, with serious implications for the field. This presentation aims to stress the importance of addressing the gaps in our current knowledge of animal vocal communication, especially in the specific context of husbandry, to ensure a valid use of sound-based sensors in farming.

One crucial aspect in the development of tools to monitor animal welfare is the interpretation of the information extracted from the sounds of animals. It is indeed critical to ensure that this ‘de-coding’ is performed correctly, which includes evaluating the honesty of the vocalizations. In other words, we need to make sure that we can trust the information obtained from the vocalizations, and that our interpretation takes into consideration the context in which they are produced. While it is farfetched to say that animals can lie, we know from several examples that they may, voluntarily or not, modulate their vocal communication based on their environment. The red deer may, for instance, lower the frequency of his vocalizations to appear larger than he is to other males from a distance, therefore asserting dominance [5]. If bioacoustic analyses were to be used to determine the weight or body size of such individual in that precise context, without understanding the vocal behaviour of that species, wrongful conclusions would likely be drawn.

Information on such aspects of communication in farm animals remain scarce. Even basic elements of animal communication remain to be established. A recent study has for instance suggested that piglets may produce vocalizations of frequencies reaching 40 kHz in life threatening situations, yet most studies record vocalizations up to 20 kHz (the maximum frequency audible to humans) [6]. Such mismatch between the intrinsic characteristics of vocalizations and the way they are recorded may impact how animals are monitored, especially, in that specific case, in relation to detection of critical events such as crushing.

Overall, despite having millions of individuals under our care, very little is known about the vocal communication of most species of farm animals. This is especially true considering the fast evolution of animal husbandry and the unique conditions in which animals are kept in most conventional systems. Modern husbandry practices often involve keeping individuals in groups with hundreds of conspecifics, housed in large and potentially loud barns, and relying on humans for feed and care – a radical change to the life of their wild ancestors. How the vocal behaviour of animals have evolved and adapted to these conditions remain for the most part unknown, but research has started to show effects of this environment on vocal communication.

For instance, it is known that noise in the environment may have various effects on the vocal communication of mammals. Modifications may include increasing the amplitude (i.e. volume) of vocalizations proportionally to the increase in noise level (also called Lombard effect), shifting the frequency of vocalizations to a band with less noise (i.e. changing pitch to avoid covering from the noise) or avoiding vocalizations until the noise is reduced [7]. Such modifications have for example been observed in sows, producing vocalizations of a higher volume when housed in a loud barn (with noise present from ventilation fans) as compared to a relatively quieter environment [8]. The level of noise in the environment can also affect vocal communication via modulation of the hearing capacities of animals. A recent study indeed showed that the ventilation and the automatic feeding system, present in most intensive husbandry systems, produce sounds up to 40 kHz [6]. While these frequencies are not audible to humans, they likely appear particularly loud to pigs that are capable of hearing them. In fact, a study performed in commercial farms in the US showed that most sows in the gestation unit had hearing damage, and some were even deaf [9]. The researchers linked that phenomenon to a high prevalence of crushing (where the sow likely doesn´t hear the screams of piglets as she sits on them). Even when placing the sows in a quieter environment, the proportion of piglet crushing did not reduce [9], potentially indicating a long-term modulation of sow-piglet communication. Besides high levels of noise in the barn, evidence show that some species, such as laying hens, may adapt their vocalizations to the presence of conspecifics or humans (so-called audience effects). In a study, researchers showed that food-deprived hens for instance adapted their vocal signaling of frustration to the state of the other hens in the ‘audience’ [10], which implies that group dynamic and composition may modulate how emotional states are conveyed in the vocalizations.

All these phenomena (audience effects, Lombard effects, hearing loss from the environment, etc..) may greatly impact several vocal characteristics, including the number and type of vocalizations produced by the animals, as well as the amplitude or frequency of the vocalizations, which are key elements in the assessment of e.g. affective states [11]. A lack of understanding and/or consideration of these effects can therefore result in wrongful assessment of their welfare. To give a concrete example, piglets may stop producing distress calls following a lack of feedback from the sow (resulting from hearing damage or desensitization), seemingly appearing not to be stressed. Alternatively, they may intend to attract the sows' attention by increasing the amplitude or frequency of their calls, seemingly appearing to be in severe distress. Overall, it seems critical to understand these dynamics to establish a correct interpretation of their vocalizations.

In addition, such environmental-driven vocal modulation may partly explain why detection tools developed in laboratory settings with few animals kept in a quiet environment may not be applicable in practice, where animals are placed in large groups with different social dynamics and, consequently, potentially different vocal behaviours.

Conclusion

Along with the development of robust algorithms usable in practice, it seems therefore critical to understand how specific husbandry conditions are shaping farm animal vocal communication, and how these modulations may affect our interpretation of their vocal signals. This knowledge is crucial to ensure a valid use of bioacoustic sensors to monitor welfare, while providing further understanding of the impact of husbandry conditions on farm animal welfare itself.

References

1. Mcloughlin, M. P., Stewart, R., & McElligott, A. G. (2019). Automated bioacoustics: Methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface*, **16**(155).
2. Manteuffel, G., Puppe, B., & Schön, P. C. (2004). Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science*, **88**(1–2), 163–182.
3. Banhazi, T. M., & Black, J. L. (2009). Precision Livestock Farming: A Suite of Electronic Systems to Ensure the Application of Best Practice Management on Livestock Farms. *Australian Journal of Multi-Disciplinary Engineering*, **7**(1), 1–14.
4. Coutant, M., Villain, A.S., Briefer, E.F. (2023). A review of the use of bioacoustics to assess various components of farm animal welfare. Under review.
5. Fitch, W.T. & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society B* **268**(1477):1669-75.
6. Heseker, P. (2023). Detecting pig screams - using animal vocalization for monitoring tail biting events. *Book of Abstract - Behavior 2023*, 91. Bielefeld, Germany.
7. Hotchkin, C. & Parks, S. (2013). The Lombard effect and other noise-induced vocal modifications: insight from mammalian communication systems. *Biological Reviews*, **88**(4), 809-824.
8. Chapel, N. M., Radcliffe, J. S., Stewart, K. R., Lucas, J. R., & Lay Jr, D. C. (2019). The impact of farrowing room noise on sows' reactivity to piglets. *Translational Animal Science*, **3**(1), 175-184.
9. Chapel, N. M. (2018). The sound science of sows: influence of auditory environment on sow hearing, piglet communication, and sow behavior in modern swine production [Doctoral thesis, Purdue University].
10. Zimmerman, P. H., Lundberg, A., Keeling, L. J., & Koene, P. (2003). The effect of an audience on the gakel-call and other frustration behaviours in the laying hen (Gallus gallus domesticus). *Animal Welfare*, **12**(3), 315-326
11. Briefer, E. F. (2020). Coding for 'Dynamic'information: Vocal expression of emotional arousal and valence in non-human animals. Coding strategies in vertebrate acoustic communication. Springer. 137-162.

Symposium: AI and Machine Learning

Real-Time Adaptive Machine Learning Systems for Personalised Bruxism Management

T.C. Dolmans[1]

[1]SØVN, Utrecht, The Netherlands. tenzing@sherpai.nl

Abstract

This paper introduces SØVN, a wearable device using in-ear sensors to detect and disrupt bruxism (teeth grinding and jaw clenching). It's a less invasive alternative to current detection methods like EMG. The device, leveraging machine learning, monitors jaw movements and provides real-time interventions like auditory stimuli. This innovative approach offers scalable, personalized treatment for bruxism, with potential applications in other health conditions requiring real-time monitoring and intervention.

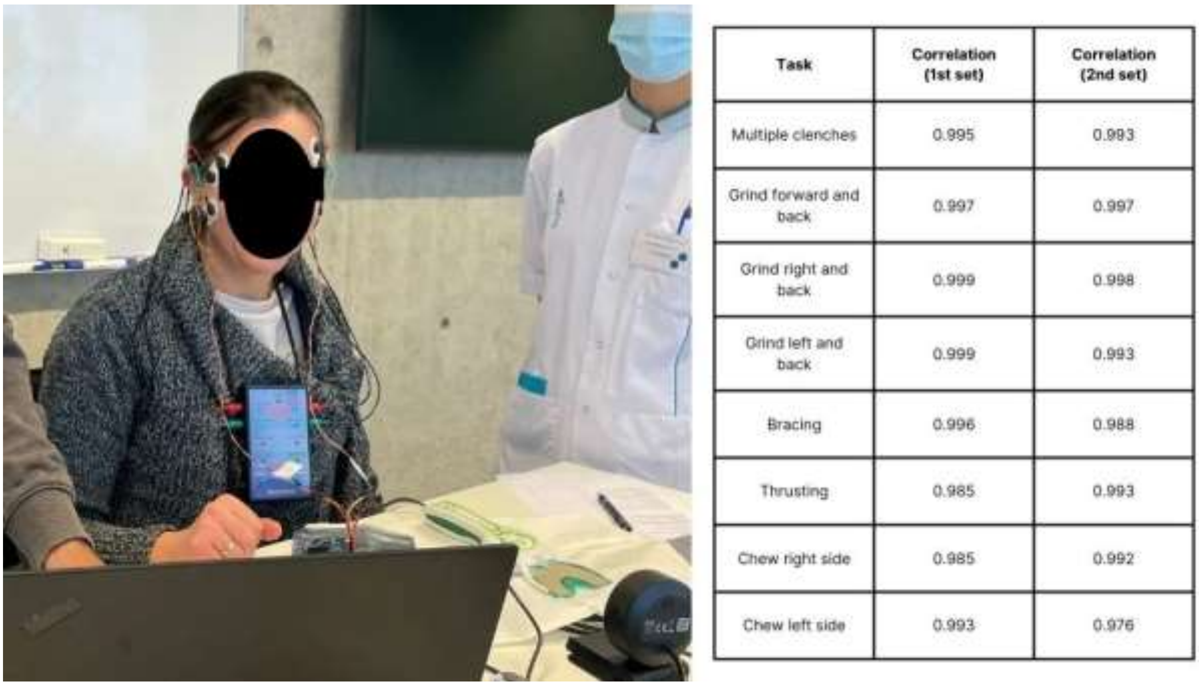
Keywords: Sleep Bruxism, Machine Learning, Deep Learning, Modelling Behaviour

Introduction

The combination of wearable sensors, an increased need for personalised healthcare, and machine learning allows for the tackling of many problems. The purpose of this short paper is twofold: to outline a framework through which these problems can be addressed and to solve a specific such problem. Bruxism, which is a subconscious behaviour characterised by teeth grinding, and jaw clenching or thrusting, affects nearly 130 million individuals in the EU and US alone [1]. SØVN is a novel device that learns to recognise and disrupt jaw movement that is associated with bruxism. To do this, in-ear sensors, also known as “earables”, collect and analyse data from users in real-time. The design principles, components, and considerations are discussed throughout this work. Though bruxism is harmless in most people, when the behaviour occurs on a frequent basis, it can cause severe problems for the teeth, jaw, and facial muscles [2]. These issues range from tooth damage, headaches, and muscle pain, among other categories [3][4][5]. Estimates put the prevalence of sleep bruxism (SB) and awake bruxism (AB) at 8% and 15% respectively [6]. Of these groups, around 40% are chronic bruxers with more than six episodes every hour every night [7]. A smaller fraction of around 15% experience severe symptoms like myofascial pain and dental complications [8].

Currently, there exist no effective solutions or standardized approach for bruxism. Mouthguards are the most popular dental strategy, they protect the teeth by allowing the user to grind the guard away, rather than the teeth, but they do not address the underlying behaviour [9]. Furthermore, custom mouthguards are costly, and the premise of the product relies on ingesting relatively large quantities of synthetic materials. The success of psychological approaches such as cognitive therapy, behavioural therapy, and stress management does not significantly different from the success of mouthguards. [9]. Prescription drugs and Botox injection approaches are also employed, though less frequently. These methods also fight symptoms rather than the underlying cause, while introducing serious adverse effects The ideal solution would enable bruxers to manage the behaviour in the first place.

The current golden standard for the detection of bruxism is electromyography (EMG). EMG is collected and reviewed by a human expert for events of clenching, grinding, and bracing. The phenotype of the EMG depends on the nature of the events. Generally, the setup of EMG can be quite cumbersome and time consuming. The SØVN device is a small in-ear wearable device capable of sensing jaw movements via the deformation of the ear canal. In order to establish whether this minimally invasive earables can truly replace EMG, a pilot study was done in which participants were asked to perform a series of movements so as to simulate various episodic behaviours. In pilot trials with the RadboudUMC in Nijmegen involving 17 participants, the Netherlands, the device is shown to have an above 99% correlation with EMG in measuring simulated jaw movements [10]. The ambulatory nature of the device allows for users to be in their own homes when participating in research and receiving intervention.



Task	Correlation (1st set)	Correlation (2nd set)
Multiple clenches	0.995	0.993
Grind forward and back	0.997	0.997
Grind right and back	0.999	0.998
Grind left and back	0.999	0.993
Bracing	0.996	0.988
Thrusting	0.985	0.993
Chew right side	0.985	0.992
Chew left side	0.993	0.976

Figure 4: A participant in the research. The right side of the image contains values for correlation between EMG and SØVN data in various tasks [10].

Goals

The first goal is to accurately detect when jaw motor activities indicative of teeth grinding, or jaw clenching occur automatically and in real-time. To mitigate the previously discussed negative effects of bruxism, it is paramount that the underlying behaviour is understood and detectable. The second goal is to automatically provide interventions which reduce event frequency, intensity, and duration. This, in turn, would lead to mitigation of painful temporomandibular disorder, chronic migraine, and severe tooth wear. To achieve these goals, we outline our methodology below in hopes of aiding others to solve similar problems.

Methods

The system contains several levels of processing that interplay to self-improve through reinforcement learning from human and machine feedback. A high-level overview is described to build an intuition about the parts, after which the calibration, detection, and intervention are discussed in more detail. Many of the system's hyperparameters are subject to optimisation, these parameters are tuned with the hyperparameter optimisation toolbox Optuna [11]. These parameters are indicated with the following notation HP_x, where HP indicates hyperparameters and subscript x indicates specific sub-sections, such as model parameters (m) or intervention parameters (i). A summary of parameters is given in table 2.

Layer	Parameters Affecting	Notation	Changed By
L1	Models, features	HP_m, HP_f	L4
L2	Windows	HP_w	L4, human expert
L3	Interventions	HP_i	L4, human expert, user
L4	Objective function	HP_l	Human expert

Table 1: Overview of parameters of each different layer of the system. What parameters affect, as well as their notation and editors are outline. HP = Hyperparameter, the subscript denotes what the parameters affect.

System Structure

At the first and fastest level (L1), a real-time classification is done with incoming data. This level receives data about participant behaviour from the device and provides a classification every second. This is done with learnt optimisation, using various machine-and deep-learning models. These models are discussed in more detail later. At level 2 (**L2**), participants' heart-and breathing rate variability (HRV & BRV), as well as their current sleep stage (CST) are continuously monitored. These metrics require more data to be reliably calculated, therefore L2 provides information on a slower basis. Where L1 operates on the order of seconds, L2 operates on the order of minutes. The next layer (**L3**) combines 'faster' and 'slower' information from L1 and L2 and determines whether an intervention should be triggered. The fourth and final layer (**L4**) assesses the effectiveness of system by considering the effects of L3 and optimises for an objective function as defined by the human expert (equation 1). Figure 2 provides a simple visualisation of the interplay and hierarchy of the layers. The behaviour of each layer is governed by their respective parameters.

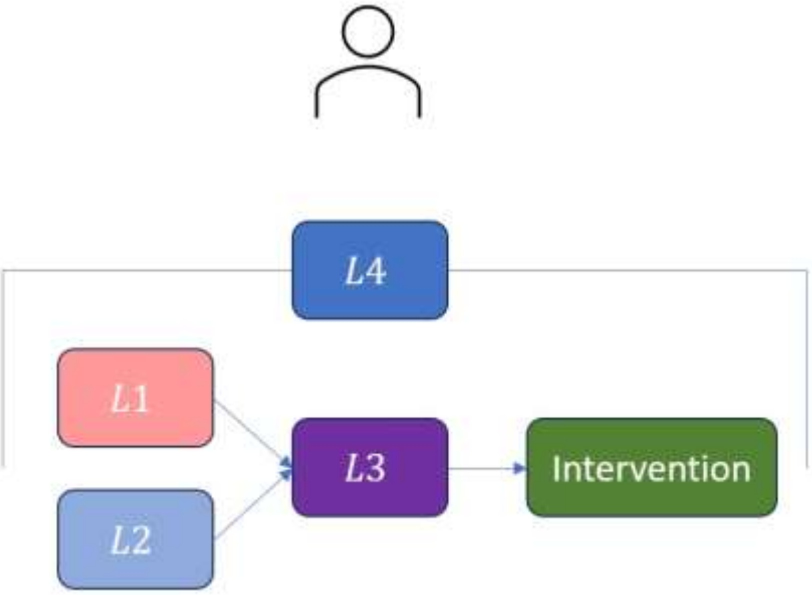


Figure 5: Layered diagram describing the system. L1 and L2 determine L3 behaviour. The effects of interventions are assessed by L4. A human expert supervises the system.

Calibration

When participants first use the device, calibration takes place to improve detection of events (taking place in L1) and the properties of interventions (relevant for L3). For detection, values for the Maximum Voluntary Contraction/Movement (MVC/MVM) are found. These values are used as boundaries for what observations can reasonably be expected during usage. As a starting point for the intensity of the intervention, the perceptual threshold is found. As the system gets used, the calibration is automatically refined.

Detection

Six participants were invited to be part of a data collection cohort for product development. In total, 16 nights of data were recorded in the participants' own beds while using the device simultaneously with a reference device; Bruxoff, collecting EMG and electrocardiograms (ECG) [14]. During these nights, a total of 289 events were identified using the reference device and confirmed by a human expert reviewing the raw EMG and ECG data, a time-consuming process. The recorded data are split into one-second windows which contain either event or non-event data and associated with a known true label. In total, there are around 1800 event windows with a duration of one second and an overlap of 0.15% between windows. The exact number of windows depends on the parameters of window creation, HP_w. Several machine-and deep-learning-based models are optimised for their accuracy, specificity, and sensitivity in the classification of the label for each window. Accuracy refers to the fraction of correctly identified windows. Specificity is defined as the fraction of true negatives divided by true negatives plus false positives. Sensitivity is defined as the fraction of true positives divided by true positives plus false negatives. In this scenario, a system should catch more events (true positives) at the expense of incorrectly classifying non-events (false positives), meaning that there exists a trade-off between specificity and sensitivity. The ‘best’ balance between these metrics depends on the goal of the system.

L1 classification is done by two classes of models: gradient-based, and non-gradient-based models. Gradient-based models are those that are comprised of parameters that are learnable through gradient descent. We use a simple multilayer perceptron (MLP), a mixture of experts (MoE), which stacks several MLPs with a gating layer, and an MoE with a self-attention-based gating layer (MoEWithSA). The self-attention is based on the famous transformer architecture as described by Vaswani et al. in 2017 [12]. This class of models is trained on pre-processed data where pre-processing consists of bandpass filtering, moving-average filtering, and normalisation. We use a support vector machine (SVM) and random forest (RF) as non-gradient-based models. This class of models is trained with extracted features as described in Bondareva et al., 2021 [13]. Classification in L1 is done with only one model at a time. Model parameters (HP_m) and which combinations of features to use (HP_f) are optimised by Optuna as part of L1 optimisation [11]. L2 insights are generated on a sliding-window basis where the windows are much longer than the windows used in L1; the exact length is subject to optimisation in HP_w.

Intervention

L3 chooses whether to perform an intervention and what type of intervention to perform using L1 and L2. Good interventions are those that result in L1 no longer indicating “event” while not leading to significant sleep disturbances as classified by L2 (expressed in HRV, BRV, and CST). In other words, good interventions are effective but not disruptive. Unsuccessful interventions are penalised, as described in the objective function (equation 1). Currently, several interventions that are being tested are: pink or brown noise, low frequency noise, verbal stimuli, or music. The type, frequency, duration, and intensity of interventions are subject to optimisation in HP_i. Lastly, users can choose to personalise their interventions.

If an intervention is triggered, any of three possible scenarios follow. First, if L1 and L2 remain unchanged there is a “null” result. The intensity and frequency of interventions go up: $HP_i \uparrow$. Second, if both L1 changes and L2 indicates disturbances, the intervention was successful but disruptive. The intensity and frequency of interventions go down: $HP_i \downarrow$. Third, if L1 changes but L2 does not, the intervention was successful and the L4 is reinforced with low loss value. A human expert determines the relative importance of each part of the objective function: parameters α and β. Increasing alpha results in more relative importance of stopping events, increasing beta puts more emphasis on not disturbing the user’s sleep.

Objective Function

The objective function for L4 is defined by combining the behaviour of L1-3. Assuming high sensitivity and specificity from L1, the value of L1 is either 0 (indicating non-event) or 1 (indicating event). The value of L1 is multiplied by alpha. L2 sleep disruption is multiplied by beta. This means that completely undisrupted sleep results in a zero-effect of the second term, leaving only the status of event or non-event as the objective. Thus, the objective function penalises disrupting participants’ sleep and rewards effective intervention without disruption, see equation 1.

$$objective = \alpha L1 + \beta \Delta L2 \downarrow$$

Equation 1: Objective function of L4. L1 indicates current status of layer 1 (values can be either “event”: 1 or “non-event”: 0). $\Delta L2$ refers to the measured sleep disturbance (values can range from “not disturbed”: 0 to “very disturbed”: 5 α and β are scalers determiend by a human expert. The down arrow indicates that the value is subject to minimisation.

Results

Preliminary results are available for the performance of L1, which are summarised in table 2. As of yet, no tests have been done that allow us to assess higher layers. Classification of a single window takes 5.4ms for gradient based models and 2.5ms for non-gradient based models. An eight-hour night’s recording can thus be classified in 72-156 seconds. The output of the system is an annotation file that contains a list of events begin and end times.

Model	Accuracy	Sensitivity	Specificity	Notes
MLP	0.73	0.14	0.05	S@S = 0.7
MoE	0.56	N/A	N/A	S@S = 0.7
MoEWithSA	0.51	N/A	N/A	S@S = 0.7
SVM	0.80	0.69	0.72	S@S = 0.7
RF	0.77	0.60	0.73	S@S = 0.7

Table 2: Results of L1. For each model, the accuracy, sensitivity, and specificity are indicated. The best results for each column are bolded. In notes, "S@S" indicates sensitivity at specificity and vice versa. The values in the respective columns and rows thus indicate a sensitivity at a specificity of 0.7 or vice versa.

Discussion

As more data is collected, system performance is expected to go up. Rather than relying solely on expert evaluation from zero, this system provides a way to review a night's data at a glance. Furthermore, since inference takes several milliseconds, real-time classification can be done to provide bruxers with real-time interventions. In the future, individualised systems can be re-trained in near real-time by allowing L4 to adjust in epochs throughout the night. With human evaluation, L1-3 can also be personalised by fine-tuning model performance on user-specific data.

The outlined system effectively wraps the modelling of behaviour in parameters than can automatically be adjusted according to individual needs. Furthermore, real-time classification in combination a comfortable earable allows for interventions that mitigate the previously discussed downsides of alternative solutions. Because of the automated nature of the system, this solution can be deployed at scale. Lastly, the outlined framework can be applied to other behavioural modelling tasks that would otherwise require large time investments of human experts. Such problems might be sleep apnoea, epilepsy, or stroke; the earable device is showing strong indications that it might be valuable in other scenarios where real time detection and intervention of events during sleep can be useful.

Ethical Statement

Participants in this research were presented with adequate information about the nature of data collection and processing, after which they each signed an informed consent form. It was made clear to participants that they could withdraw from this research at any time, without having to explain their decision.

References

1. McKinsey&Company. "Feeling good: The future of the $1.5 trillion wellness market." https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/feeling-good-the-future-of-the-1-5-trillion-wellness-market. Published on April 8, 2021, Accessed on December 29, 2023.
2. Zhou, Y., Gao, J., Luo, L., & Wang, Y. (2016). Does bruxism contribute to dental implant failure? A systematic review and meta-analysis. *Clinical implant dentistry and related research*, **18(2)**, 410-420.
3. Reissmann, D. R., John, M. T., Aigner, A., Schön, G., Sierwald, I., & Schiffman, E. L. (2017). Interaction between awake and sleep bruxism is associated with increased presence of painful temporomandibular disorder. *Journal of oral & facial pain and headache*, **31(4)**, 299-305.
4. Fernandes, G., Franco, A. L., Gonçalves, D. A., Speciali, J. G., Bigal, M. E., & Camparis, C. M. (2013). Temporomandibular disorders, sleep bruxism, and primary headaches are mutually associated. *J Orofac Pain*, **27(1)**, 14-20.
5. Li, Y., Yu, F., Niu, L., Hu, W., Long, Y., Tay, F. R., & Chen, J. (2018). Associations among bruxism, gastroesophageal reflux disease, and tooth wear. *Journal of clinical medicine*, **7(11)**, 417.
6. Wetselaar, P., Vermaire, E., Lobbezoo, F., & Schuller, A. A. (2019). The prevalence of awake bruxism and sleep bruxism in the Dutch adult population. *Journal of oral rehabilitation*, **46(7)**, 617-623.

7. Rompré, P. H., Daigle-Landry, D., Guitard, F., Montplaisir, J. Y., & Lavigne, G. J. (2007). Identification of a sleep bruxism subgroup with a higher risk of pain. *Journal of dental research*, **86(9)**, 837-842.
8. Ohlmann, B., Waldecker, M., Leckel, M., Bömicke, W., Behnisch, R., Rammelsberg, P., & Schmitter, M. (2020). Correlations between sleep bruxism and temporomandibular disorders. *Journal of clinical medicine*, **9(2)**, 611.
9. Yap, A. U., & Chua, A. P. (2016). Sleep bruxism: Current knowledge and contemporary management. *Journal of conservative dentistry*: JCD, **19(5)**, 383–389. https://doi.org/10.4103/0972-0707.190007
10. Kleisman, F., Sparreboom- Kalaykova, S. I. (2023). Detectie van gesimuleerd bruxisme met slimme oordoppen. *Master Thesis, Tandheelkunde, RadboudUMC, Nijmegen.*
11. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *In Proceedings of the **25**th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30.**
13. Bondareva, E., Hauksdóttir, E. R., & Mascolo, C. (2021). Earables for Detection of Bruxism: a Feasibility Study. *In Adjunct Proceedings of the **2021** ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (pp. 146-151).
14. Saczuk, K., Lapinska, B., Wilmont, P., Pawlak, L., & Lukomska-Szymanska, M. (2019). The Bruxoff Device as a Screening Method for Sleep Bruxism in Dental Practice. Journal of clinical medicine, 8(7), 930. https://doi.org/10.3390/jcm8070930

Measuring equine respiration in the field: an exploration of microphone data and deep learning detectors

J.I.M. Parmentier[1,2,3*], R.M. Aarts[2], F.M. Serra Bragança[2,4], B.J. van der Zwaag[1,3]

[1]Pervasive Systems Group, Edge Research Centre, EEMCS, University of Twente, Enschede, The Netherlands,
[2]Clinical Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands,
[3]Inertia Technology B.V., Enschede, The Netherlands
[4]Sleip AI, Stockholm, Sweden
***Corresponding author: j.i.m.parmentier@utwente.nl**

Introduction

An important physiological aspect in exercising horses is respiration and its associated parameters (e.g., breathing pattern, respiratory rate, and locomotor-respiratory coupling). Quantifying variations during training in these parameters has a huge potential on understanding mechanisms linked to equine performance [1], but also to potentially detect respiratory anomalies, and ultimately improve equine welfare. Related to this, several studies have explored using microphones to study respiratory sounds showing its potential to detect upper-airway disorders [2]–[4]. This supports the use and exploration of microphone data to monitor equine respiration during training.

However, segmentation of respiratory cycles was limited to manually counting respiratory sounds and most studies were conducted on treadmills, where environmental noises are not representative of training conditions and using a sulky is impossible. Due to increasing available computational power and new data processing techniques (e.g., deep learning), there are opportunities to develop near-real-time quantifications of the respiratory patterns and respiratory rates of (harness) horses during training.

In this study, we explore deep learning models to automatically detect exhalation events in microphone data obtained on exercising harness trotters. We also evaluate the effect of downsampling the microphone data on the performances of our exhalation events detectors.

Materials and Methods

Fifteen warmblood harness trotters in full training were equipped with an omni-directional microphone (ECM-LV1, Sony, 44100 Hz) taped between the nostrils and a voice recorder (ICD-Px470, Sony) attached to the bridle. All horses underwent a standardized exercise test for harness trotters on an oval track, with increasing trotting speeds. The last segment of high-speed trot (37.9±1.5km/h on average) was used for this study as respiration sounds were more distinguishable than at lower speeds. High-speed trot is also the condition under which the respiratory system of the trotters is under pressure and during which respiratory sounds are the loudest.

As exhalation sounds were generally louder and more distinct than inhalation sounds, especially in presence of environmental noise (tack, hoof beats, wind), exhalation events were chosen for this study. Exhalation events were labelled by one user in Audacity 3.4.2. [5] through sound and visually inspecting both microphone channels and their Mel spectrogram and labelled microphone data were then imported in MATLAB R2022b (MathWorks, Natick, Massachusetts, USA) for training the detectors.

Temporal Convolutional Network (TCN) detectors [6] of different depths (4, 8 and 12 dilation blocks) were trained to classify sound signals into exhalation (1) and no exhalation (0) in a sequence-to-sequence classification fashion. The detectors were trained with either one of the two channels or both channels together, with different downsampling factors (1, 2, 5, 10 and 40, thus sampling frequencies of 44100 Hz, 22050 Hz, 8820 Hz, 4410 Hz and 1102.5 Hz respectively). Each model was trained with the data of thirteen horses, with one horse withheld for validation (early stopping with validation patience of five epochs) and another one for test (leave-one-out performance evaluation). The F1-scores were then computed to compare the performance of the different TCNs. Detectors with higher median F1-score with lower interquartile range (IQR) were considered performing better.

Results

When training with shallower TCN detectors, the best performance was obtained with lower sampling frequencies (see Table 1). However, at higher sampling frequencies, most false positives were loud sound events detected as exhalations. Increasing the depth of the TCNs decreased the sensitivity to the sampling frequencies, as shown in Table 2. This is explained by the learning mechanisms lying behind the TCN architecture: deeper TCN learn from a wider receptive field in the training data. For a receptive field of ten samples, the duration of the observed events is of 0.0002 seconds at 44100Hz against 0.001 seconds at 8820Hz. Overall, better results were obtained when training with Channel 2 or Both Channels compared to Channel 1. This was regardless of the downsampling factor (see Table 2). Exhalation events were more recognisable in Channel 2 compared to Channel 1 during the labelling process, explaining the different results and further showing the importance of correct labelling in deep learning methodologies.

Table 3, Median and interquartile ranges (IQR) F1-scores for 4 blocks TCN detectors trained with either Channel 1, Channel 2 or Both Channels with high sampling frequencies (22050Hz and higher) or low sampling frequencies (8820Hz and lower).

Sampling frequency (TCN 4 Blocks)	Channel 1	Channel 2	Both Channels
High	0.72 IQR [0.66-0.77]	0.77 IQR [0.73-0.80]	0.79 IQR [0.66-0.82]
Low	0.81 IQR [0.73-0.85]	0.86 IQR [0.84-0.90]	0.86 IQR [0.85-0.90]

Table 4, Median and interquartile ranges (IQR) F1-scores for TCN detectors trained with either Channel 1, Channel 2 or Both Channels with different depths, for all sampling frequencies.

TCN depth (All sampling frequencies)	Channel 1	Channel 2	Both Channels
4 blocks	0.77 IQR [0.71-0.82]	0.84 IQR [0.77-0.87]	0.84 IQR [0.80-0.88]
8 blocks	0.84 IQR [0.79-0.89]	0.89 IQR [0.85-0.93]	0.90 IQR [0.86-0.92]
12 blocks	0.86 IQR [0.78 -0.90]	0.90 IQR [0.87-0.93]	0.91 IQR [0.88-0.92]

Conclusion

This work shows that by using a microphone simply taped to the nose of horses it is possible to automatically detect exhalation events at high-speed trot with deep learning models, despite the environmental noise. We also showed that decreasing the sampling frequency improves the performance of the detectors, especially for less complex models which are less prone to overfitting. Using only one channel of data seems to be sufficient for the automatic detection of equine exhalation events in microphone data, but we would like to be cautious with the application of this result as some measured sound events might be buried in environmental noise. This would be depending on where and how the microphone is placed on the horse's nose. Further comparisons of microphone placements need to be conducted to conclude on the most robust and reliable channel(s) to use for this purpose. In the future, we will explore the performance of our detectors at lower trotting speeds, as well as evaluating the detectors' outputs to compute physiological information like respiratory rate and compare our results to a simpler signal processing method published in the human literature [7].

Ethical Statement

All animal experiments in this work were approved by the Swedish Ethics Committee (diary number 5.8.18-04197/2022) in agreement with Swedish legislation on animal experiments.

Acknowledgements

The authors would like to thank Zala Zgank, Ebba Zetterberg and Guilherme de Camargo Ferraz for their help during the data collection, as well as the Menhammar Stuteri drivers and grooms. They would also like to thank Elin Hernlund, Marie Rhodin, Johan Hellander, Mihai Marin-Perianu and Raluca Marin-Perianu for their input in the project. This study was partially funded by the research project "Varenne" E! 114697.

References

1. C. Cotrel, C. Leleu, and A. Courouce-Malblanc, 'Factors influencing variation in locomotor-respiratory coupling in Standardbred Trotters in the field', *Equine Vet. J.*, vol. 38, no. S36, pp. 562–566, 2006, doi: 10.1111/j.2042-3306.2006.tb05605.x.
2. G. R. G. Barnes, M. Brennan, B. E. Goulden, and J. Kirkland, 'Sound spectography in the diagnosis of equine respiratory disorders: a preliminary report', *N. Z. Vet. J.*, vol. 27, no. 7, pp. 145–146, Jul. 1979, doi: 10.1080/00480169.1979.34629.
3. F. J. Derksen, S. J. Holcombe, W. Hartmann, N. E. Robinson, and J. A. Stick, 'Spectrum analysis of respiratory sounds in exercising horses with experimentally induced laryngeal hemiplegia or dorsal displacement of the soft palate', *Am. J. Vet. Res.*, vol. 62, no. 5, pp. 659–664, May 2001, doi: 10.2460/ajvr.2001.62.659.
4. S. H. Franklin, S. G. Usmar, J. G. Lane, J. Shuttleworth, and J. F. Burn, 'Spectral analysis of respiratory noise in horses with upper airway disorders', *Equine Vet. J.*, vol. 35, no. 3, pp. 264–268, May 2003, doi: 10.2746/042516403776148228.
5. 'Audacity ® | Free Audio editor, recorder, music making and more!' Accessed: Dec. 31, 2023. [Online]. Available: https://www.audacityteam.org/
6. S. Bai, J. Z. Kolter, and V. Koltun, 'An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling'. arXiv, Apr. 19, 2018. doi: 10.48550/arXiv.1803.01271.
7. C. Romano *et al.*, 'Respiratory Rate Estimation during Walking and Running Using Breathing Sounds Recorded with a Microphone', *Biosensors*, vol. 13, no. 6, Art. no. 6, Jun. 2023, doi: 10.3390/bios13060637.

Challenging Machine-Learning with Underwater Whisker Tracking in South African Fur Seals

A.Elder[1*], C. Brassey[1], E. McLoughlin[2], K.Todd [1,2] and R. Grant[1]

1. Manchester Metropolitan University, Manchester UK

2. Seaquarium of Rhyl, Rhyl, UK

Pinnipeds, including seals, sea lions and walrus, have the most prominent and sensitive whiskers of any mammal [1, 2]. Pinniped whiskers are an active touch sensing system, which means that they are actively controlled to maximise the amount of relevant information they can gather from an object [2]. For example, Pinnipeds identify the texture of an object by making stroking movements with their whiskers, and size by moving their whiskers around the edges of an object [2]. Being able to identify sensory movement strategies by tracking whiskers during tactile exploration can further our understanding of how Pinnipeds sense so successfully within their challenging, underwater environments, addressing questions about their function and their importance for survival.

Video tracking allows us to track whiskers and quantify their movements by calculating metrics, such as whisker angular position, spread, amplitude and asymmetry. Studies often use automated whisker tracking systems including the Automated Rodent Tracker (ART) [3], Biotact Whisker Tracking Tool [4] and WhiskEras [5], but these only work in constrained environments, such as laboratories, with controlled backlighting; and are only really designed to work on rodents. When it comes to a zoo or field setting, and animals that are not rodent-like, manual tracking is adopted, such as the Manual Whisker Annotator (MWA), [6]. However, manual tracking is labour intensive and time consuming, which can limit sample sizes [2]. Working within a zoo or field setting, is much less controlled, with backgrounds, lighting and accessibility all varying between individuals and enclosures [7]. However, working in these environments is important for accessing more diverse species for comparative and evolutionary behavioural studies.

DeepLabCut (DLC) is an AI-based markerless tracking system, developed to extract the posture of animals, using anatomical landmarks, and especially when backgrounds are dynamically changing [8,9]. This makes DLC an ideal candidate to explore zoo-based behaviour research. DLC is a well-established method, and has been showcased on fast moving sensory systems, including the rodent pupil [10], rodent forelimbs [11] and rodent whisker movements [12]. DLC can also be used in relatively dynamic environments, such as tracking fish swimming underwater, from overhead cameras [12]. However, DLC has yet to be applied underwater in a challenging zoo environment, where cameras are submerged underwater, resulting in footage with air bubbles, changeable lighting and dark shadows, which generates increased visual noise into the images. With AI at the forefront of scientific research it is important to validate the method fully before applying it. The aim of this study is to test DLC with a challenging use-case tracking South African fur seal whiskers, underwater in a zoo. To do this we will initially train a network, to validate DLC, by firstly checking how DLC performs in relation to the manually selected landmarks, by test-training the data set and extracting pixel errors within that network. Secondly, by validating the metrics of the whiskers including angular position, spread, amplitude and asymmetry, via extracting them from the landmarks and interpreting the accuracy of those metrics. Therefore, validating DLC using both the calculated pixel error and by calculating the whisker metrics with our own code, using landmarks from DLC. As a behavioural scientist the validation of accurate metrics are just as important, therefore, validating not only DLC but also the markerless workflow that follows is essential. This work has ethical approval from both committees at Manchester Metropolitan University and SeaQuarium Rhyl.

Video footage from a group of four South African fur seals *(Arctocephalus pusillus*) housed at Seaquarium Rhyl, was collected using two GoPro Hero10 cameras (120 fps) to capture whiskers movements underwater while the fur seals completed a visual discrimination task (see Figure 1a, b). The cameras were positioned on a rig submerged underwater recording from a top-down view and a side view (see Figure 1). Video footage was reviewed and selected for tracking, including examples where there was no yaw movements by the seal resulting in all the whiskers being visible on both sides of the face, from approach to contact with the stimuli; the seal did not pre-emptively choose the target before the rig was placed in the water, the seal gave the correct answer and there was suitable lighting to see all the whiskers. DeepLabCut version 2.3.6 [13] was used to annotate landmarks recorded from the two cameras. For the top-down view fourteen points were selected, including the nose and head of the fur seal, and two points on a total of six whiskers spread across the mystacial pad; one point at the base of the selected whiskers and one point on the whisker shaft, (see Figure 1b). For the side view, a total of eight points were selected including the head, nose and two points along three whiskers; again, one point at the base of the selected whiskers and one point on the whisker shaft, (see Figure 1a). The whiskers present a unique challenge. While it is easy to locate a whisker base point, the shaft point needs to be a relatively repeatable landmark on the whisker shaft (i.e., a point on a curve), (see Figure 1c). We began by generating a training dataset with three hundred frames, using a p-cutoff of 0.6, specifying the threshold of the likelihood and helps distinguish likely body parts from uncertain ones, meaning the network plots points it is confident about [13]. Preliminary findings suggest that DLC could not consistently track all the whiskers. Some points which, included the more delicate whisker shaft and the head were continuously not found by the trained network. Several environmental factors, including bubbles, limited lighting and shadows, introduced challenging visual noise into the dataset, challenging marker positioning (see Figure 1c). However, refining the network by manually correcting and adjusting the labels improved the feature detectors, increasing the training dataset by an additional one hundred frames, resulting in an improvement, and could potentially offer a useful alternative to manual tracking (see Figure 1d).

Initial observations from our study show encouraging results from the application of DLC to a challenging underwater whisker tracking use-case. Additional challenges, such as lighting and shadows can be reduced by adding in underwater lighting, but additional desensitisation of the animals used would be needed. Furthermore, training a data set that has more challenging backdrops or visual noise could generate a more robust network, instead of those set under laboratory conditions. Finally, having the ability to generate and create an underwater pre-trained network for Pinnipeds could reduce the extensive labour and time required for tracking videos manually. DLC reveals positive findings in this challenging use-case, suggesting that its possible applications are vast. Being able to track posture and movement in animals within an unconstrained zoo or field setting will enable behavioural researchers to conduct comparative studies with higher sample sizes, which has important applications for the study of evolutionary behaviour and function.

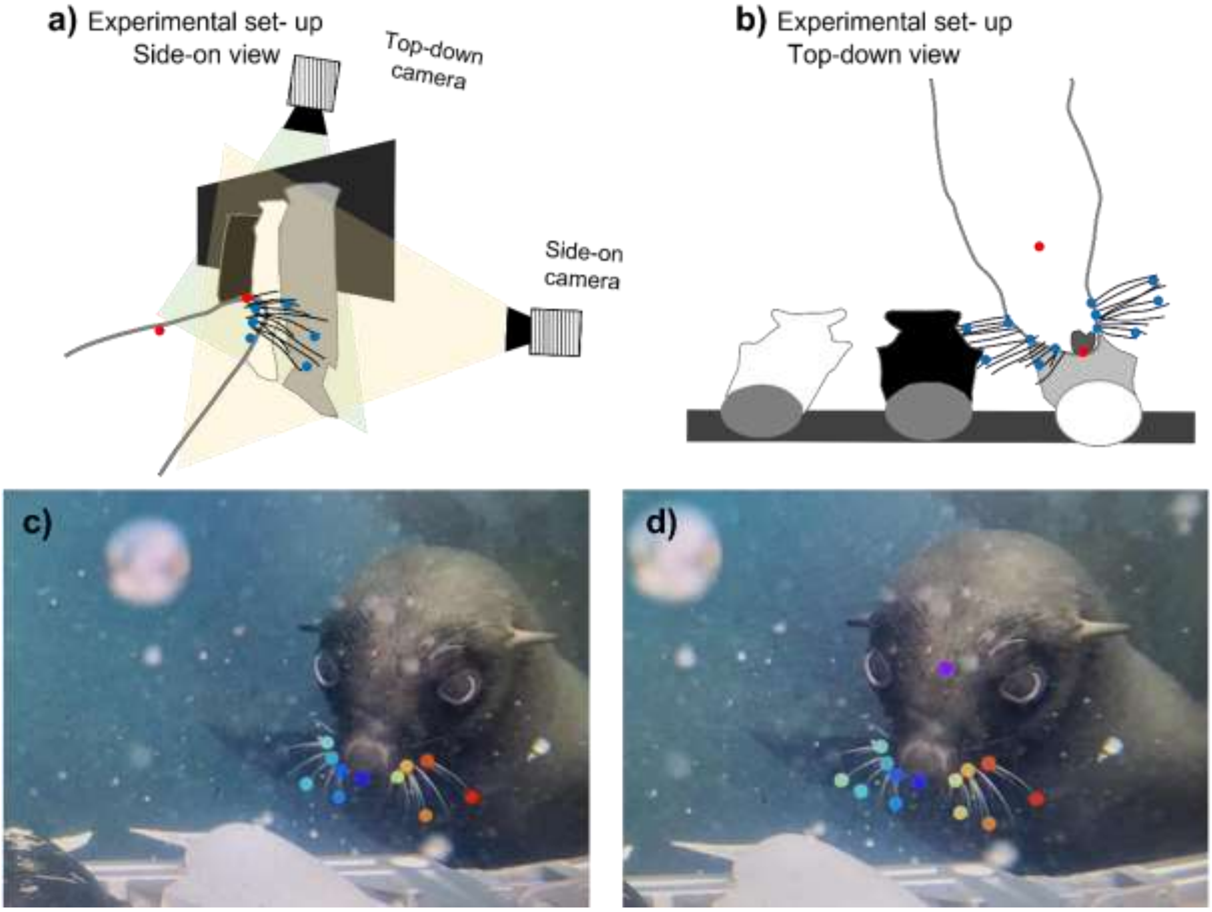


Figure 1:
Tracking South African fur seals whiskers using DeepLabCut, while completing a visual discrimination task underwater: a) Discrimination Task set up, with camera positions, showcasing the side view tracking points on the fur seal b) Discrimination Task set up, showcasing the top-down view tracking points on the fur seal; c) Screen shot of initial pre-trained results using DeepLabCut, with three points missing; d) Screen shot of re-trained data in DeepLabCut with all points now evident and much improved tracking positioned. Coloured markers indicate tracking points including nose, head, and three whiskers on each side of the face, each with a base and shaft point.

References

1. Grant and Goss (2022). What can whiskers tell us about mammalian evolution, behaviour, and ecology? Mamm. Rev., 52 (2022), pp. 148-163, 10.1111/MAM.12253.

2. Milne, A.O., Jones G.C., Black, C.H, Orton, L.D., Sullivan, M.S. and Grant, R.A. (2021) Active touch sensing in the California sea lion (*Zalophus californianus*). Journal of Experimental Biology. https://doi.org/10.1242/jeb.243085.

3. Hewitt, B. M., Yap, M. H., Hodson-Tole, E. F., Kennerley, A. J., Sharp, P. S., & Grant, R. A. (2017). A novel automated rodent tracker (ART), demonstrated in a mouse model of amyotrophic lateral sclerosis. Journal of Neuroscience Methods. https://doi.org/10.1016/j.jneumeth.2017.04.006 .

4. Perkon, I., Košir, A., Itskov, P.M., Tasič, J. and Diamond, M.E., (2011). Unsupervised quantification of whisking and head movement in freely moving rodents. Journal of Neurophysiology, 105(4), pp.1950-1962.

5. Betting JLF, Romano V, Al-Ars Z, Bosman LWJ, Strydis C, De Zeeuw CI. WhiskEras: A New Algorithm for Accurate Whisker Tracking. Front Cell Neurosci. 2020 Nov 17;14:588445. doi: 10.3389/fncel.2020.588445. PMID: 33281560; PMCID: PMC7705537.

6. Hewitt, B, Yap, MH and Grant, RA (2016) Manual Whisker Annotator (MWA): A Modular Open-Source Tool. Journal of Open Research Software, 4 (1). ISSN 2049-9647.

7. Grant, R.A., Ryan, H., Breakell, V. (2023) 'Demonstrating a measurement protocol for studying comparative whisker movements with implications for the evolution of behaviour.' *Journal of Neuroscience Methods*, 384pp. 109752-109752.

8. Mathis, A., Mamidanna, P., Cury, K.M. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* 21, 1281–1289 (2018). https://doi.org/10.1038/s41593-018-020.

9. Nath, T., Mathis, A., Chen, A.C. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* 14, 2152–2176 (2019). https://doi.org/10.1038/s41596-019-0176.

10. Privitera M., Ferrari K.D., von Ziegler L.M., Sturman O., Duss S.N., Floriou-Servou A., Germain P.L.L., Vermeiren Y., Wyss M.T., De Deyn P.P., et al. A complete pupillometry toolbox for real-time monitoring of locus coeruleus activity in rodents. *Nat. Protoc.* 2020;15:2301–2320. doi: 10.1038/s41596-020-0324-6.

11. Bova, A., Kernodle, K., Mulligan, K., Leventhal, D. Automated Rat Single-Pellet Reaching with 3-Dimensional Reconstruction of Paw and Digit Trajectories. <em>J. Vis. Exp.</em> (149), e59979, doi:10.3791/59979 (2019).

12. Lauer, J., Zhou, M., Ye, S. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat Methods* 19, 496–504 (2022). https://doi.org/10.1038/s41592-022-01443-0.

13. DeepLabCut, https://deeplabcut.github.io/DeepLabCut/docs/recipes/DLCMethods.html

Enhancing Understanding Of An AI-based Markerless Tracking Approach For Gait Analysis In Domestic Dogs (*Canis lupis familiaris*)

H. Gill[1], R. Grant[1], J. Gardiner[1], K. Bates[2], J. Charles[2], C. Brassey[1]

1Department of Natural Sciences, Manchester Metropolitan University, Manchester, United Kingdom.
H.gill@ad.mmu.ac.uk, Robyn.Grant@mmu.ac.uk, James.Gardiner@mmu.ac.uk, C.Brassey@mmu.ac.uk

2Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, United Kingdom.
K.T.Bates@liverpool.ac.uk, J.Charles@liverpool.ac.uk

Kinematic gait analysis typically relies on infrared motion capture (mo-cap) cameras and reflective surface markers placed on specific anatomical landmarks. However, this is mostly applicable to domestic animals in laboratory settings, due to having to physically apply markers to the animal. An emerging field of study, offering a potential alternative to the traditional mo-cap systems, includes markerless AI-based motion capture. This method has the potential to overcome several of the challenges faced by traditional mo-cap [1], operating beyond standard lab conditions, without the need for marker attachment, which would make it possible to obtain kinematic parameters from under-researched species, including in a zoo or field setting. This technology has great potential to benefit animal welfare. However, the validation of AI-based motion tracking for animal gait analysis is still in its infancy as a field of research. Domestic dogs (*Canis lupis familiaris*) are a well-studied group within the field of animal biomechanics, particularly with respect to their gait [2]. Given their steady temperament and trainability for consistent data collection, alongside their diverse body shapes, dogs present a useful model for assessing AI-based tracking.

The purpose of this research was to assess the machine learning tool DeepLabCut (DLC) [3] for extracting parameters of canid gait (see Figure 1), by comparing the output to results collected from a markered motion capture system. The focus of this talk pertains to the methodological process used in deploying bespoke neural networks in animal tracking research. Within this we aim to discuss the challenges associated with conducting markerless motion tracking research, including considering the trade-off between time investment and performance of neural networks, as well as external factors, such as differences in coat colouration. Finally, we aim to present agreement between AI-based markerless tracking and markered motion-capture findings. The research has been granted ethical approval from the committees at Manchester Metropolitan University, the host institution, and the University of Liverpool, the site of data collection. Kinematic data from 35 individuals was collected across seven breeds of domestic dog of varying body shapes, to include Border collie, Dachshund, English spaniel, French bulldog, German shepherd, Labrador, and West highland terrier. Using a BlackMagic 4k Pocket Camera, we captured on-lead steady-state movement at 120 fps in the sagittal plane. Each individual was filmed walking with and without infrared markers attached. Three methods of landmark coordinate extraction were then applied to the video data. The first method of extraction involves the use of traditional mo-cap software on footage with reflective surface markers attached to the animal. The second method used DLC's freely available pretrained 'SuperAnimal-quadruped' network [4], trained on >40k landmarked images of a broad range of mammalian quadrupeds, on markerless footage from our dataset. Finally, using DLC (version 2.3.6) [3,5] we generated a bespoke convolutional neural network, trained on ~500 landmarked images of 35 individuals from our own markerless lab dataset (Using a 95:5 split for training and evaluation, respectively). This network was then used to analyse videos from similar experimental settings. DLC requires a user-defined confidence score to condition the X,Y coordinates of each tracked landmark, denoted by a p-cutoff in this research of 0.6. Landmark tracking scored by DLC above this confidence threshold is viewed as a positive result. Several kinematic variables, such as stance time and joint angles, were extracted using custom Matlab code to assess the agreement between mo-cap and AI-based techniques. Preliminary findings suggest pretrained networks, without additional training, may not be capable of tracking canid gait. However, bespoke networks offer potential as a legitimate alternative to marker-based tracking. The test pixel error from this analysis (~4) falls below the value denoted as satisfactory (<5) in recent work published on tracking rodent movement [3]. In addition, overlap between certain kinematic parameters calculated from markered motion capture and bespoke-network tracking, namely duty factor, provides further encouragement for the potential of markerless tracking. However, reliably and accurately tracking foot swing still proves challenging for markerless tracking software.

Validating the use of AI-based markerless motion tracking methods have numerous benefits to both lab and field research, across several fields of science. Given the removal of physical markers, markerless tracking in lab environments could yield valuable additional time for researchers, offering opportunities to increase both the number and diversity of species. Including additional breeds of domestic dog, to encompass individuals with diverse coat patterns and morphometrics is a key step in the continued validation of this technology. In captivity, accurate markerless tracking of zoo species could lead to advancements in gait analysis for the benefit of animal welfare, in addition to providing automated sources of data important for captive species management, such as space utilisation, activity budget and behavioural repertoire. Finally, identifying appropriate areas for application of pretrained networks would off-set the extensive labour and time investment typically required for neural networks, making this technology more accessible to industry professionals. However, given the relatively recent emergence of AI-based markerless tracking, more robust validation is required before researchers can begin reliably utilising this method of data extraction from animal motion.

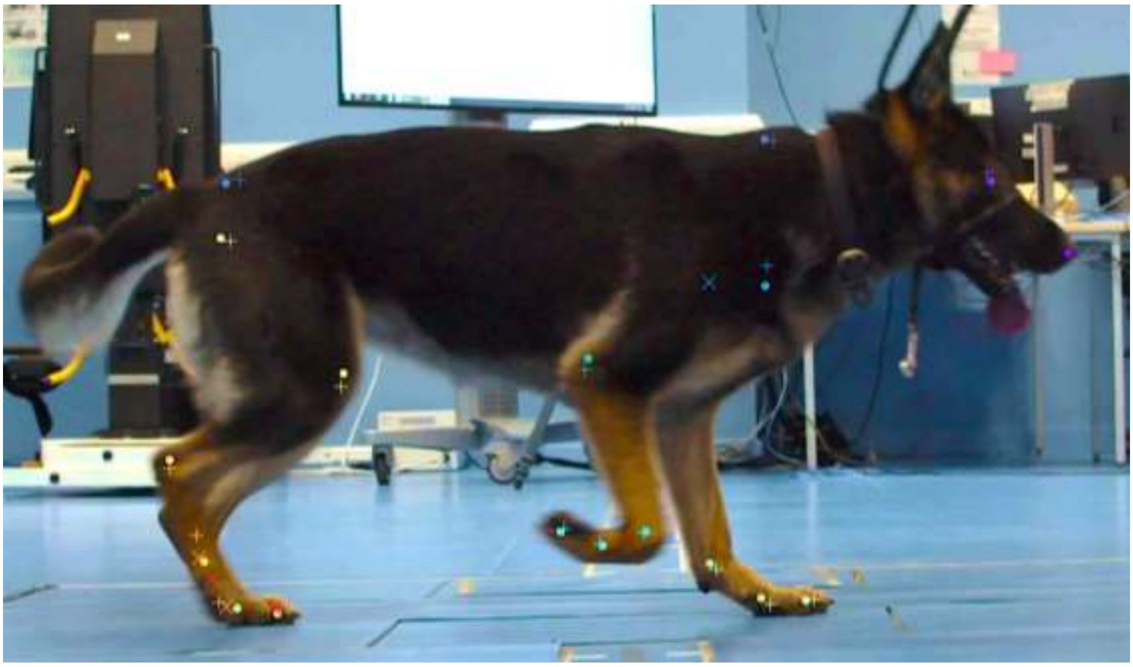


Figure 1 Diagram presenting validation methods of AI-based markerless tracking. The image depicts a snapshot of our bespoke convolutional neural networks estimated landmark locations from a lab-based gait trial on a domestic dog (German shepherd). Human-tracked landmark locations are denoted by the solid-coloured circles, and DLC landmarks are represented by the cross and plus symbols. The plus symbol represents DLC's landmark tracking when this surpassed the predetermined confidence threshold (p>0.6), with the cross-symbol highlighting tracking below this limit of confidence.

References

1. Sellers, W. I., & Hirasaki, E. (2014). Markerless 3D motion capture for animal locomotion studies. *Biology open*, ***3***(7), 656-668.

2. Sandberg, G.S., Torres, B.T. and Budsberg, S.C. (2020). Review of kinematic analysis in dogs. *Veterinary Surgery*, ***49***(6), 1088-1098.

3. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W. and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, ***21***(9), 1281-1289.

4. Ye, S., Filippova, A., Lauer, J., Vidal, M., Schneider, S., Qiu, T., ... & Mathis, M. W. (2022). SuperAnimal models pretrained for plug-and-play analysis of animal behavior. *arXiv preprint arXiv:2203.07436*.

5. Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature protocols*, ***14***(7), 2152-2176.

Symposium: Novel Methods in Measuring Animal Affective States

Using Facial Action Unit and Geometric Morphometric approaches to quantify animal facial movements and their interpretability in the context of affective states

L.R. Finka

Cats Protection, Haywards Heath, United Kingdom

Short abstract

This talk will cover some of the practical, methodological, empirical and conceptual challenges as well as solutions in the quantification of affect based on animals' facial appearances. Examples will include the development of Facial Action Coding Systems (FACS) informed Geometric Morphometric approaches in the context of pain assessment in domestic cats. The benefits of machine learning applications and suggested future directions for broader welfare assessment using these methods will also be discussed.

Paper

The development of objective, standardized methods to facilitate accurate and reliable evaluation of animal affective states is the holy grail of animal welfare research. A very broad range of approaches spanning many scientific disciplines and specialisms have contributed to this diverse field of work. In particular, there is a growing body of research focused on the social and emotional relevance of facial shape changes across various mammalian taxa. Facial Action Coding Systems (FACS) have proved a very useful method in the categorization of the movement of groups of specific facial muscles via 'Action Units'. These systems are species-specific and enable the objective manual coding of the more common facial shape changes that are exhibited in a given species across various social, emotional and functional contexts [1-3] and also enable cross-species comparisons [4-5]. However, FACS approaches to measuring animal facial movements are time and labour intensive and potentially subject to human coder error. These systems of coding have primarily been developed with the intention that they are applied to dynamic visual observations of the face rather than to static single representations, and the coding methods do not easily facilitate the quantification of movement intensity, merely whether movements have occurred or not. Additionally, due to the inherent complexity of facial musculature, the repertoire of current species-specific unit and descriptor codes available by which to label facial movements are unlikely to be fully exhaustive. In relation to the measuring of specific affective states, 'Action Unit/Descriptor' inspired approaches (i.e. grimace scales) have been used to differentiate between typically static examples of facial feature changes across variations in clinically induced and controlled and also spontaneously occurring painful conditions [6-12]. However, these methods are in the most part based on extrapolations from the human based facial coding systems and are as such not intrinsically grounded in species-specific facial anatomy or relevant Action Units and Descriptors.

Recent applications of species-specific FACS and facial anatomy informed Geometric Morphometric approaches may provide solutions to some of these methodological challenges [13]. Geometric Morphometrics utilize landmark based measurements that can be placed in any location on a target object in order to capture either 2D or 3D representations of shape and subsequently it's subtle variation over a given context, population or temporal sequence. These methods have been pioneered to successfully measure subtle differences in global facial shape changes corresponding to varying intensities of acute post operative pain in domestic cats using manually annotated static images [13]. Although in some respects methodologically quicker and more simplistic than a FACS approach, manual annotation of sufficient images to support statistical analysis is still relatively time and labour intensive as well as impractical to facilitate real time inferences about the affective state of a given animal. An extension of this work has therefore been to develop automated machine learning approaches to image annotation and pain/no pain categorization using the geometric morphometric and also deep learning approaches; these have to date achieved good levels of performance across several different cat pain datasets [14-15].

Despite these considerable methodological advances, the mapping and isolation of specific groups of facial movements (sensu FACS) or geometric configurations against specific affective states is a relatively imprecise science. Leading perspectives within affective neuroscience refute the notion that discrete emotion categories are regionally localized within the brain and that there is a corresponding 'facial expression' system that can directly and reliably capture the outputs of these affective processes [16]. For example, there is the potential for facial movements (and their visual presentation) to have elements of emotional, social and functional overlap, to be concomitantly and/or independently impacted by factors additional to internal affective processes, to vary across contexts, cultures, individual differences such as facial morphology, and to convey social information that is not emotion-specific [17-18]. These issues are further compounded by the fact that animals obviously cannot self report on their internal experiences to allow for more accurate operationalization of contexts where facial movements can be measured in the presence of a given affective state and to a desired degree of intensity. Additional postural and behavioral information as well as social and environmental contextual factors in combination with various validated psychometric tools (where available) and carefully designed studies may be necessary to facilitate the more precise inferences about the relevance of specific facial movements to different affective experiences. A potential methodological solution to start to empirically disentangle some of these inherent issues in applied contexts is to extend Geometric Morphometric approaches to the whole of the animal body in order to create a more comprehensive representation of posture and movement and to analyze these data in combination with a suite of other sources of behavioral and contextual (i.e. social and environmental) information. Large scale manual annotation of all relevant landmarks on animals' faces and bodies, is however, impractical and therefore similar machine learning solutions to the automation of landmark annotation as well as data categorization will be extremely valuable.

References

1. Bennett, V., Gourkow, N., Mills, D.S. (2017). Facial correlates of emotional behaviour in the domestic cat (Felis catus). *Behavioural processes*. **141**, 342-50.
2. Bremhorst, A., Mills, D.S., Würbel, H., Riemer, S. (2022). Evaluating the accuracy of facial expressions as emotion indicators across contexts in dogs. *Animal cognition*. **25**, 121-36.
3. Correia-Caeiro, C., Burrows, A., Wilson, D.A., Abdelrahman, A., Miyabe-Nishiwaki, T. (2022). CalliFACS: the common marmoset facial action coding system. *PloS one* **5**.
4. Correia-Caeiro, C., Guo, K., Mills, D.S. (2017). Dogs and humans respond to emotionally competent stimuli by producing different facial actions. *Scientific reports* **7**.
5. Vick, S.J., Waller, B.M., Parr, L.A., Smith Pasqualini, M.C., Bard, K. (2007). A Cross-species Comparison of Facial Morphology and Movement in Humans and Chimpanzees Using the Facial Action Coding System (FACS). *Journal of Nonverbal Behavior* **31**, 1–20.
6. Sotocinal, S. G., Sorge, R.E., Zaloum, A., Tuttle, A.H., Martin, L.J., Wieskopf, J.S., Mapplebeck, J.C.S., Wei, P., Zhan, S., Zhang, S., McDougall, J.J., King, O.D., Mogil, J.S. (2011). The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular pain* **7**, 1744-8069.
7. Keating, S. C., Thomas, A. A., Flecknell, P. A., Leach, M.C. (2012). Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. *PloS one* **7**.
8. Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., Leach, M.C. (2014). Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS one* **9**.
9. Holden, E., Calvo, G., Collins, M., Bell, A., Reid, J., Scott, E.M., Nolan, A.M. (2014). Evaluation of facial expression in acute pain in cats. *Journal of Small Animal Practice* **55**, 615–621.
10. Di Giminiani, P., Brierley, V. L., Scollo, A., Gottardo, F., Malcolm, E. M., Edwards, S. A., & Leach, M. C. (2016). The Assessment of Facial Expressions in Piglets Undergoing Tail Docking and Castration: Toward the Development of the Piglet Grimace Scale. *Frontiers in veterinary science* **3**, 100.
11. McLennan, K. M., Rebelo, C. J., Corke, M. J., Holmes, M. A., Leach, M. C., Constantino-Casas, F. (2016). Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Applied Animal Behaviour Science* **176**, 19–26.

12. Reijgwart, M. L., Schoemaker, N. J., Pascuzzo, R., Leach, M. C., Stodel, M., de Nies, L., Hendriksen, C.F.M., van der Meer, M., Vinke, C.M., van Zeeland, Y. R (2017). The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PloS one* **12**.
13. Finka, L.R., Luna, S.P., Brondani, J.T., Tzimiropoulos, Y., McDonagh, J., Farnworth, M.J. (2019). Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Scientific Reports.* **9**.
14. Feighelstein, M., Shimshoni, I., Finka, L. R., Luna, S. P., Mills, D. S., Zamansky, A. (2022). Automated recognition of pain in cats. *Scientific Reports* **12**.
15. Feighelstein, M., Henze, L., Meller, S., Simshoni, I., Hermoni, B., Berko, M., Twele, F., Schütter, A., Dorn, N., Kästner, S., Finka, L.R., Luna, S.P., Mills, D.S.M., Volk, H.A., Zamansky, A. (2023). Explainable automated pain recognition in cats. *Scientific reports* **13**.
16. Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences* **35**, 121-43.
17. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* **20**, 1-68.
18. Finka, L.R., Luna, S.P., Mills, D.S., Farnworth, M.J. (2020). The Application of Geometric Morphometrics to Explore Potential Impacts of Anthropocentric Selection on Animals' Ability to Communicate via the Face: The Domestic Cat as a Case Study. *Frontiers in Veterinary Science* **7**.

Automated recognition of facial expression of frustration and disappointment in horses during feeding period

Claire Ricci-Bonot[1], Marcelo Feighelstein[2], Hana Hasan[3], Hallel Weinberg[3], Tidhar Rettig[3], Maya Segal[4], Tomer Distelfeld[4], Ilan Shimshoni[2], Daniel S. Mills[1], Anna Zamansky[2]

1. Department of Life Sciences, Joseph Banks Laboratories, University of Lincoln, Lincoln, UK

2. Information Systems Department, University of Haifa, Haifa, Israel

3. Computer Science Department, University of Haifa, Haifa, Israel

4. Faculty of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel

Introduction

In the past animal welfare research was essentially focused on negative emotional states (e.g. fear, distress [1]), with the objective of reducing these to improve the animals' quality of life. However, it is now recognised that animals also need to experience positive emotional states for their wellbeing [2] and research is increasingly focusing on this topic [3,4].

One way of studying these emotions if to use facial expressions, which is a major source of information in many species (cat [5]; cow [6]; pig [7]; rat [8]; sheep [9]). Facial Action Coding System (FACS) developed by Ekman et al. [10] is an objective tool to record facial expressions based on the movements of underlying facial muscles, to avoid bias due to holistic evaluation. It has been developed in several species (cats [11], dogs [12], and several primates (chimpanzees [13], common marmosets [14], gibbon [15], orang outangs [16], rhesus macaques [17]) and horses [18]).

Horses are gregarious animals, who live within a complex social system [19,20], communicating through subtle visual signals (ear positioning, eyes direction and facial expressions) [18,21]. Despite suppositions of them having an emotionally complex world, there are very few research on the recognition of their emotions. Research using facial expressions has mainly been focusing on the context of pain.

A recent research by Ricci-Bonot and Mills [22] has been focusing on identifying facial markers of emotional states during the feeding period. To achieve this goal 30 horses were tested in three different situations, considered positive or negative, involving the potential availability of food: anticipation of food for the positive one, and frustration at waiting for food or disappointment at the loss of it for the negative ones. Horses were tested using a feeding device fixed on the outside of their stable (Figure 1). Video recordings of their face were analysed using Horse FACS (EquiFACS), as well as the horses' behaviours. Researchers found that the occurrence of 9 actions and behaviours differed significantly between the frustration and the disappointment situations, the two scenarios predicted to induce negative emotional states. 'Eye white increase' (AD1), 'ear rotator' (EAD104), and 'biting feeder' had higher likelihood for the frustration situation than the disappointment one. 'Blink' (AU145), 'nostril lift' (AUH13), 'tongue show' (AD19), 'chewing' (AD81) and 'licking feeder' had higher likelihood for the disappointment situation than the frustration one. Specific facial markers associated with anticipation could not be characterised.

An important limitation of these FACS tools is the 'human factor', i.e. the errors and bias that may arise from human coding [23]. It is also time consuming to perform and requires more than one certified coder in order to ensure the reliability of the coding, making it challenging to analyse large volumes of video recordings. Computer Vision based approaches could provide a solution for overcoming these issues [24,25].

In this new study, the dataset of Ricci-Bonot and Mills [22] was used for training AI models for recognition of equine emotional states via two different routes: (1) deep learning, where videos were taken as input, analysed frame by frame and aggregated for prediction of an emotional state, and (2) machine learning, for making a prediction of an emotional state using the EquiFACS coding of the videos as input.

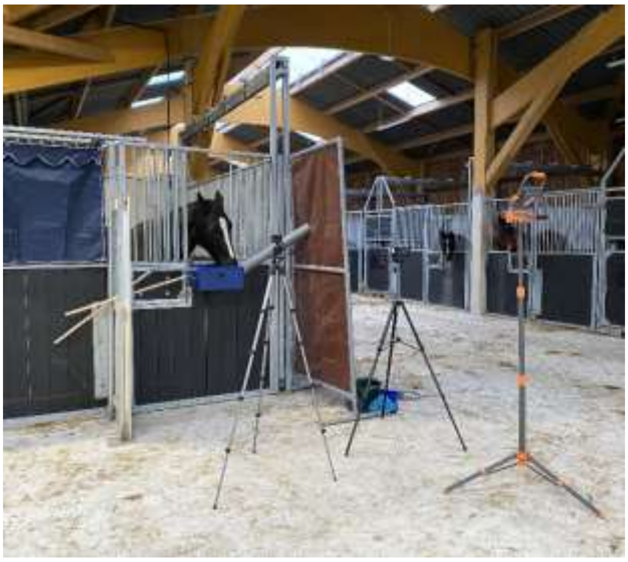


Figure 1: Experimental set-up with the feeder (blue box and grey pipe) attached outside the stable.

Material and methods

The dataset used in this study was collected as part of a previous research by Ricci-Bonot and Mills [22]. This research was approved by the delegated authority of the University of Lincoln Research Ethics Committee (UoL2021_6910). All methods were carried out in accordance with the University Research Ethics Policy and the ethical guidelines of ISAE47. Written informed consent was obtained from the owner of all horses used in the research. No further ethical approval was required for the current computer based work.

Two AI pipelines for classification of the four emotional states (frustration, disappointment, positive anticipation and baseline) were developed. The first deep-learning pipeline, presented on Figure 2 takes as input (3 sec long) videos and uses the GrayST method to incorporate temporal information and a frame selection technique – further details on the approach are described in [26]. The second pipeline takes as input the EquiFACS coding and applies a simple and explainable machine learning technique using decision trees.

Results

The deep learning pipeline outperforms the machine learning one, while the latter offers explainability (in the form of information on which equiFACS action units are important for classification). The former obtains 76% accuracy in separating the four situations: baseline (when the horse is at rest), anticipation, frustration and disappointment, whereas machine learning obtains a score of 69% accuracy. These scores can be improved by combining anticipation and frustration together and asking to differentiate between baseline, anticipation + frustration, and disappointment, resulting in an accuracy of 90% for deep learning and 88% for machine learning. Indeed, anticipation and frustration were more difficult to separate for both deep learning and machine learning systems, with respectively 61% and 46% accuracy.

Discussion

The present study shows that deep learning video based can separate the four situations (baseline, anticipation, frustration and disappointment) with 76% accuracy whereas EquiFACS based reaches 69%. These results suggest that raw video contain more information which have not been captured by the EquiFACS annotation system. Moreover, in the research of Ricci-Bonot and Mills [22], the authors only coded the actions and behaviours as absence/presence, and so including start-end information could also be helpful for the machine. This is emphasised by the fact that the latter only has 46% accuracy in distinguishing anticipation from frustration, difficult cases which could not be separated in the previous study of Ricci-Bonot and Mills [22], whereas deep learning still manages 61% accuracy. However one advantage of the machine learning approach over deep learning is explainability, i.e., understanding the rationale behind the machine's classification of the emotional states [27]. When looking at the machine learning decision tree, we can see the decision making is not only based on facial expressions but also on movements of the head, suggesting that the deep learning system also uses the wider movements of the head.

References

1. Sneddon, L.U., Elwood, R.W., Adamo, S.A., Leach, M.C. (2014). Defining and assessing animal pain. *Animal Behaviour* **97**, 201-212.

2. Taylor, K., Mills, D. (2007). Is quality of life a useful concept for companion animals? *Animal welfare* **16**, 55-65.

3. Duncan, I.J.H. (1996). Animal welfare defined in terms of feelings. *Acta Agriculturae Scandinavica Section A, Animal Science, Supplement* **27**, 29-35.

4. Boissy, A., Arnould, C., Chaillou , E., Désiré, L., Duvaux-Ponter, C., Greiveldinger, L., Leterrier, C., Richard, S., Roussel, S., Saint-Dizier, H., Meunier-Salaün, M.C., Valance, D., Veissier, I. (2007). Emotions and cognition: a new approach to animal welfare. *Animal Welfare* **16**, 37-43.

5. Finka, L.R., Luna, S.P., Brondani, J.T., Tzimiropoulos, Y., McDonagh, J., Farnworth, M.J., Ruta, M., Mills, D.S. (2019). Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Scientific Reports* **9**, 1-12.

6. Sandem, A.I., Braastad, B.O., Bøe, K.E. (2002). Eye white may indicate emotional state on a frustration-contentedness axis in dairy cows. *Applied Animal Behaviour Science* **79**, 1-10.

7. Camerlink, I., Coulange, E., Farish, M., Baxter, E.M., Turner, S.P. (2018). Facial expression as a potential measure of both intent and emotion. *Scientific Reports* **8**, 17602.

8. Finlayson, K., Lampe, J.F., Hintze, S., Würbel, H., Melotti, L. (2016). Facial indicators of positive emotions in rats. *PLoS One* **11**, e0166446.

9. Reefmann, N., Wechsler, B., Gygax, L. (2009). Behavioural and physiological assessment of positive and negative emotion in sheep. *Animal Behaviour* **78**, 651-659.

10. Ekman, P., Friesen, W., Hagar, J. (2002) Facial Action Coding System. Salt Lake City, UT.

11. Caeiro, C.C., Burrows, A.M., Waller, B.M. (2017). Development and application of CatFACS: Are human cat adopters influenced by cat facial expressions? *Applied Animal Behaviour Science* **189**, 66-78.

12. Waller, B.M., Peirce, K., Caeiro, C.C., Scheider, L., Burrows, A.M., McCune, S., Kaminski, J. (2013). Paedomorphic facial expressions give dogs a selective advantage. *PLoS One* **8**, e82686.

13. Vick, S.J., Waller, B.M., Parr, L.A., Smith Pasqualini, M.C., Bard, K.A. (2007). A cross species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS). *Journal of Nonverbal Behavior* **31**, 1-20.

14. Caeiro, C., Burrows, A., Wilson, D.A., Abdelrahman, A., Miyabe-Nishiwaki, T. (2022). CalliFACS: The common marmoset Facial Action Coding System. *PLoS One* **17**, e0266442.

15. Waller, B.M., Lembeck, M., Kuchenbuch, P., Burrows, A.M., Liebal, K. (2012). GibbonFACS: a muscle-based facial movement coding system for hylobatids. *International Journal of Primatology* **33**, 809-821.

16. Caeiro, C.C., Waller, B.M., Zimmermann, E., Burrows, A.M., Davila-Ross, M. (2013). OrangFACS: A musclebased facial movement coding system for Oragnutans (Pongo spp). *International Journal of Primatology* 34, 115-129.

17. Parr, L.A., Waller, B.M., Burrows, A.M., Gothard, K.M., Vick, S.J. (2010). MaqFACS: A muscle-based facial movement coding system for the rhesus macaque. *American Journal of Physical Anthropology* **143**, 625-630.

18. Wathan, J., Burrows, A.M., Waller, B.M., McComb, K. (2015). EquiFACS: The equine facial action coding system. *PLoS One* **10**, e0131738.

19. Feh, C. (2005). Relationships and communication in socially natural herds. In Mills, D., McDonnell, S. (Eds.), *The Domestic Horse, the Evolution, Development and Management of Its Behaviour* (pp. 83-93). Cambridge, UK: Cambridge University Press.

20. Cozzi, A., Sighieri, C., Gazzano, A., Nicol, C.J., Baragli, P. (2010). Post-conflict friendly reunion in a permanent group of horses (Equus caballus). *Behavioural Processes* **85**, 185-190.
21. Wathan, J., McComb, K. (2014). The eyes and ears are visual indicators of attention in domestic horses. *Current Biology* **24**, R677-R679.
22. Ricci-Bonot, C., Mills, D.S. (2023). Recognising the facial expression of frustration in the horse during feeding period. *Applied Animal Behaviour Science* **265**, 105966.
23. Anderson, D.J., Perona, P. (2014). Toward a science of computational ethology. *Neuron* **84**, 18-31.
24. Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology* **36**, 253-263.
25. Cohn, J.F., Ekman, P. (2005). Measuring facial action. In Harrigan, J., Rosenthal, R., Scherer, K. (Eds.), *The New Handbook of Methods in Nonverbal Behavior Research* (pp. 9-64). Oxford, UK: Oxford University Press.
26. Zamansky, A., Feighelstein, M., Luna, S., Silva, N., Trindade, P., van der Linden, D. (2023). Pain Assessment in Animals: human experts, make way for AI!. <https://www.researchsquare.com/article/rs-3402387/latest> Accessed 18th December 2023.
27. Escalante, H.J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baró, X., Ayache, S., Viegas, E., Güçlütürk, Y., Güçlü, U., van Gerven, M.A.J., van Lier, R. (2017). Design of an explainable machine learning challenge for video interviews. *Proceedings of International Joint Conference on Neural Networks (IJCNN)* (Anchorage, AK, USA, 14-19 May 2017), 3688-3695.

A new observational tool for measuring facial movement in gorillas (*Gorilla spp.*): GorillaFACS - The Gorilla Facial Action Coding System

C. Correia-Caeiro[1,2*], R. Costa[3], M. Hayashi[3], A. Burrows[4,5], J. Pater[4], T. Miyabe-Nishiwaki[6], J.L. Richardson[7], M.M. Robbins[8] and K. Liebal[1,2,9]

1Human Biology & Primate Cognition Department, Institute of Biology, Leipzig University, Leipzig, Germany. catia_caeiro@hotmail.com

2Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

3Research Department, Japan Monkey Center, Inuyama, Japan.

4Department of Physical Therapy, Duquesne University, Pittsburgh, USA.

5Department of Anthropology, University of Pittsburgh, Pittsburgh, USA.

6Center for the Evolutionary Origins of Human Behavior (EHuB), Kyoto University, Inuyama, Japan.

7Center for the Advanced Study of Human Paleobiology, Department of Anthropology, The George Washington University, Washington, DC, USA.

8Department of Primate Behavior and Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

9Children & Nature, LeipzigLab, Leipzig University, Leipzig, Germany.

Abstract

Facial expressions are central for communication and emotion expression in social species. Traditionally, facial expressions have been studied in a gestalt way or using subjective labels of emotion. However, this introduces bias in the research questions and misses subtle but important cues, particularly when looking at facial expressions of non-human animals. The Facial Action Coding System (FACS), a widely used tool to measure facial movement in humans, solves these issues, and because it is based on the facial anatomy, allows comparisons between species. It also avoids issues of a priori assumptions of meaning since it is based on observed movement only. Several FACS tools have been published over the years for different species including all apes, but not yet for gorillas. Hence, here we developed a new objective tool to precisely measure facial expressions in gorillas, the GorillaFACS. First, we will briefly introduce what are the FACS tools and then present the step-by-step protocol to develop GorillaFACS, which included the dissection of the gorilla facial musculature, the analysis of spontaneous gorilla facial behavior, and finally, the classification of each individual facial movement of this species. By using objective and anatomically-based tools such as FACS to investigate facial expressions in a comparative and evolutionary way, we not only can better understand human and other animals' evolution of communicative and emotional processes, but also directly apply this knowledge to improve species welfare and human-animal relationships.

Introduction

What is FACS?

Facial expressions are central for communication and emotion expression in social species. Traditionally, facial expressions have been studied in a gestalt way or using subjective labels of emotion. However, this introduces bias in the research questions and misses subtle but important cues, particularly when looking at facial expressions of non-human animals. In addition, facial expressions are short and very subtle behaviours, typically with meaning that varies by context and species [1], which makes them difficult to study.

FACS (Facial Action Coding System) is an observation tool that was developed in the 70s by Paul Ekman and colleagues [2,3] to study the human face from an anatomical point of view. Each facial movement in the human face is produced by a particular muscle that produces observable changes on the face. These appearance changes can be used to identify and code Action Units (AUs), such as for example, AU12 - Lip Corner Puller, produced by the zygomaticus major muscle that pulls the lip corners towards the ears. This FACS tool allows quantification of each movement, and importantly, an objective way of measuring facial movements, without attributing meaning or emotion. In the field of human facial expression research, FACS is considered the gold standard.

Adapting the human FACS to non-human species: the AnimalFACS development

In 2007, Vick and colleagues adapted the human FACS to chimpanzees, and published the first AnimalFACS: the ChimpFACS [4]. Since then, ten other AnimalFACS [5–14] have been created for a range of primates and domestic species (please see www.animalFACS.com for access to all AnimalFACS tools developed till date]. All these AnimalFACS tools are made available online free of charge for anyone to use. Each AnimalFACS includes a training manual with video and picture examples of each facial movement, and requires users to pass a certification test (all managed by the team at www.animalFACS.com). The certification test ensures all FACS coders are reliable between them, independently of the study, experience in the target species or in facial expression research; The certification test ensures the tool can be used in a systematic way. All systems are based on the facial muscles of each species, followed the same methodology for its development, and thus can be used in parallel to compare facial expressions across species.

FACS tools have been published over the years for different species including all apes, but not yet for gorillas. Hence, here we developed a new objective tool to precisely measure facial expressions in gorillas, the GorillaFACS.

Why developing a GorillaFACS for gorillas

Despite being considered to have weaker social bonds than chimpanzees [15], and low facially expressivity among primates [16–18], gorillas often make use of communicative signals to manage intra and inter-group contacts [19–22]. For instance, they often use agonistic signals like prolonged gazing, chest beating, and other visual displays, aimed at managing negative or tense social interactions [23]. Facial expressions, gestures, and body movements are part of these important visual displays, often indicating which interactions are aggressive, sexual, or affiliative [24]. The ability to read and mimic facial expressions between gorillas has been demonstrated to synchronise and promote play interactions, highlighting the significance of facial communication in fostering social bonds and understanding emotional cues within their intricate social dynamics [25]. In addition to intra-specific communication, gorillas also need to rely on such signals during interactions with humans, especially in the context of close proximity to tourists [26] or zoo visitors [27,28]. In fact, gorillas use facial expressions in response to the attentional state of a human experimenter, meeting the criteria of intentional communication with humans [29].

Given this information about gorilla facial expressions, it is surprising we still know so little about this taxa communication. Facial expressions in gorillas have rarely been studied in social contexts, and not at all in interactions with humans. Furthermore, the few studies of gorilla facial expressions either describe facial expressions in a holistic way (i.e., one label for the full facial display that might include several facial movements) or are identified by broadly descriptive terms (e.g., "stare") or emotionally-loaded terms (e.g., "frown"), which make objective comparisons between studies and species difficult. While this initial work on gorilla facial expressions has laid the groundwork for us to understand their diverse and complex behavioural repertoire, to date, a more detailed examination of their facial expressions has yet to be undertaken. Hence, this calls for the development of a more objective tool to accurately record and analyse gorillas' facial expressions, eliminating anthropomorphic and emotional biases [10].

Methods

We employed a three-step methodology to develop the GorillaFACS: the first step consisted of determining the facial muscular plan of the gorilla; in the second step we analysed videos of spontaneous behaviour of gorillas to identify facial movements; finally, in the third step, we combined the anatomical information with the observed facial movements to classify each of these facial movements into Action Units (AUs - movements from mimetic muscles) and Action Descriptors (ADs - movements from non-mimetic muscles, but that affect the AUs appearance changes).

Determination of the facial muscular plan

A facial muscle dissection was performed (by AB and JP) on one gorilla head specimen (*Gorilla gorilla gorilla*) to document the presence, variability, and general morphology of each muscle in comparison with humans [30]. The specimen was stored as part of a research collection at the Duquesne University. The presence and variability in the facial muscles are noted in the Results section - **Table 1**.

Identification of facial movements

The facial movement in gorilla spp. was analysed by watching video recordings (total of approximately 23 hours and 23 minutes) of spontaneous facial movements of gorillas frame-by-frame and extracting short clips to illustrate each AU/AD. The videos analysed included over 200 wild and captive individuals, from all ages, both sexes, and all stages of maturity. The videos included individuals from a variety of populations and three of the four gorilla subspecies (i.e., Western Lowland gorillas housed in zoos, Eastern Lowland gorillas housed at the GRACE sanctuary, as well as Eastern Mountain gorillas from wild populations), and contexts (including potentially positive, negative, and neutral contexts: e.g., resting, grooming, feeding, play, aggression, copulation, human interaction). The videos were reused from other ethically approved research projects unrelated to the present work, or sourced from online public databases (e.g., YouTube.com, Gracegorillas.org, all with a Creative Commons Licence or approval from the video owners). Therefore, this work was purely observational and no negative contexts (e.g., pain, distress) were induced during the current work and/or solely for the purpose of developing GorillaFACS.

Classification of facial movements into Action Units, Action Descriptors and Ear Action Descriptors

Based on the information obtained during the dissection, an effort was made to identify how each muscle functioned in changing the surface appearance of the face by using functional homologies with the human facial muscles. These appearance changes were described with reference to the basic morphological features or landmarks of the face (**Fig. 1**). Hence, the anatomical plan and behavioural video analysis were combined to describe the facial movements observed in the videos using specific directional and anatomical terms. These were then classified according to codes used in previous FACS (including AUs and ADs), following functional muscular homologies wherever possible, or creating new codes whenever the homologies were not identified. All the AUs found in gorillas are presented in **Table 1** along with the corresponding muscles.

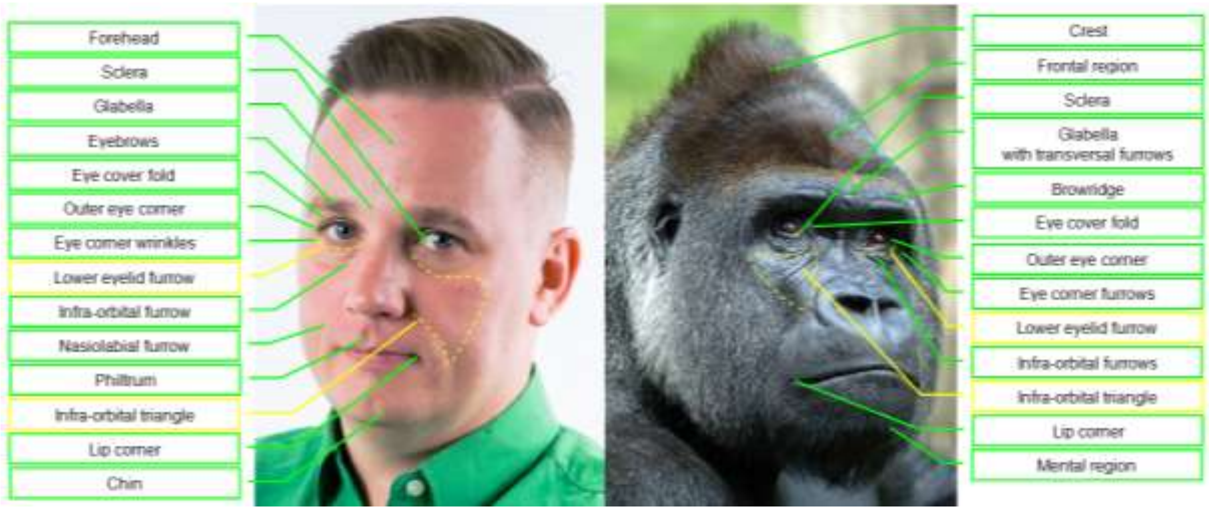


Fig. 1 Comparison of facial landmarks of humans and gorillas (CC licensed images from Pixabay users tarasnesterenko1 and willems_87).

GorillaFACS reliability

We tested inter-observer reliability between two FACS coders (CC: certified in HumanFACS [2] and in all the AnimalFACS developed to date [4–11,31]; RC: certified in ChimpFACS [4]) by coding 25 short clips (not used to describe the AUs). Inter-observer reliability was used to: [1] confirm all coders could reliably identify AUs included on the GorillaFACS manual, and [2] to refine the descriptions of AUs through discussion when agreement between coders on a particular AU was low. This was followed by additional rounds of coding using the same 25 clips to confirm that inter-observer reliability had sufficiently improved based on the improvements of AUs descriptions until an overall level of at least 70% (according to Wexler's index [32]).

Results

The results here presented are intended not only as a short report of the facial movements found when developing the GorillaFACS for gorillas, but will also be published in an extended open access manuscript in the near future, with the description of each AU and AD in detail, along video and picture examples. This will serve as a manual for future GorillaFACS coders to learn to identify these facial movements and a guide for any GorillaFACS coding post-certification.

Table 1. Comparison between FACS Action Units (AU) for humans [2] and gorillas (dissections from the current work and from [33,34]) according to underlying musculature. ✓- present, x - absent. Cells highlighted in grey are present in gorillas.

AU code	AU name	Underlying muscle	Human	Gorilla
AU1	**Inner Brow Raiser**	Frontalis (medial)	✓	x
AU2	**Outer Brow Raiser**	Frontalis (lateral)	✓	x
AU1+2	**Brow Raiser**	Frontalis	✓	✓
AU4	**Brow Lowerer**	Procerus, Depressor supercilii, Corrugator supercilii	✓	✓
AU5	**Upper Lid Raiser**	Orbicularis oculi	✓	x
AU6	**Cheek Raiser**	Orbicularis oculi, pars orbitalis	✓	✓
AU7	**Lid Tightener**	Orbicularis oculi, pars palpebralis	✓	✓
AU43	**Eye closure**		✓	✓
AU45	**Blink**		✓	✓
AD47	**Half-blink**		x	✓
AU8	**Lips Towards Each Other**	Orbicularis oris	✓	x
AU9	**Nose Wrinkler**	Levator labii superioris alaeque nasi	✓	✓
AU10	**Upper Lip Raiser**	Levator labii superioris	✓	✓
AU11	**Nasiolabial Furrow Deepener**	Zygomatic minor	✓	x
AU12	**Lip Corner Puller**	Zygomatic major	✓	✓
AU13	**Cheek Puffer**	Caninus (or Levator anguli oris)	✓	x
AU14	**Dimpler**	Buccinator	✓	✓
AU15	**Lip Corner Depressor**	Depressor anguli oris	✓	x
AU16	**Lower Lip Depressor**	Depressor labii inferioris	✓	✓
AU160	**Lower Lip Relax**	Relaxation of orbicularis oris/lower lip	x	✓
AU17	**Chin Raiser**	Mentalis	✓	x
AU18	**Lip Pucker**	Incisivii labii (superioris and inferioris), Orbicularis oris	✓	✓
AU20	**Lip Stretcher**	Risorius[1]	✓	x
AU21	**Neck Tightener**	Platysma myoides	✓	x
AU22	**Lip Funneler**	Orbicularis oris	✓	✓
AU122	**Lower Lip Inner Curl**		x	✓
AU222	**Lower Lip Extension**		x	✓

AU23	**Lip Tightener**		✓	x
AU24	**Lip Pressor**		✓	✓
AU25	**Lips Parted**	Orbicularis oris, Levator labii superioris, Depressor labii inferioris, non-mimetic muscles	✓	✓
AU26	**Jaw Drop**		✓	✓
AU27	**Mouth Stretch**		✓	✓
AU28	**Lip Suck**	Orbicularis oris	✓	✓
AU38	**Nostril Dilator**	Nasalis	✓	✓
AU39	**Nostril Compressor**	Nasalis, Depressor septi nasi	✓	✓
AU138	**Nose Shield Expander**	Nasalis	x	✓
AU139	**Nose Shield Flattener**	Nasalis	x	✓
AU238	**Nose Downwards**	Nasalis	x	✓

[1] Muscle not identified in the current work dissection, but identified in Diogo et al (2010) **[33]** dissection.

Discussion

The GorillaFACS will be a crucial tool for researchers to investigate the complexity of facial displays in this genus in terms of function and evolution. Furthermore, due to the anatomical nature of FACS, it is possible to apply FACS for each taxon to compare the meaning and context of facial displays between species and determine species-specific facial expression meaning. Gorillas have been fairly well studied in terms of their gestural [35] and vocal repertoire [36], but the facial behaviour repertoire has not yet been as well studied in gorillas as in other primate taxa [37]. Given that facial expressions are important good indicators of welfare status in individuals [38], understanding gorilla communication and emotion has direct applications for the species management and welfare, particularly in captivity where human observers need to constantly monitor individuals health and welfare indicators.

References

1. Correia-Caeiro, C., Guo, K., Mills, D.S. [2017]. Dogs and humans respond to emotionally competent stimuli by producing different facial actions. *Scientific Reports* **7**, 15525.

2. Ekman, P., Friesen, W.V., Hager, J.C. [2002]. Facial Action Coding System [FACS]: manual. Salt Lake City: Research Nexus.

3. Ekman, P., Friesen, W.V. [1978]. Facial coding action system [FACS]: A technique for the measurement of facial actions. Palo Alto, CA: Consulting Psychologists Press.

4. Vick, S.J., Waller, B.M., Parr, L.A., Pasqualini, M.C.S., Bard, K.A. [2007]. A cross-species comparison of facial morphology and movement in humans and chimpanzees using the Facial Action Coding System [FACS]. *Journal of Nonverbal Behavior* **31**[1], 1–20.

5. Correia-Caeiro, C., Waller, B.M., Zimmermann, E., Burrows, A.M., Davila-Ross, M. [2013]. OrangFACS: A muscle-based facial movement coding system for orangutans [*Pongo spp.*]. *International Journal of Primatology* **34**[1], 115–29.

6. Parr, L.A., Waller, B.M., Burrows, A.M., Gothard, K.M., Vick, S.J. [2010]. Brief communication: MaqFACS: A muscle-based facial movement coding system for the rhesus macaque. *American Journal of Physical Anthropology* **143**[4], 625–30.

7. Waller, B.M., Peirce, K., Correia-Caeiro, C., Scheider, L., Burrows, A.M., McCune, S., et al. [2013]. Paedomorphic facial expressions give dogs a selective advantage. *PLOS ONE* **8**[12], e82686.

8. Correia-Caeiro, C., Burrows, A.M., Waller, B.M. [2017]. Development and application of CatFACS: Are human cat adopters influenced by cat facial expressions? *Applied Animal Behaviour Science* **189**, 66–78.

9. Waller, B.M., Lembeck, M., Kuchenbuch, P., Burrows, A.M., Liebal, K. [2012]. GibbonFACS: A muscle-based facial movement coding system for hylobatids. *International Journal of Primatology* **33**[4], 809–21.

10. Correia-Caeiro, C., Burrows, A., Wilson, D.A., Abdelrahman, A., Miyabe-Nishiwaki, T. [2022]. CalliFACS: The common marmoset Facial Action Coding System. *PLOS ONE* **17**[5], e0266442.

11. Wathan, J., Burrows, A.M., Waller, B.M., McComb, K. [2015]. EquiFACS: The Equine Facial Action Coding System. *PLOS ONE* **10**[8], e0131738.

12. Correia-Caeiro, C., Holmes, K., Miyabe-Nishiwaki, T. [2021]. Extending the MaqFACS to measure facial movement in Japanese macaques [*Macaca fuscata*] reveals a wide repertoire potential. *PLOS ONE* **16**[1], e0245117.

13. Julle-Danière, É., Micheletta, J., Whitehouse, J., Joly, M., Gass, C., Burrows, A.M., et al. [2015]. MaqFACS [Macaque Facial Action Coding System] can be used to document facial movements in Barbary macaques [*Macaca sylvanus*]. *PeerJ* **3**, e1248.

14. Clark, P.R., Waller, B.M., Burrows, A.M., Julle-Danière, E., Agil, M., Engelhardt, A., et al. [2020]. Morphological variants of silent bared-teeth displays have different social interaction outcomes in crested macaques [*Macaca nigra*]. *American Journal of Physical Anthropology* **173**[3], 411–22.

15. Cordoni, G., Norscia, I., Bobbio, M., Palagi, E. [2018]. Differences in play can illuminate differences in affiliation: A comparative study on chimpanzees and gorillas. *PLOS ONE* **13**[3], e0193096.

16. Dobson, S.D. [2009]. Allometry of facial mobility in anthropoid primates: Implications for the evolution of facial expression. *American Journal of Physical Anthropology* **138**[1], 70–81.

17. Dobson, S.D. [2009]. Socioecological correlates of facial mobility in nonhuman anthropoids. *American Journal of Physical Anthropology* **139**[3], 413–20.

18. Brann, D. [1999]. The study of facial expressions in gorilla [*Gorilla gorilla gorilla*] communicative behaviors. Ann Arbor, USA: The Union Institute Graduate College, University of Michigan. [Doctoral Thesis].

19. Pika, S., Liebal, K., Tomasello, M. [2003]. Gestural communication in young gorillas [*Gorilla gorilla*]: Gestural repertoire, learning, and use. *American Journal of Primatology* **60**[3], 95–111.

20. Waller, B.M., Cherry, L. [2012]. Facilitating play through communication: Significance of teeth exposure in the gorilla play face. *American Journal of Primatology* **74**[2], 157–64.

21. Tanner, J.E., Byrne, R.W. [1993]. Concealing facial evidence of mood: Perspective-taking in a captive gorilla? *Primates* **34**, 451–7.

22. Tanner, J.E. [2004]. Gestural phrases and gestural exchanges by a pair of zoo-living lowland gorillas. *Gesture* **4**[1], 1–24.

23. Harcourt, A.H., Stewart, K.J. [2007]. Gorilla society: What we know and don't know. *Evolutionary Anthropology: Issues, News, and Reviews* **16**[4], 147–58.

24. Palagi, E., Norscia, I., Cordoni, G. [2019]. Lowland gorillas [*Gorilla gorilla gorilla*] failed to respond to others' yawn: Experimental and naturalistic evidence. *Journal of Comparative Psychology* **133**, 406–16.

25. Bresciani, C., Cordoni, G., Palagi, E. [2022]. Playing together, laughing together: rapid facial mimicry and social sensitivity in lowland gorillas. *Current Zoology* **68**[5], 560–9.

26. Costa, R.F.P., Romano, V., Pereira, A.S., Hart, J.D.A., MacIntosh, A., Hayashi, M. [2023]. Mountain gorillas benefit from social distancing too: Close proximity from tourists affects gorillas' sociality. *Conservation Science and Practice* **5**[1], e12859.

27. Lewis, R.N., Chang, Y.M., Ferguson, A., Lee, T., Clifforde, L., Abeyesinghe, S.M. [2020]. The effect of visitors on the behavior of zoo-housed western lowland gorillas [*Gorilla gorilla gorilla*]. *Zoo Biology* **39**[5], 283–96.

28. Carder, G., Semple, S. [2008]. Visitor effects on anxiety in two captive groups of western lowland gorillas. *Applied Animal Behaviour Science* **115**[3], 211–20.

29. Botting, J., Bastian, M. [2019]. Orangutans [*Pongo pygmaeus* and hybrid] and gorillas [*Gorilla gorilla gorilla*] modify their visual, but not auditory, communicative behaviors, depending on the attentional state of a human experimenter. *International Journal of Primatology* **40**[2], 244–62.

30. Diogo, R., Wood, B.A., Aziz, M.A., Burrows, A. [2009]. On the origin, homologies and evolution of primate facial muscles, with a particular focus on hominoids and a suggested unifying nomenclature for the facial muscles of the Mammalia. *Journal of Anatomy* **215**[3], 300–19.

31. Correia-Caeiro, C., Holmes, K., Miyabe-Nishiwaki, T. [2021]. Extending the MaqFACS to measure facial movement in Japanese macaques [*Macaca fuscata*] reveals a wide repertoire potential. *PLOS ONE* **16**[1], e0245117.

32. Wexler, D.A. [1972]. Method for unitizing protocols of descriptions of emotional states. *Journal of Supplemental Abstracts Service, Catalogue of Selected Documents in Psychology, American Psychological Association* **2,** 116.

33. Diogo, R., Potau, J.M., Pastor, J.F., dePaz, F.J., Ferrero, E.M., Bello, G., et al. [2010]. Photographic and Descriptive Musculoskeletal Atlas of Gorilla: With Notes on the Attachments, Variations, Innervation, Synonymy and Weight of the Muscles. Taylor & Francis. 120 p.

34. Rotenstreich, L., Marom, A. [2023]. 'Untying the knot': The primitive orofacial muscle architecture in the gorilla [*Gorilla gorilla*] as a key to the evolution of hominin facial movement. *FASEB Journal* **37**[9], e23137.

35. Pika, S., Liebal, K., Tomasello, M. [2003]. Gestural communication in young gorillas [*Gorilla gorilla*]: Gestural repertoire, learning, and use. *American Journal of Primatology* **60**[3], 95–111.

36. Salmi, R., Hammerschmidt, K., Doran-Sheehy, D.M. [2013]. Western Gorilla Vocal Repertoire and Contextual Use of Vocalizations. *Ethology* **119**[10], 831–47.

37. Waller, B.M., Whitehouse, J., Micheletta, J. [2017]. Rethinking primate facial expression: A predictive framework. *Neuroscience & Biobehavioral Reviews* **82,** 13–21.

38. Descovich, K., Wathan, J.W., Leach, M.C., Buchanan-Smith, H.M., Flecknell, P., Farningham, D., et al. [2017]. Facial expression: An under-utilized tool for the assessment of welfare in mammals. *ALTEX: Alternatives to Animal Experimentation* **34**[3], 409–29.

Dogs' Tail as a Metronome of Their Emotional States and its Use as a Communicative Tool: Effects of Intra- and Inter-specific Audiences During a Frustration Condition

Y. Ouchi[1,2,*], G. Pedretti[3,*], F. Range[1], P. Valsecchi[3], S. Marshall-Pescini[1,†] and T. Monteiro[1,4,†]

[1]Domestication Lab, Konrad-Lorenz-Institute for Ethology, University of Veterinary Medicine Vienna, Vienna, Austria

[2]Degree Program in Intelligent and Mechanical Interaction Systems, University of Tsukuba, Tsukuba, Japan

[3] University of Parma, Department of Medicine and Surgery, Parma, Italy

[4] William James Center for Research, Universidade de Aveiro, Aveiro, Portugal

[*]co-first authors [†]co-senior authors

Abstract

In this study, we aim to investigate tail movement patterns in domestic dogs during a frustration test with the presence or absence of a social partner. The study involves observing dogs in social (with conspecifics and humans) and non-social conditions while analysing the tail position and movement kinematics. Based on previous literature, we hypothesise that the presence of a social partner will elicit distinct tail movements compared to non-social conditions.

Introduction

Domestic dogs communicate through visual signals modifying the position of parts of their body such as legs and tail postures [1, 2], facial expressions [3, 4] and performing displacement behaviours (i.e., lips licking, yawning, scratching, shaking, blink – [5, 6]). The tail is a crucial part of domestic dogs' body, while originally evolved in vertebrates for locomotion, during canids' evolution it seems to have lost its primary function related to agility manoeuvres [7] and, instead, acquired a function linked to visual communication [8]. The tail is one of the main indicators used by behavioural specialists as a proxy for dogs' emotional states and communicative intents [9] and yet the underlying mechanisms of its movements and function are poorly studied. A general interpretation of canids' inner state based on the tail is that, when the tail is carried with a high position, dogs express confidence and/or aggression while, when the tail is carried in a low position, dogs express fear and submissive attitudes [10]. However, the tail conveys information on dogs' emotional states and motivation through a complex interaction between carriage position (high or low), speed of movement, amplitude and side of wagging.

Recent studies showed how a simplistic interpretation of the information conveyed by the tail based on position and occurrence of movement (i.e., wagging: yes/no; [e.g., 4, 11]) might not be accurate and that tail movements can be characterised by more sophisticated measures (e.g., angle of wagging, speed, dynamics; [e.g., [12]). For example, domestic dogs displayed behavioural asymmetry in tail wagging with a right-side bias (in amplitude) when facing humans, while presenting a leftwards bias when faced with an unfamiliar conspecific, but this study investigating lateral biases in tail wagging analysed the angle between tail and body axis, annotated manually, frame-by-frame, from video recordings [13]. This kind of analysis can be performed with a normal camera placed on the ceiling of the experimental room, however it is preferable if the body of the animal is still during the experimental paradigm and the coding procedure is extremely time consuming [2, 13]. New methods involving 3D camera systems and deep learning allow for automatic tracking of the tail of a freely moving dog in a room provided with a complex calibrated system. Although very precise, this tracking system needs a specific set up of 6 to 8 fixed cameras kept in the same position throughout the duration of the experiment [14].

To help face these challenges, we implemented an alternative approach for tracking the tail of a dog in an experimental room using a commercial wide angle camera and an open source deep-learning based framework for animal pose tracking [15]. In this study, we aim to investigate tail movement patterns in domestic dogs during a frustration test with the presence or absence of a social partner. The study involves observing dogs in social (with conspecifics and humans) and non-social conditions while analysing the tail position and movement kinematics. Based on previous literature [13], we hypothesise that the presence of a social partner will elicit distinct tail movements compared to non-social conditions.

Methods

We tested 46 pet dogs (23 males and 23 females) at the laboratory of Domestic Dog Ethology of the University of Parma. Each dog came to the laboratory 4 times on different days (within subject design) and tests were conducted in an indoor room (4x7m) divided by a wooden apparatus with a 50x100 cm open window in the middle at 30cm height from the ground (see [4] for detailed description of the apparatus). One camera (GoPro HERO+) was placed on the top of the apparatus filming the animal from a 45° angle. In the first session dogs underwent habituation trials only. In the remaining three sessions subjects were exposed to three experimental conditions: non-social frustration (FN) condition where the food reward was visible for 5s, then moved away for another 5s before closing the opaque panel; human frustration condition (FH), similar to FN but with a visible human experimenter staring at the food; dog frustration condition (FD), similar to FN but with a visible conspecific staring at the food.

We tracked dogs to acquire their movement trajectories, body orientations, and the tail movements during the stimulus period using SLEAP [15]. A network was trained to detect 7 body parts: nose, head, withers, back, tail-base, tail-middle and tail tip. After outlining the body orientation of the dogs during the different conditions, we will quantify tail movements, amplitude and angular velocity in each condition, testing hypotheses about the differences in the social and non-social conditions and between the conspecific condition and the human condition. We will evaluate possible future steps and potential areas of application such as research on dogs' emotional state and visual communication.

Ethical statement

The experimental procedure was approved by the Animal Welfare committee of the University of Parma in accordance with Animal Welfare organisation (OPBA) and ARRIVE guidelines and regulations (Protocol number PROT. N. 6/CESA /2022). Owners signed informed experimental procedure consent and privacy consent for publishing images and findings.

References

1. Siniscalchi, M., D'Ingeo, S., Minunno, M., Quaranta, A. (2018). Communication in Dogs. Animals 8: 131.

2. Siniscalchi, M., Lusito, R., Vallortigara, G., Quaranta, A. (2013). Seeing Left- or Right-Asymmetric Tail Wagging Produces Different Emotional Responses in Dogs. Current Biology 23: 2279-2282.

3. Kaminski, J., Hynds, J., Morris, P., Waller, B.M. (2017). Human attention affects facial expressions in domestic dogs. Scientific Reports 7: 12914.

4. Pedretti, G., Canori, C., Marshall-Pescini, S., Palme, R., Pelosi, A., Valsecchi, P. (2022). Audience effect on domestic dogs' behavioural displays and facial expressions. Scientific Reports 12: 9747.

5. Mariti, C., Falaschi, C., Zilocchi, M., Fatjó, J., Sighieri, C., Ogi, A., et al. (2017). Analysis of the intraspecific visual communication in the domestic dog (Canis familiaris): A pilot study on the case of calming signals. Journal of Veterinary Behavior 18: 49-55.

6. Pedretti, G., Canori, C., Biffi, E., Marshall-Pescini, S., Valsecchi, P. (2023). Appeasement function of

displacement behaviours? Dogs' behavioural displays exhibited towards threatening and neutral humans. Animal Cognition 26: 943-952.

7. Rottier, T., Schulz, A.K., Söhnel, K., McCarthy, K., Fischer, M.S., Jusufi, A. (2022). Tail wags the dog is unsupported by biomechanical Modeling of Canidae Tails Use during Terrestrial Motion. bioRxiv. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.12.30.522334 (accessed 2024 Apr 16).

8. Leonetti, S., Cimarelli, G., Hersh, T.A., Ravignani, A. (2024). Why do dogs wag their tails? Biological Letters. Available from: https://royalsocietypublishing.org/doi/10.1098/rsbl.2023.0407 (accessed 2024 Apr 16).

9. Tami, G., Gallagher, A. (2009). Description of the behaviour of domestic dog (Canis familiaris) by experienced and inexperienced people. Applied Animal Behaviour Science 120: 159-169.

10. Kleiman, D.G. (1972). Social behavior of the maned wolf (chrysocyon brachyurus) and bush dog (speothos venaticus): A study in contrast. Journal of Mammalogy 53: 791-806.

11. Bremhorst, A., Sutter, N.A., Würbel, H., Mills, D.S., Riemer, S. (2019). Differences in facial expressions during positive anticipation and frustration in dogs awaiting a reward. Scientific Reports 9: 19312.

12. Ren, W., Wei, P., Yu, S., Zhang, Y.Q. (2022). Left-right asymmetry and attractor-like dynamics of dog's tail wagging during dog-human interactions. iScience 25: 104747.

13. Quaranta, A., Siniscalchi, M., Vallortigara, G. (2007). Asymmetric tail-wagging responses by dogs to different emotive stimuli. Current Biology 17: R199-R201.

14. Völter, C.J., Starić, D., Huber, L. (2023). Using machine learning to track dogs' exploratory behaviour in the presence and absence of their caregiver. Animal Behaviour 197: 97-111.

15. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., et al. (2022). SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods 19: 486-495.

The Development of a Facial Landmark Scheme for Dogs

Greta Abele[1], Annika Bremhorst[1], Chiara Canori[2], Nareed Farhat[3], Giulia Pedretti[2], George Martvel[3], Anna Zamansky[3]

1. **Dogs and Science, Switzerland**

2. University of Parma, Italy,

3. University of Haifa, Israel

Facial expressions, characterized by the movements of facial muscles, have been recognized as potential indicators of emotional states in both humans [1] and non-human animals [2]. This link between facial expressions and emotions has spurred growing interest in studying these behaviours within the broader scope of animal emotion and welfare [3]. Furthermore, facial expressions may have evolved as visual communicative signals, informing social partners about the emitters' motivations and future actions [4] and providing a promising vehicle for understanding social interactions in animals.

An essential milestone for the objective analysis of facial expressions was the first anatomically based tool for identifying visually distinguishable facial movements, the Facial Action Coding System (FACS) [5]. This tool led to extensive research on facial behaviours in humans [6, 7]. FACS has been modified for use with primates such as chimpanzees (chimpFACS – 8), macaques (MaqFACS – 9), hylobatids (GibbonFACS – 10) and orangutans (OrangFACS - 11) as well as other non-primate domesticated mammalian species: dogs (dogFACS– 12), cats (catFACS - 11) and horses (EquiFACS - 14).

An increasing number of studies have been conducted on the association between facial expressions in animals and different affective states (including different emotions such as fear, disgust, frustration, happiness, but also pain states – for a review see 15). Even if some attempts have been made to automatize the analysis of facial expressions in animals [16], most of the behavioural coding is still performed by certified FACS coders, employing enormous amount of manual work.

Furthermore, some domesticated animals, such as cats and dogs, present both unique challenges and opportunities in the context of facial movement analysis. Due to the considerable diversity, resulting from selective breeding, cats, and even more so, dogs display a wide variety of breed types and morphological characteristics. This diversity can significantly influence the form of visual communication [17] as well as their facial expressions [18; 12], making the analysis of their facial displays a complex task. This complexity, along with the potential subtlety of facial movements, underscores the need for automated, accurate and objective methods for their measurement and analysis.

Geometric morphometrics is a powerful tool for this purpose, employing landmarks as proxies for shape. This approach has recently been effectively applied to objectively assessing feline facial expressions [20; 21]. These studies developed a 48-landmark scheme grounded in cat facial anatomy. This scheme, illustrated in Figure 1, represents a significant advancement in the objective analysis of feline facial expressions, offering a structured method to understand their emotional and social cues.

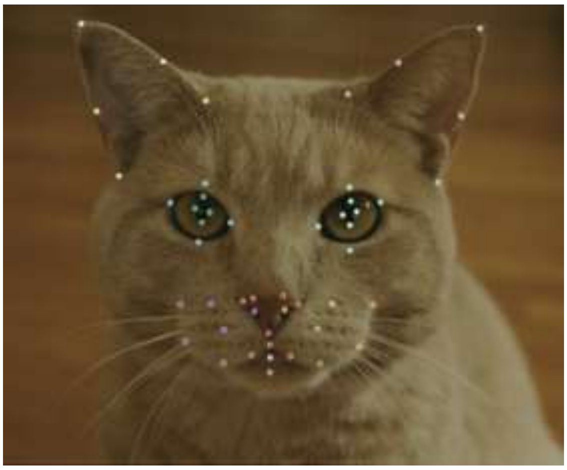


Figure 1: The landmarking scheme of 48 points. Image is taken from [22].

The idea is that the locations of these landmarks represent their relative positions, and differences in these positions across objects indicate the extent of shape variation. Finka et al. [21] used this technique to measure changes in the facial shape of cats associated with pain. They analyzed images of 29 domestic short-haired female cats undergoing surgery (ovariohysterectomy), marking them with 48 specific landmarks corresponding to underlying facial muscles/anatomy, combined with a comprehensive list of cat facial behaviours (as described by the Cat Facial Action Coding System CatFACS [13]) and that are important for cat-specific facial movements. They found a significant correlation between changes in facial shape due to pain and the UNESP-Botucatu MCPS tool [23], a validated pain assessment method. Finka et al. [21] later extended this approach to examine how variations in cat breeds and head shapes affect the positioning of facial landmarks. They observed significant differences in the baseline configurations of facial landmarks among a variety of common domestic cat breeds and head shapes.

These findings indicate that a facial landmark scheme contains meaningful data for quantifying subtle facial changes which can be used for pain recognition, and inter-breed fine-grained comparisons. However, the method is hindered by its reliance on the manual annotation of landmarks, a process both time-consuming and labor-intensive. For example, skilled annotators in Martvel et al.'s study [28] were reported to spend on average over 5.5 minutes annotating a single facial image.

AI models can streamline this process, apart from providing additional benefits. To develop these, facial landmarks can serve as a starting point for the development of automation of pain recognition in cats [Feighelstein et al., 2022, Feighelstein et al., 2023], where AI models used landmark positions as inputs. Using this approach, Martvel et al. [Martvel et al., 2023a] developed an automated detector of cat facial landmarks. To do this, they annotated a large dataset of cat faces [Martvel et al., 2023b] with the 48 landmark scheme and used it to train their model. Its performance was comparable to, and in some cases better than, other models used for human facial landmark localization [Martvel et al., 2023a]. The detector has been already evaluated on a number of benchmark tasks, including breed and cephalic type recognition and pain recognition [27]. Its use has also been evaluated not only on single images, but on whole videos [28], which led to more accurate pain estimation.

The aim of the current study is to develop a similar architecture for domestic dogs that will enhance the application of geometric morphometric inspired approaches with an automated facial landmark detector. This is expected to be much more complicated than for cats, due to the extreme variety of breeds and morphological features in these species. For instance, the tip of the ear in floppy-eared dogs is placed very differently than in pointy ears (see Figure 2).

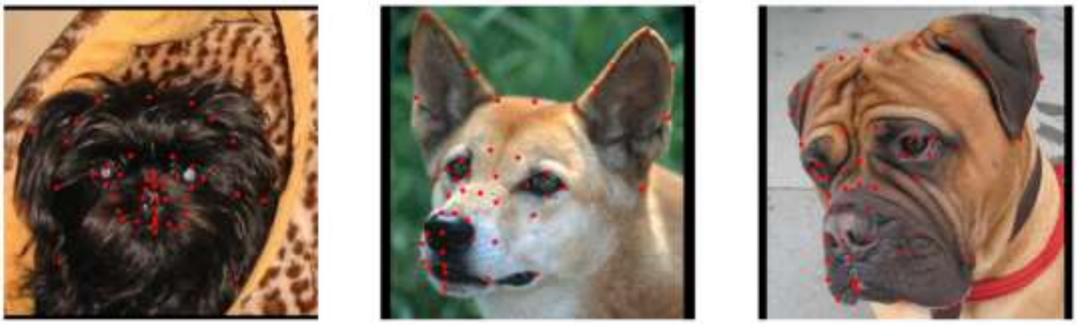


Figure 2: Differences in landmarking scheme for dogs with different morphology. Images are taken from [29].

To this end, we intend to follow the steps that were taken in the development of the infrastructure for cats. This includes the development of a landmark scheme for dogs, the creation of annotated dog faces datasets, and the training of an automated detector for dog facial landmarks. This paper presents our preliminary results, including the development of a 48-landmark scheme. The resulting scheme is presented on Figure 3. Below, we describe the procedure we followed to develop the scheme.

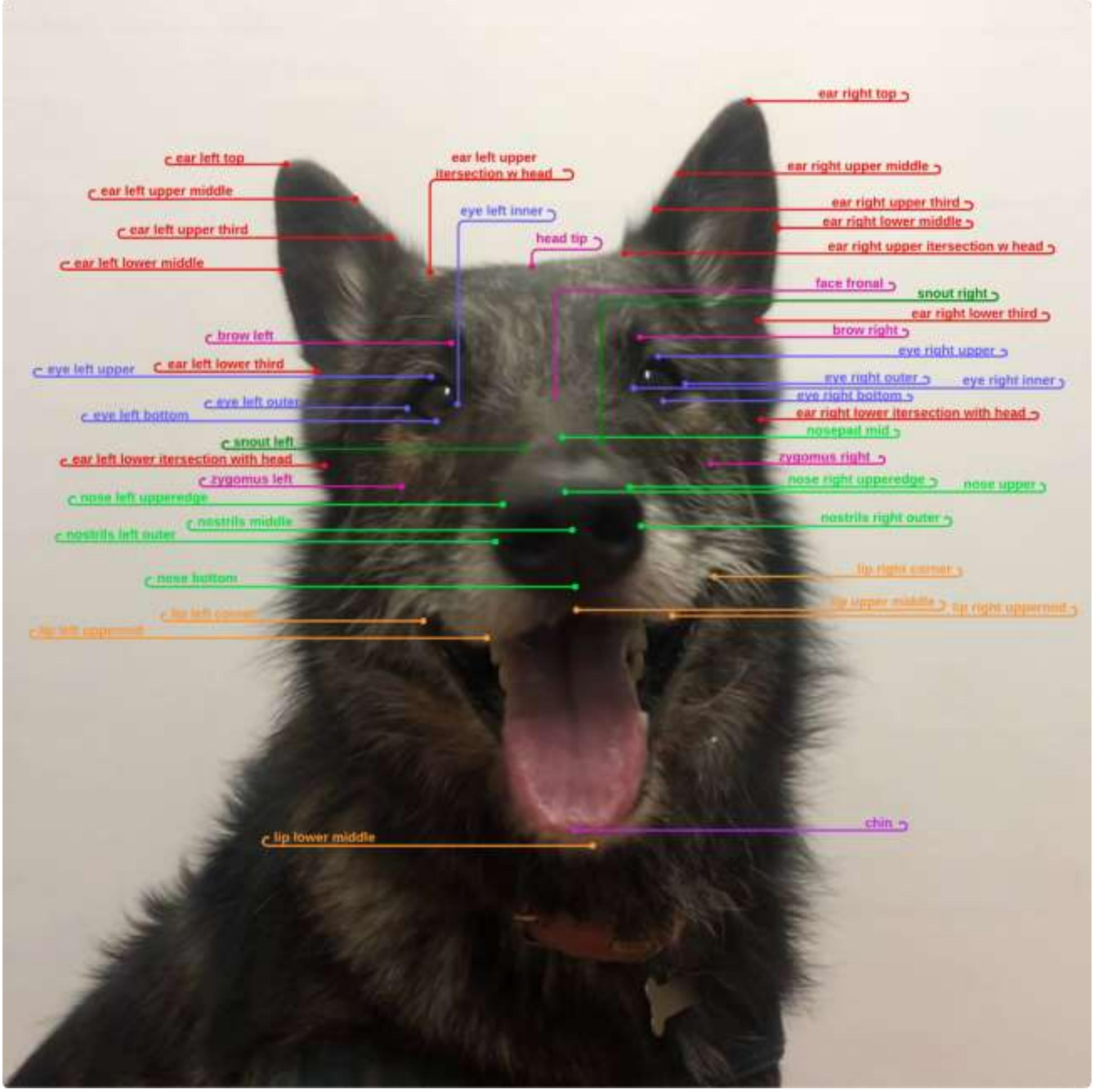


Figure 3: 48 landmarks for dogs

Procedure

The development of the dog facial landmark scheme was undertaken with the aim of creating a comprehensive framework for analyzing canine facial expressions. The first step of the process consisted in determining which points of the facial region of the dogs could be suitable for landmarks placing. In order to establish these points, three canine science experts worked independently, creating their own scheme of landmarks. These experts have extensive experience in behavioural research, are well-versed with dogs, and are certified coders of the Dog Facial Action Coding System (DogFACS) [Waller et al., 2013]. Analogous to how the cat landmarks development was guided by CatFACS, the dog adaptation was informed by DogFACS. The dog landmark framework was based on the anatomy of canine facial musculature and the range of expressions generated by facial action units as defined byDogFACS.

In the determination of the landmarks, AB was largely inspired by the landmark scheme developed for cats [Finka et al., 2019], adapted to accommodate the anatomical and expressive differences in dogs. The process of AB involved using the Cat Landmark Description File [Finka et al., 2019] as a baseline. This file detailed the facial action units (as described by CatFACS) related to each landmark. For each action unit that was also described by DogFACS, we evaluated whether it had the same muscular basis. If so, the same landmark used for cats was applied to dogs. Additionally, as described by Finka et al., Type II landmarks (not specific to muscle insertions but capturing general facial shape changes) were placed. GP and CC based the determination of landmarks on establishing which dogs' facial point are involved in each DogFACS action units and action descriptors. To do so, GP and CC analyzed each facial region (ears, upper face and lower face) establishing which point from dogs of different morphology could have capture the movements described by the DogFACS manual. Each point was established either as a reference for the general face shape (settling on main bones or skin structures) either because it indicated a bundle of muscles involved in action units and action descriptors movements.

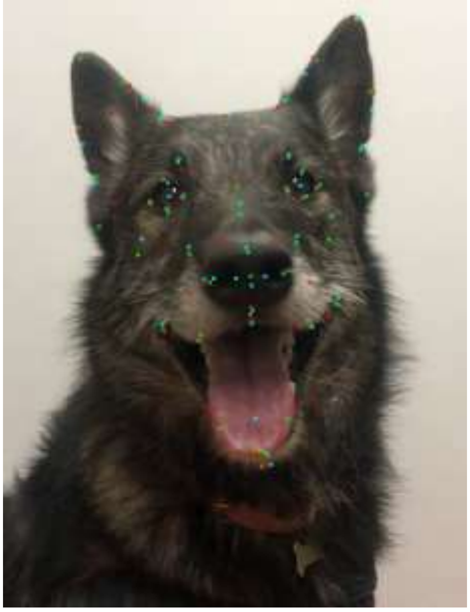


Figure 4: Output of the landmarks determination for the three independent coders. Green points are GP's landmarks, light blue points are CC's landmarks and red point are AB's landmarks.

Convergence and expert agreement

After independently developing the two landmark schemes, the points were converged. The experts compared the three set landmarks and those that were consistent across both approaches were retained, while differing landmarks were discussed thoroughly to reach an expert agreement.

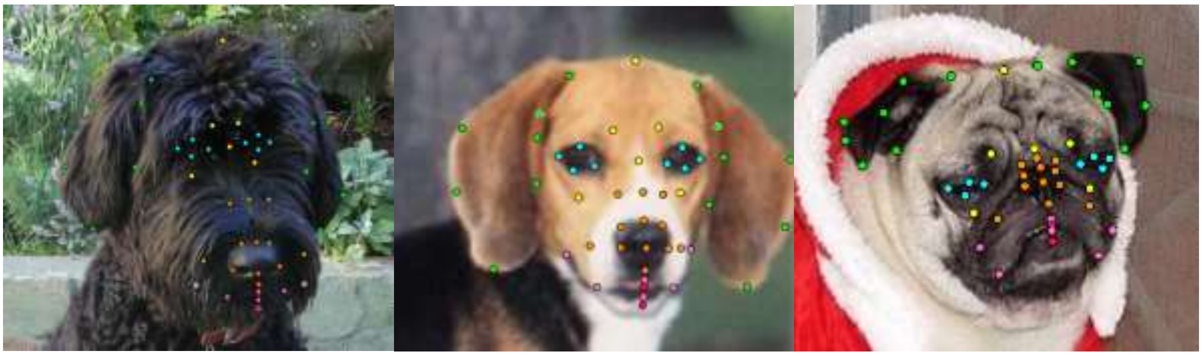


Figure 5: examples of the decided landmarks

Similar to the cat study, we also developed a comprehensive final manual detailing landmarks' placement and its relevance to facial musculature and action units. This manual serves as a guide for accurately annotating canine facial landmarks, ensuring consistency and reliability in future research and applications.

References

1. Ekman, P., & Rosenberg, E. L. (Eds.). (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001
2. Lezama-García, K., Orihuela, A., Olmos-Hernández, A., Reyes-Long, S., & Mota-Rojas, D. (2019). Facial expressions and emotions in domestic animals. *CABI Reviews*, (2019), 1-12.
3. Désiré, L., Boissy, A., & Veissier, I. (2002). Emotions in farm animals: a new approach to animal welfare in applied ethology. *Behavioural processes*, *60*(2), 165-180
4. Crivelli, C., & Fridlund, A. J. (2018). Facial displays are tools for social influence. *Trends in cognitive sciences*, *22*(5), 388-399.
5. Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.
6. Ekman, P. (1973). Universal facial expressions in emotion. Studia Psychologica, 15(2), 140-147.
7. Russell, J. A., Bachorowski, J. A., & Fernández-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual review of psychology*, *54*(1), 329-349.
8. Parr, L. A., Waller, B. M., Vick, S. J., & Bard, K. A. (2007). Classifying chimpanzee facial expressions using muscle action. *Emotion*, *7*(1), 172.
9. Parr, L. A., Waller, B. M., Burrows, A. M., Gothard, K. M., & Vick, S. J. (2010). Brief communication: MaqFACS: a muscle-based facial movement coding system for the rhesus macaque. *American journal of physical anthropology*, *143*(4), 625-630.
10. Waller, B. M., Lembeck, M., Kuchenbuch, P., Burrows, A. M., & Liebal, K. (2012). GibbonFACS: a muscle-based facial movement coding system for hylobatids. *International Journal of Primatology*, *33*, 809-821.
11. Caeiro, C. C., Waller, B. M., Zimmermann, E., Burrows, A. M., & Davila-Ross, M. (2013). OrangFACS: A muscle-based facial movement coding system for orangutans (Pongo spp.). *International Journal of Primatology*, *34*, 115-129.
12. Waller, B., Caeiro, C. C., Peirce, K., Burrows, A., & Kaminski, J. (2013). DogFACS: the dog facial action coding system.
13. Correia Caeiro, C., Waller, B., & Burrows, A. (2013). The Cat Facial Action Coding System Manual (CatFACS).
14. Wathan, J., Burrows, A. M., Waller, B. M., & McComb, K. (2015). EquiFACS: The equine facial action coding system. *PLoS one*, *10*(8), e0131738.
15. Descovich, K. A., Wathan, J., Leach, M. C., Buchanan-Smith, H. M., Flecknell, P., Framingham, D., & Vick, S. J. (2017). Facial expression: An under-utilised tool for the assessment of welfare in mammals. *Altex*.
16. Boneh-Shitrit, T., Feighelstein, M., Bremhorst, A., Amir, S., Distelfeld, T., Dassa, Y., ... & Zamansky, A. (2022). Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. *Scientific reports*, *12*(1), 22611.
17. Goodwin, D., Bradshaw, J. W., & Wickens, S. M. (1997). Paedomorphosis affects agonistic visual signals of domestic dogs. *Animal behaviour*, *53*(2), 297-304.
18. Pedretti, G., Canori, C., Marshall-Pescini, S., Palme, R., Pelosi, A., & Valsecchi, P. (2022). Audience effect on domestic dogs' behavioural displays and facial expressions. S*cientific Reports*, 12(1), 9747. https://doi.org/10.1038/s41598-022-13566-7
19. Waller, B. M., Peirce, K., Caeiro, C. C., Scheider, L., Burrows, A. M., McCune, S., & Kaminski, J. (2013). Paedomorphic facial expressions give dogs a selective advantage. *PLoS one*, *8*(12), e82686.
20. Finka, L. R., Luna, S. P., Brondani, J. T., Tzimiropoulos, Y., McDonagh, J., Farnworth, M. J., ... & Mills, D. S. (2019). Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Scientific reports*, *9*(1), 9883.

21. Finka, L. R., Luna, S. P., Mills, D. S., & Farnworth, M. J. (2020). The Application of Geometric Morphometrics to Explore Potential Impacts of Anthropocentric Selection on Animals' Ability to Communicate via the Face: The Domestic Cat as a Case Study. *Frontiers in Veterinary Science*, *7*, 1070.
22. Martvel, G., Shimshoni, I. & Zamansky, A. Automated detection of cat facial landmarks (2023a). In review, 2310.09793.53.
23. Brondani, J. T., Mama, K. R., Luna, S. P., Wright, B. D., Niyom, S., Ambrosio, J., ... & Padovani, C. R. (2013). Validation of the English version of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in cats. *BMC Veterinary Research*, *9*, 1-15.
24. Feighelstein, M., Shimshoni, I., Finka, L. R., Luna, S. P., Mills, D. S., & Zamansky, A. (2022). Automated recognition of pain in cats. *Scientific Reports*, *12*(1), 9575.
25. Feighelstein, M., Henze, L., Meller, S., Shimshoni, I., Hermoni, B., Berko, M., ... & Zamansky, A. (2023). Explainable automated pain recognition in cats. *Scientific reports*, *13*(1), 8973.
26. Martvel, G., Farhat, N., Shimshoni, I. & Zamansky, A. (2023b). Catflw: Cat facial landmarks in the wild dataset. CV4Animals Workshop, CVPR 2023, arXiv preprint arXiv:2305.04232
27. Martvel, G., Lazebnik, T., Feighelstein, M. et al. Automated Pain Recognition in Cats using Facial Landmarks: Dynamics Matter, 17 December 2023c, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3754559/v1]
28. Martvel, G., Lazebnik, T., Feighelstein, M., Meller, S., Shimshoni, I., Finka, L., Luna, S., Mills, D., Volk, H., & Zamansky, A. (2023). Automated Landmark-Based Cat Facial Analysis and its Applications. 10.21203/rs.3.rs-3776553/v1
29. Khosla, A., Jayadevaprakash, N., Yao, B. and Fei-Fei, L. (2011). Novel dataset for Fine-Grained Image Categorization. First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Symposium: Digital Innovations in Home cage monitoring

Digital Innovations in Home Cage Monitoring: Advancing Animal Welfare and Pharmaceutical Development

Stefano Gaburro[1], Fabrizio Scorrano[2], Michael Tsoory[3], Thomas Svava Nielsen[4]

1Tecniplast S.p.A., Buguggiate, Italy. stefano.gaburro@tecniplast.it

2Novartis AG, Basel, Switzerland. fabrizio.scorrano@novartis.com

3Weizmann Institute of Science, Department of Veterinary Resources, Rehovot, Israel. michael.tsoory@weizmann.ac.il

4Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. thomas.nielsen@sund.ku.dk

Abstract

Rack-based home cage monitoring systems significantly enhance animal welfare and research quality in pharmaceutical studies. These systems offer a non-invasive method to continuously observe animals in their natural environment, adhering to ethical research principles. They are particularly effective in reducing stress-related variables in chronic disease models, thereby improving the reliability of data. Their application demonstrates a notable improvement in operational efficiency and scientific outcomes across the several examples of mouse models presented.

Home cage rack-based technology: welfare and scientific applications

The advent of rack-based home cage monitoring systems initially conceptualized through technologies like Digital Ventilated Cages (DVC®), has ushered in a new era in preclinical research. These systems represent a significant advancement in the ethical study of animal models, particularly in the context of continuous, non-invasive monitoring. Traditional methods of animal monitoring in research have often been limited by the need for intermittent, direct observation or invasive procedures, potentially inducing stress and behavioral alterations in the subjects. The rack-based system circumvents these limitations by allowing for the continuous observation of animals in their natural habitat. This approach aligns with the 3Rs principle, aiming to refine research methods to minimize discomfort and stress to animal subjects. The scalability of these systems is a critical feature, enabling the simultaneous monitoring of a broad range of cages. This aspect is particularly beneficial in large-scale studies where consistent data collection across numerous subjects is paramount and gender differences should be evaluated. The integration of micro electromagnetic field technology ensures that data integrity is maintained, even in environments enriched for the natural behavior of the animals.

The non-intrusive nature of rack-based monitoring systems has profound implications for animal welfare. By reducing the stress associated with traditional monitoring methods, these systems provide a more humane approach to animal research. This aspect is not only ethically significant but also scientifically crucial, as stress can be a confounding factor in behavioral and physiological studies. In terms of research outcomes, the continuous data collection facilitated by these systems offers a more comprehensive understanding of animal behavior and disease progression. This is particularly evident in chronic disease models and infectious disease studies, such as those involving Covid-19, ALS or other disease models where early detection of subtle physiological changes can be critical. As the field of preclinical research continues to evolve, the integration of advanced monitoring technologies like rack-based systems will likely become more prevalent. These systems offer a promising pathway towards more ethical, accurate, and comprehensive animal research. However, it is essential to continuously assess and refine these technologies to ensure they meet the highest standards of animal welfare and scientific rigor. In conclusion, rack-based home cage monitoring systems represent a significant stride forward in the realm of animal research. By enabling continuous, non-invasive observation of animal models, they enhance both the welfare of the subjects and the quality of the data obtained, thereby contributing to the advancement of biomedical research.

In these three examples below we will present several animal models evaluated with the rack-based technology in different settings.

Detecting Sciatic nerve injury (SNI) induced motor impairments and recovery using the DVC system; a core facility unit user's perspective

Laboratory rodent models of nerve injury rely heavily on repeated assessments of motor functions away from the home–cage, where the laboratory animals are often forced to walk to allow assessments of gait and stride, for example. Such assessments lead to substantial disturbance of the animals' routine and cause them some discomfort that might mask the experimental manipulation effects. In addition, these evaluations are labor-intensive and require time-consuming post-hoc analyses. Therefore, the current study sought a home-cage-based alternative and in a series of experiments assessed the DVC® system's capacity to detect sciatic nerve injury-induced motor impairments and recovery dynamics as reflected by changes in voluntary, spontaneous, activity in the home cage.

In the first experiment, DVC® Activations and Distance data were collected from single-housed mice (N=31) over 4 weeks. The first two weeks served as a baseline period. Immediately following the base-line period all mice underwent hind leg unilateral (right) surgical manipulations, either sciatic nerve injury (SNI, hereafter) by crush (n=16), or a control procedure (Sham) that did not inflict any nerve damage yet did involve invasive manipulations of the thigh's skin and muscle (n=15). The home–activity was then recorded for two more weeks in order to identify both the immediate effects of the above-noted surgical manipulations and the 'return to baseline' levels, as an index of recovery.

In comparison with Sham mice, SNI mice exhibited a significantly greater reduction in activity (as compared to baseline) and a slower rate of return to baseline level. These patterns of effects were evident on both DVC® activity indices.

The second experiment juxtaposed the DVC® activity indices with gait and stride data that were collected using the CATWALK system (Noldus; NL). The experiment followed the same timeline as the first one yet included four groups. Two groups, Sham and SNI, were monitored using only the DVC® system. The other two, Sham and SNI, were monitored by the DVC® system, but also underwent frequent gait and stride assessments using the CATWALK (CW, hereafter) system. Overall: 1) Sham DVC only (n=4); 2) SNI DVC only (n=4); 3) Sham DVC + CW (n=4); 4) SNI DVC + CW (n=3).

The stride and gait data (CATWALK indices: paw print area, Stand and Duty Cycle) indicated that as compared with Sham mice, SNI mice exhibited a significantly greater reduction in the use of the manipulated leg (as compared to baseline) and a slower rate of return to baseline level.

The home – cage activity DVC® indices data indicate dynamics of recovery processes, reflected by "return to baseline levels", that correspond to those indicated by stride and gait analyses.

Additional discussion will address the advantages and challenges of monitoring home cage activity from a core facility user's perspective.

Use of rack base technologies in Pharmaceutical Settings

The growing trend in the pharmaceutical industry toward digital home-cage monitoring solutions has become a must and is underscoring their significance in the context of clinical trials and human application. In recent years, there has been a notable shift in pharmaceutical clinical research from centralized clinical trials, conducted in established research sites and hospitals, to more dynamic approaches involving mobile clinics. These mobile setups facilitate patient services like blood sampling, simple tests, and monitoring of physiological parameters. This evolution is not only enhancing the quality and quantity of data collected but also providing broader access to diverse population substrates. It underscores the growing emphasis on comfort, ethics, and the integration of technology in research.

Parallel to these developments in clinical research, a similar transformation is observed in in vivo research with rodents. Initially, rodent studies relied on standard, one-snapshot behavioral or functional testing methods, such as catwalk for gait analysis, grip strength tests, open field, and plus maze assessments. These traditional tests required both the scientist and the animals to undergo specific training. Typically conducted during daylight, when rodents are less active, these methods often resulted in short observation periods, low data collection, and potential errors due to manual recording by scientists. Moreover, the lack of standardization across different institutions presented additional challenges.

With the introduction of AI and Machine Learning, a trend towards "do-it-yourself solutions" emerged in many laboratories. These solutions involve building custom test arenas, often equipped with food and bedding, where animals are observed over extended periods, such as 2-3 hours. Behaviors and activities are automatically recorded using either self-developed algorithms or open-source solutions like Deeplabcut. While these approaches allow for more longitudinal data collection, they still necessitate animal training, isolation, and handling during their inactive daylight hours. Moreover, achieving standardization and scalability with these solutions remains a challenge, as they are typically developed and validated within a single laboratory for specific animal models.

In response to these limitations, temporary on-bench solutions like Phenomaster, Phenotyper, Intellicage, and others were developed. These commercial systems enable animals to be monitored for extended periods (over 24 hours) as they provide essentials such as food and water in an environment similar to standard open or IVC cages. The advantages of these systems include the ability to automatically collect more standardized data with reduced disturbance, and their scalability and consistency compared to the more variable do-it-yourself solutions. However, these systems are not without limitations. For most, the animal still requires isolation and training, and, depending on IACUC and internal regulations, can only be housed in these systems for a limited duration.

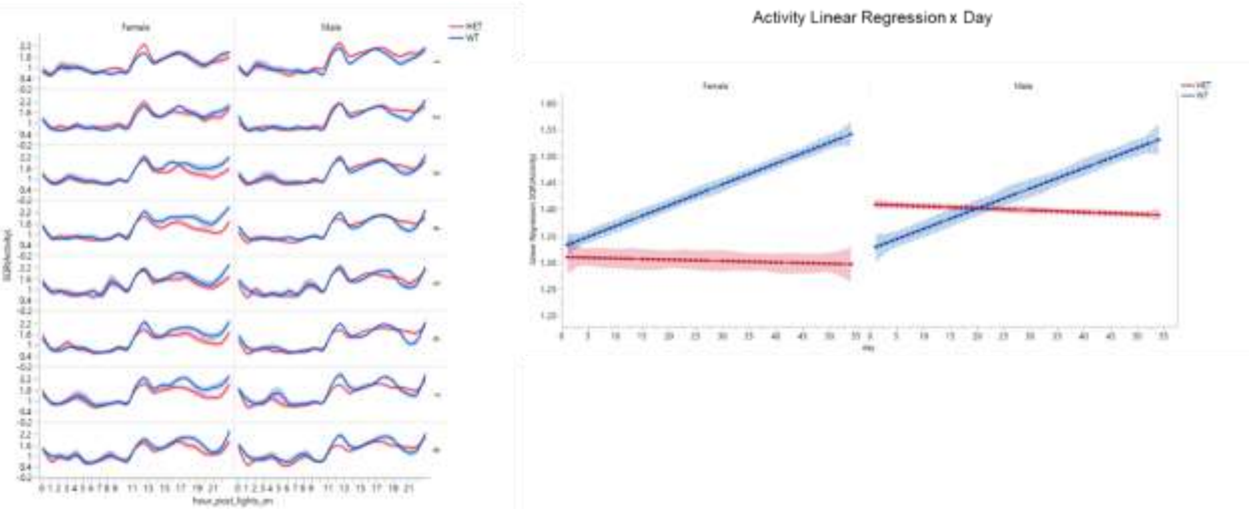
The focus then shifts to the latest advancement in this field: digital home-cage monitoring solutions. These solutions utilize standard IVC cages where animals can be socially housed, with enrichment, food, and water provided, enabling monitoring throughout their entire lifespan, if desired. A prime example of this technology is the DVC Tecniplast system, available since 2017 at Novartis. This system is seen as the most comprehensive solution for operational, welfare, and scientific purposes. It is fully scalable within a facility but requires a robust IT infrastructure. However, currently, this system is only available for mice.

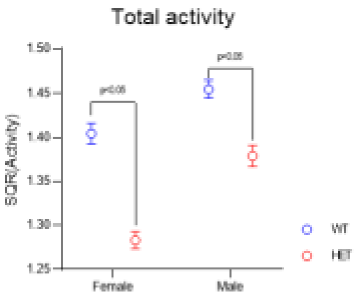
The DVC system uses a standard IVC cage, available in an 80-slot Green Line version. Each cage is equipped with an electromagnetic field board containing 12 electrodes, activated depending on the position of the animals in the cage. The system's capabilities include tracking activity per cage and providing detailed metrics for a digital running wheel, such as the number of bouts, distance, and speed. It also features several welfare alarms for detecting anomalies in activity, fighting, cage conditions (such as flooding or low food), and automated bedding moisture detection, aiding in determining optimal times for cage changes. Integration with personal animal management software is also possible, allowing for automated cage census and tracking.

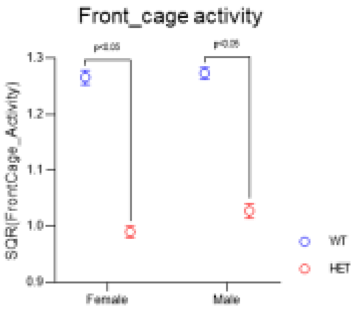
To illustrate the practical application of this system, consider an experiment conducted in Cambridge, USA, focused on preclinical modeling of a neurodevelopmental disorder. This disorder, in its human manifestation, is known to exhibit symptoms such as developmental retardation, maladaptive and impulsive behavior, sleep disturbances, and abnormal EEG patterns. Initially, researchers considered using the Phenomaster system for this study. However, this would have required isolating the animals and limiting their time in the cage to two weeks, whereas an eight-week monitoring period was necessary to fully characterize the model. Consequently, the DVC caging system was chosen for this study.

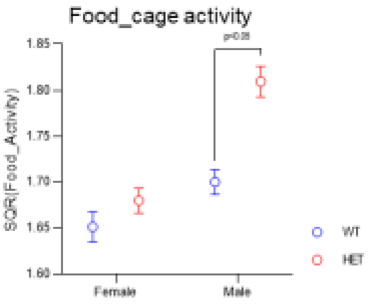
The study included both wild-type and heterozygous mice with microdeletions, encompassing both sexes, to monitor general activity, spatial activity distribution in the cage, and the rest disturbance index, a validated indicator of the animals' rest disturbances during daylight. Additionally, two behavioral tests were performed at the beginning, middle, and end of the experiment. Body weight and food intake were monitored weekly, and clinical scoring was conducted three times per week.

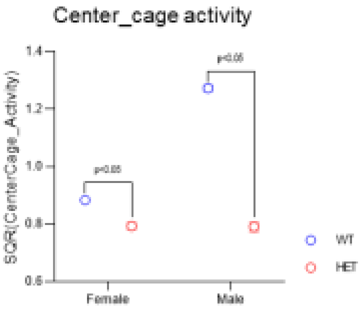
Analyzing the data, a plot (left side) was generated representing the circadian activity evolution per week for both heterozygous (HET) and wild-type (WT) mice, differentiated by sex (females on the left and males on the right). The plot's horizontal axis represents time, starting from when lights are turned on at 6 AM and continuing until 5 AM the next day. The vertical axis shows the average activity levels. Notably, the activity of the HET mice, marked in red, decreases over the weeks during the dark phase. For male mice, this reduction in activity becomes apparent from the fourth week, particularly noticeable in the activity peak occurring when the lights turn off. On the right side of the plot, a linear regression of daily activity is presented. This analysis reveals a flattening and slight reduction over time in the activity of both HET genotypes. In contrast, WT mice of both sexes exhibit an increase in activity levels. The differences in male mice become visible from week 4, indicating that the activity difference manifests later in time. This finding highlights the importance of considering both sexes in such studies.

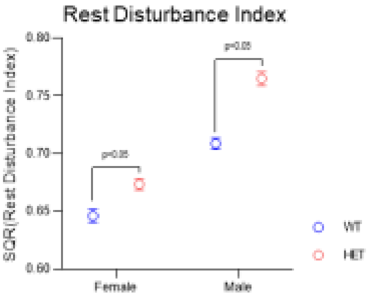












The results also provide insights into the use of the cage. WT mice were more active at the front and frequently crossed the center of the cage. An interesting observation was made regarding female HET mice, who spent more time under the food grid. Although no difference in food intake was noted in this model, the mice developed a malformation of the muzzle, primarily in females. It is speculated that the HET mice spent more time under the grid to eat the pellets. The Rest Disturbance Index was higher in HET mice compared to WT mice, as expected, showing a disturbance of the rest phase during daylight.

In conclusion, this section highlights the significant role of digital home-cage monitoring solutions in advancing pharmaceutical research. By providing detailed, continuous data on rodent behavior and physiology, these systems enhance our understanding of disease mechanisms and treatment efficacy. The DVC Tecniplast system demonstrates how technology can improve operational efficiency, animal welfare, and scientific output. Its ability to integrate with animal management software and provide comprehensive data across a range of metrics makes it an invaluable tool in the pharmaceutical industry. This technology not only improves the quality of data collected but also accelerates the process of translating preclinical findings into clinical insights, thus expediting the development of new therapies and treatments.

Urination in the Home Cage: Development of a Digital Biomarker for Sample-Free Diagnosis of Diabetes in Mice

Blood glucose is arguably one of the most common and most important in vivo parameters in metabolic research. Given the pivotal role of dysregulated glucose homeostasis in metabolic disease, particularly in diabetes, numerous pharmacological and genetic animal models are available which exhibit elevated blood glucose (known as hyperglycemia) and diabetes. However, in most models there is a large individual variation in the incidence, severity, and time course of the development of the hyperglycemic phenotype. This necessitates frequent measurements of blood glucose to monitor disease progression, which typically involves repeatedly cutting or puncturing the tail vein of the mice for collection of blood samples.

One of the hallmark features of sustained hyperglycemia in diabetes is increased urination, known as polyuria. When blood glucose concentrations are normal, the glucose entering the urine during renal filtration is effectively reabsorbed via epithelial transport as the urine passes through the proximal tubules of the kidney. However, in conditions of sustained severe hyperglycemia, such as in uncontrolled diabetes, the blood glucose concentration reaches a level where the capacity for renal glucose reabsorption becomes saturated. The concomitant increase in the urinary glucose concentration causes osmotic diuresis, whereby additional water is drawn into the urine. The resulting excessive fluid loss presents as polyuria, and is accompanied by a compensatory increase in thirst, known as polydipsia. Together with an increase in appetite (called polyphagia) these symptoms are known as the 3P's of diabetes, and they represent the most common symptoms of undiagnosed diabetes in humans.

In mouse models of diabetes, severe polyuria is also a common feature, and in cages with diabetic mice it is often necessary to change the bedding several times per week. Consequently, we hypothesized that by monitoring and analyzing the temporal changes in bedding moisture in mouse cages, we could identify the appearance of diabetes symptoms as an indicator of the onset of disease.

Monitoring of bedding moisture in the home cage is an integral part of the functionality of the Tecniplast DVC system, where the Bedding Status Index (BSI) is generated from such measurements. The BSI can be used to guide the frequency of cage changes and to warn about flooding events from leaky water bottles, but since it is recorded continuously, we investigated whether it can also be used to evaluate the average rate of urination by the mice in the cage. Consequently, we tracked the changes in BSI along with the development in blood glucose and the consumption of food and water in mouse models of both obesity/prediabetes, type 1 diabetes, and type 2 diabetes, and we found a robust relationship between increased changes in BSI and sustained hyperglycemia, which allows for accurate identification of the onset of disease.

Thus, by leveraging the continuous monitoring of BSI by the DVC system, we have developed a digital biomarker, that can be used as an alternative to repeated blood glucose measurements to reliably and accurately detect the appearance of diabetes symptoms based on the increase in bedding moisture. We believe that this novel sample-free and non-invasive approach to identify the onset of diabetes in mice while they are ambulatory in their home cage represents a significant refinement and improvement of animal welfare as it holds the potential to drastically reduce the requirement for manual monitoring of blood glucose in diabetes research.

Ethical Statement

All animal experiments were approved by the local authorities (Switzerland, France, Germany, Denmark) and where applicable in line with the EU-directive 2010/63/EU. The nerve injury experiments were approved by the Weizmann Institute of Science institutional animal care and use committee.

Symposium: Considerations in behavioural phenotyping of genetic mouse models of Alzheimer's disease and frontotemporal dementia

Behavioural, cognitive and sensory phenotyping of knock-in mouse models of Alzheimer's disease and frontotemporal dementia

S. T. Boyanova [1,2], G. Banks [1,2], R. S. Bains [2], M. Stewart [2], S. Wells [2], and F. Wiseman [1,2]

1 UK Dementia Research Institute, University College London, London, UK. s.boyanova@ucl.ac.uk

2 The Mary Lyon Centre at MRC Harwell, Oxfordshire, UK.

Introduction

Robust phenotyping of animal models of dementia-causing diseases is critical to understanding the cellular and molecular mechanisms that cause disease [1]. It is important both that the underlying pathological mechanisms of the model recapitulate clinical processes and that the consequences of this on sensory, motor, and cognitive functions are understood to maximise the translational value of research findings. Here we discuss the application of these principles to physiological mouse models of amyloid-β accumulation in Alzheimer's disease (AD), and amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD).

AD is characterised by memory deficits, a decline in other cognitive functions, and is often associated with behavioural and psychological symptoms (BPSD). The pathology of the disease includes accumulation of extracellular amyloid-β (Aβ) plaques, intracellular, and misfolded and hyperphosphorylated neurofibrillary tau tangles, which are associated with a neuroimmune response. This is then followed by neuronal loss mainly in the cerebral cortex and the hippocampus as the disease progresses [1].

AD is classified as early-onset (EOAD), occurring before the age of 65 years, or late-onset (LOAD) with onset after the age of 65 years of age, both forms of the disease share common underlying neuropathology and clinical features. However, EOAD is less common and includes autosomal dominant cases caused by mutations in the presenilin 1 (*PSEN1*), presenilin 2 (*PSEN2*), and amyloid precursor protein (*APP*) genes. These mutations affect APP processing and result in an altered production of the Aβ peptide [2]. How, the accumulation of Aβ leads to the cascade of neuropathological changes and the development of AD clinical features is not well understood.

Another common cause of early-onset dementia is ALS/FTD, these two diseases have overlapping clinical, pathological, and genetic origins, and are considered to have common underlying mechanisms. FTD is characterised by the development of BPSD symptoms, including impulsivity, personality changes, apathy, and repetitive behaviours. A GGGGCC repeat expansion in the first intron of the *C9ORF72* gene is the most common genetic cause of ALS and FTD [3,4]. One potential disease mechanism and a major pathogenic feature of ALS/FTD, is the expression of aberrant dipeptide repeat (DPR) proteins generated from the hexanucleotide repeat by the process of Repeat Associated Non-AUG (RAN) translation [5]. In addition, nearly all cases of ALS, and approximately half of FTD and AD cases, develop cytoplasmic mislocalised and aggregates of the TAR DNA-binding protein (TDP-43) [6]. Over 50 *TARDBP* disease-associated mutations have been identified in ALS/FTD. What contribution C9ORF72-DPR and TDP-43 neuropathology make to the development of FTD/AD clinical features is not well understood.

The known EOAD and FTD causal mutations have been modelled in transgenic mice that overexpress the disease associated proteins. These models have been useful to understand aspects of disease and they played a key role in the early development of amyloid-β immunotherapy. However, they exhibit some artefactual phenotypes because of the non-physiological highly elevated level of expression, and temporal and spatial misexpression of the disease associated proteins which is not representative of clinical disease. Therefore, a new generation of knock-in gene-edited mouse models have been created, in which the endogenous mouse loci have been altered to better recapitulate underlying disease biology [5,7,8]. These mice express disease-associated protein at wild type levels while producing elevated disease associated forms of these proteins. For example, the *App* gene has been partially humanised such that the mouse generates the human version of AD-associated Aβ alongside AD causal mutations – such as the Swedish (KM670/671NL) and Beyreuther/Iberian (I716F) mutations in the *App*$^{NL\text{-}F/NL\text{-}F}$ mice, or three mutations – Swedish, Beyreuther/Iberian, and the Arctic mutation (E693G) in the *App*$^{NL\text{-}G\text{-}F/NL\text{-}G\text{-}F}$ mice which promotes the peptides aggregation [7]. Humanisation of the Aβ region leads to changes in APP/Aβ biology in the rodent brain, independently of the effect of AD causal mutations, thus the use of a control line with only humanisation such as the (*App*$^{em1bdes/em1bdes}$) model is valuable to understand the cause of phenotypic changes [9]. Similarly, knock-in mouse models of TDP-43 and C9orf72 DPR neuropathology have been recently developed, including the *Tardbp*$^{Q331K/Q331K}$ [8] and the *C9orf72*$^{GR400/+}$ [5]. DPR mouse models.

Here we use a side-by-side longitudinal cognitive and sensory behavioural study to characterise *App*$^{em1bdes/em1bdes}$, *App*$^{NL-F/NL-F}$, *App*$^{NL-G-F/NL-G-F}$, *C9orf72*$^{GR400/+}$ and *Tardbp*$^{Q331K/Q331K}$ mouse models, with a focus on understanding how these models can be used to understand BPSD features of disease.

Methods

The mice underwent a standardised pipeline of longitudinal behavioural testing. Separate cohorts of the *App* knock-in series are also undergoing touch screen cognitive testing to be completed in 2025. All protocols used in this study were done in accordance with the Animals (Scientific Procedures) Act 1986 (UK) Amendment Regulations 2012 (SI 4 2012/3039) at the Mary Lyon Centre at MRC Harwell.

Results

We will present intermediate results from home cage analysis of behaviour, the three-chamber test of social motivation, and the Sanderson forced Y-maze test of short-term memory. Moreover, we tested olfaction and visual function (optokinetic drum test) in the mice to understand if sensory function confounds affect test performance. We observed short-term memory deficits in the *App*$^{NL-G-F/NL-G-F}$ mice compared to wild type in the Y-maze from 33 weeks of age. In addition, we found a significant effect in the *Tardbp*$^{Q331K/Q331K}$ model in the three-chamber test of social recognition and a significant age-related genotype effect in the *C9orf72*$^{GR400/+}$ animals in the Y-maze.

Conclusion

These data can be used to select the most useful model and time-point to inform fundamental research experimental design to understand the cellular and molecular processes that contribute to aspects of AD/FTD clinical disease, and the design of proof-of-concept intervention preclinical experiments to facilitate the development of new therapies for these devastating diseases.

References

1. Sasaguri H, Nilsson P, Hashimoto S, Nagata K, Saito T, De Strooper B, et al. APP mouse models for Alzheimer's disease preclinical studies. EMBO J. 2017;36:2473–87.
2. Cannavo C, Tosh J, Fisher EMC, Wiseman FK. Using mouse models to understand Alzheimer's disease mechanisms in the context of trisomy of chromosome 21. Prog Brain Res. 2020;251:181–208.
3. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011;72:257–68.
4. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011;72:245–56.
5. Milioto C, Carcolé M, Giblin A, Coneys R, Attrebi O, Ahmed M, et al. PolyGR and polyPR knock-in mice reveal a conserved neuroprotective extracellular matrix signature in &lt;em&gt;C9orf72&lt;/em&gt; ALS/FTD neurons. bioRxiv [Internet]. 2023;2023.07.17.549331. Available from: http://biorxiv.org/content/early/2023/07/17/2023.07.17.549331.abstract
6. Meneses A, Koga S, O'Leary J, Dickson DW, Bu G, Zhao N. TDP-43 Pathology in Alzheimer's Disease. Mol Neurodegener. 2021;16:84.
7. Saito T, Matsuba Y, Mihira N, Takano J, Nilsson P, Itohara S, et al. Single App knock-in mouse models of Alzheimer's disease. Nat Neurosci. 2014;17:661–3.

8. White MA, Kim E, Duffy A, Adalbert R, Phillips BU, Peters OM, et al. TDP-43 gains function due to perturbed autoregulation in a Tardbp knock-in mouse model of ALS-FTD. Nat Neurosci. 2018;21:552–63.
9. Serneels L, T'Syen D, Perez-Benito L, Theys T, Holt MG, De Strooper B. Modeling the β-secretase cleavage site and humanizing amyloid-beta precursor protein in rat and mouse to study Alzheimer's disease. Mol Neurodegener. 2020;15:60.

A simple task with mixed findings: factors to consider in preclinical behavioural research

Szu-Han Wang

Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom. s.wang@ed.ac.uk

Introduction

Dry-land mazes are commonly used for behavioural phenotyping and/or for characterising learning and memory functions in rodents or in rodent genetic models of diseases. One of the dry-land maze tasks that is often used to assess working memory is spontaneous alternation in a Y-maze and has been used in phenotyping of genetically modified mouse models of Alzheimer's disease [1]. It is considered a simple task from three perspectives: First, the apparatus is in a simple design that is composed of three equilateral straight alleys with an equilateral centre zone and is relatively easy to build. Second, the procedure for assessing behaviour is simple by placing the animal in the Y-maze and let the animal freely explore for a short period of time. Finally, the measurement of the behaviour is simple by counting the sequence of alley visiting.

The simplicity of the Y-maze task prompts the popularity of its use in phenotyping of genetic modified mice of disease models. For example, a knock-in mouse model carrying Swedish, Beyreuther/Iberian, and Arctic mutations in the genes of amyloid precursor protein (APP) is developed to capture amyloidosis in Alzheimer’s disease without the overexpression of APP [2]. Reduction in the percentage of spontaneous alternation in the Y-maze task has been reported in this APP NL-G-F mouse model at the age of 6 months old. This reduction, while significant compared to the wildtype group, remains highly significantly above chance [2].

Following the initial finding, later research that uses the same line of mice and the wildtype littermate control does not support an impairment in the Y-maze task. For example, we recently report that when the APP NL-G-F mice are previously trained with another extensive, cognitive task, no impairment is seen [3]. Another study that uses untrained mice in an attempt to replicate the initial report also does not show an impairment in the APP NL-G-F mice [4]. Facing mixed results, it is pertinent to identify key factors that contribute to the discrepancy. This paper aims to provide both quantitative and qualitative analyses of studies that reported the presence or absence of impairment in spontaneous alternation in the Y-maze in the APP NL-G-F mice. These analytical results would provide insights on factors to consider in comparison of preclinical findings and factors to control for in future studies.

Methods

Literature searches were performed in both Web of Science and PubMed databases. The search keywords included APP knock-in (KI) or APP NL-G-F, mice, and Y-maze. This analysis study did not involve ethical committee approval, while ethical statements for animal research could be found in cited publications. Abstracts of searched results were further screened for relevance. The inclusion criteria were: original studies, using APP NL-G-F mice, using the Y-maze task, and published in English. Full texts of relevant papers were obtained and data figures containing results of the Y-maze task were identified. The percentages of spontaneous alternations (SA%), animal, numbers, and the standard error of means or standard deviation were extracted from published figures or graphs. The SA% was calculated based on: Numbers of sequential alternation of 3 arms/ (Total numbers of arm visits - 2) * 100. The chance level from random alternation would be 22.22%. The group means (M), sizes (n), and standard deviations (s) were used to determine the effect sizes (Hedges’ g [5]) of genotype difference in the Y-maze task.

Hedges’ g = $(M_{KI} - M_{WT}) / S_{pooled}$

S_{pooled} = square root of $\{[(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2] / (n_1 + n_2 - 2)\}$

Factors that may contribute to the inter-study difference were extracted from each paper. These included: age of the mice, sex of the mice, source of wildtype mice, task length, handling numbers prior to the Y-maze task, behavioural training prior to the Y-maze task, and the performance in the wildtype mice. The influence of categorical or descriptive factors (e.g. sex of the mice, source of wildtype mice, and behavioural training prior to the Y-maze task) on whether an impairment in APP-KI mice was observed would be compared across papers. Quantitative analyses were performed to determine the correlation between each quantitative factor and the effect size (e.g. age of the mice, task length, handling numbers prior to the Y-maze task, and the performance in the wildtype mice).

Results

After applying the inclusion criteria described above, 6 papers were identified from 2014 to 2023. They were compared based on 3 categorical or descriptive factors. First, sex of mice: Two papers used both female and male mice and one reported no impairment of SA% in APP-KI mice [3], while the other one reported an impairment that was seen in both female and male mice [6]. Two papers that used male mice also showed mixed results [2,4]. An additional paper that used male mice did not include wildtype control [7], while one paper that used female mice showed an impairment of SA% in APP-KI mice [8]. Second, source of wildtype mice: Two papers used wildtype littermates and showed mixed results [2,3]; same with 3 other papers that did not clarify or did not use wildtype littermates [4,6,8]. Third, behavioural training prior to the Y-maze task: Two papers used experimentally-naïve mice and showed mixed results [2,4]; same with 3 other papers in which mice received prior behavioural procedures such as an appetitive delayed matching-to-place task [3], an open-filed and a novel object recognition task [6], or exposure to an elevated plus maze [8].

Effective sizes of genotype difference in the Y-maze task were calculated from 5 papers that included data from the wildtype control [2-4,6,8]. As shown in Figure 1, 3 papers supported an impairment of SA% in APP-KI mice [2,6,8], while the other 2 did not [3,4]. All these papers used the same task length which was 5-min, so this factor did not contribute to the mixed results.

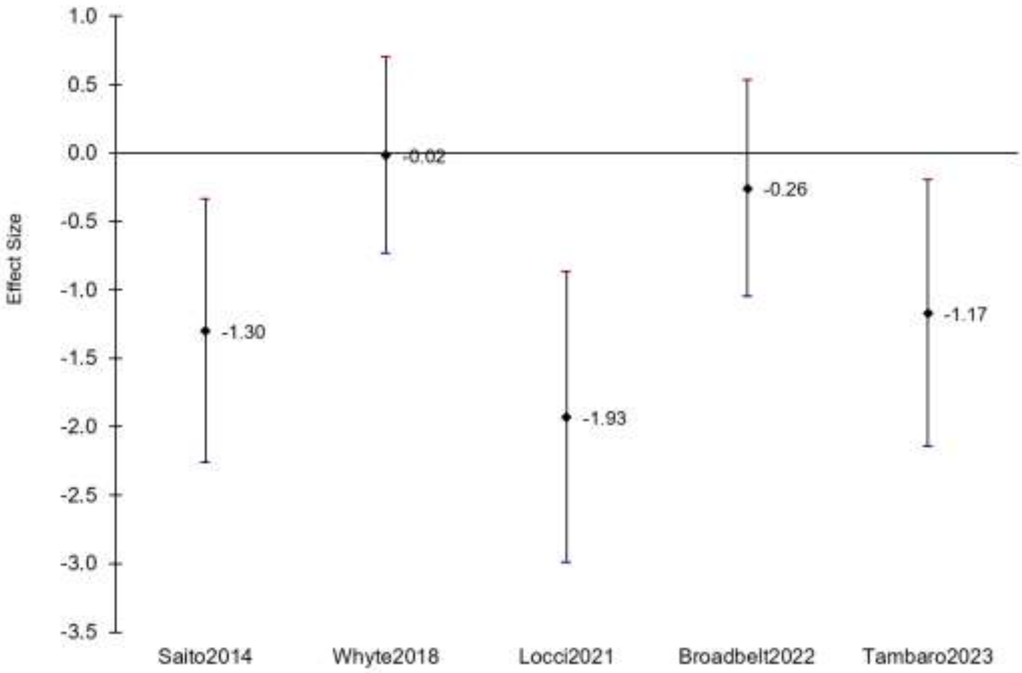


Figure 1. Effective sizes of genotype difference in the Y-maze task from 5 papers [2-4,6,7] in a chronological order. The first author and published year are indicted at the lower part of the figure. Negative effect sizes indicate an impairment in spontaneous alternation in the APP NL-G-F mice. Both Hedges' g values (number and diamond) and confidence intervals (bars) are presented.

As data from 2 age groups [2] or 2 sexes were provided [3] in 2 of the papers, correlational analyses could be performed across 7 studies in 5 papers [2-4,6,8]. The correlation between the age of the mice, ranging from 26 to 52 weeks, and the effect size was weak and insignificant (r = -0.13, p > 0.4). Due to a lack of reporting of handling in one paper [8], the correlation between handling numbers and the effect size was perform in 6 studies of 4 papers [2-4,6]. The result showed a positive correlation (r = 0.66), which was not significant (p = 0.078) with the small number of studies so far. See Figure 2. This analysis across 6 studies would suggest more handing being associated with less impairment in the APP NL-G-F group in spontaneous alternation in the Y-maze.

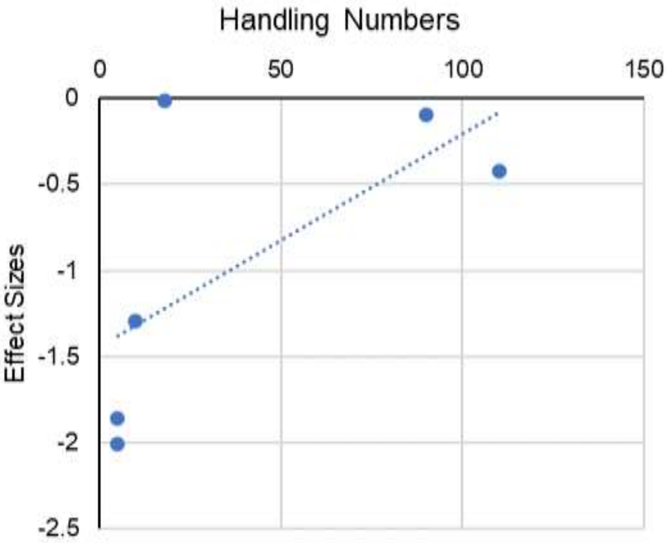


Figure 2. Correlation between the number of handing and the effect sizes of genotype difference in the Y-maze task from 6 studies in 4 papers [2-4,6]. More handling is associated with less impairment in spontaneous alternation in the APP NL-G-F mice.

Finally, the performance in the wildtype group could be identified from all 7 studies in 5 papers [2-4,6,8]. The correlation between the performance in the wildtype mice and the effect size (r = -0.69) was significant (p = 0.04, see Figure 3). This analysis across studies would suggest a higher performance in the wildtype group being associated with a stronger impairment in the APP NL-G-F group in spontaneous alternation in the Y-maze.

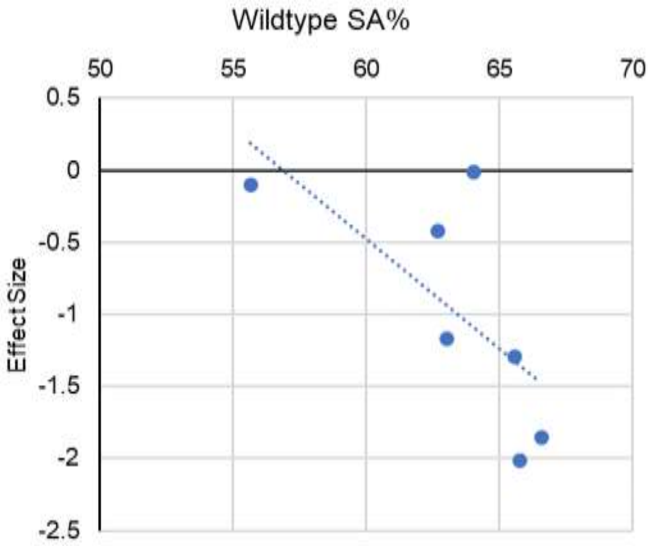


Figure 3. Correlation between the wildtype performance and the effect sizes of genotype difference in the Y-maze task from 7 studies in 5 papers [2-4,6,8]. Performance in the wildtype mice is indicated by the percentage of spontaneous alternation (SA%). Negative effect sizes indicate an impairment in spontaneous alternation in the APP NL-G-F mice.

Discussion

Factors to consider. Focusing on a simple task such as spontaneous alternation in the Y-maze task in a relatively new APP NL-G-F mouse model of Alzheimer's disease, this paper compares across a selective range of factors and identifies ones that critically contribute to mixed results that show either no impairment (effect sizes near zero), or a significant impairment (large negative effect sizes) in working memory (see Figure 1). Analyses here reveal 2 factors that most likely enable dissociate of the findings across studies.

The first factor is 'the number of handling' prior to the Y-maze task. Our recent research uses cognitively pretrained mice and reports no impairment in APP NL-G-F mice in the Y-maze task, while there is an impairment in a cognitive process called behavioural tagging [3]. In addition to characterising the phenotype of APP NL-G-F mice in a wide range of behavioural tasks such as spatial learning and memory in an appetitive delayed matching-to-place task and in a Barnes maze, novel object recognition, and objection location recognition, this study also aims to reduce the translational gap between rodent research that involves limited to no explicit prior cognitive training and human research where participants are already enriched with cognitive experiences. It is conceivable that previous cognitive training delays cognitive impairment in APP NL-G-F [3,9] or in other AD mice [10].

Cognitive training in rodents usually involves more handing in standard behavioural tasks. Through analyses across studies, the current finding reveals that the number of handling prior to the Y-maze task may play an important role in determining if an impairment is seen. For example, 2 studies that show no impairment had handled the mice 18 times [4] or substantially more times before the Y-maze task [3]. On the contrary, 2 other studies, which show large APP NL-G-F impairment, had handled the mice 5-10 times prior to the task [2,6]. While the moderate correlation is not yet statistically significant due to a small number of studies (Figure 2), including future research could reveal and substantiate this relationship. It is vital to prevent potential confounding from anxiety phenotypes or subtle motivation difference in APP NL-G-F mice, compared to the wildtype animals. Hence, it is recommended that sufficient handling should be performed before behavioural tasks.

The second factor is 'performance in the wildtype group'. According to 3 studies that show the largest negative effect sizes, wildtype mice would spontaneously alternate at > 65% [2,5]. When falling below 65% in the wildtype group, mixed results are apparent [2,4,6]. See Figure 3. This is unlikely due to a 'floor effect' as 3 studies with mixed results show similar wildtype performance at a similar range of 62-64%. APP NL-G-F performance in some of the studies [2,6,8] can also go lower than one study that show 56% in the wildtype group [3]. It could be that when the control group performs at a higher level, subtle changes in motivation or attention [3,6,9,11] in APP NL-G-F may contribute to reduction in spontaneous alternation. Although this factor cannot be easily controlled *a priori* and is only revealed after the wildtype mice have undergone this task, it is a factor that needs to be considered in drawing comparison and conclusion across studies.

Interpretations and future work. One limitation of current work is that the study size is fairly small and not all factors being investigated are reported in all published studies. This may contribute to some of the effects being only approaching significance. For example, the factor of handling is correlated with the effect size but the p value is only approaching significance. Further substantiation of this correlation can be achieved through obtaining performance from individual animals in published papers and by including future studies.

Despite the large negative effect sizes in some of the publication [2,6,8], the magnitude of impairment in spontaneous alternation in APP NL-G-F mice may not be as drastic as it seems. In all these 3 studies [2,6,8], the performance in the APP NL-G-F mice (43-51%) remain highly significant above chance (22%). Of note, the tendency of spontaneously 'stay' or 'alternate' in rodents may be task dependent [12].

The current approach involves descriptive analysis of 3 factors and quantitative analyses of 4 factors, although 1 of which is consistent across papers. These are treated as independent factors although it could be that interactions between factors would affect the effect size of the phenotype. This approach can be expanded to include testing of the interaction effects or modelling of multiple factors. Other factors that involve descriptive comparison across 2 of earlier studies have been reported [4]. Future work can include additional factors for analyses, provided that they are all measured and reported in publication with proper wildtype control. Examples of these factors include the measurement of total arm visits, the level of locomotion, and the level of anxiety or stress.

Spontaneous alternation in the Y-maze in rodent research take advantage of rodents' tendency in exploring less recently visited arms. It has been suggested as an indicator of working memory, but should not be assumed as a measurement of memory alone [13]. The concept of working memory is traditionally investigated in humans and monkeys using different models. To what degree the spontaneous alternation in mice reflect the same working memory concept that requires holding object or item information in active minds in humans needs clarification.

Application and conclusions. In summary, handling is a key factor in interpreting mixed phenotyping results. It is recommended that sufficient handling should be conducted so the finding is less confounded by stress, anxiety, or unfamiliarity. Performance in the wildtype animals should be taken into consideration when comparing findings across studies. Our approach of combining quantitative and qualitative analyses can be applied to other behavioural models, disease models, or research questions to supplement review articles that take descriptive approaches. These in-depth analyses can provide new insights on informing better practice in future preclinical research.

References

1. Kobayashi, K., Chen, D. (2005). Behavioral phenotypes of amyloid-based genetically modified mouse models of Alzheimer's disease. *Genes, Brain and Behavior*, 4, 173-196.
2. Saito, T., Matsuba, Y., Mihira, N., Takano, J., Nilsson, P., Itohara, S., Iwata, N., Saido, T. C. (2014). Single app knock-in mouse models of alzheimer's disease. *Nature Neuroscience*, 17(5), 661-664.
3. Broadbelt, T., Mutlu-Smith, M., Carnicero-Senabre, D., Saido, T. C., Saito, T., & Wang, S-H. (2022). Impairment in novelty-promoted memory via behavioral tagging and capture before apparent memory loss in a knock-in model of alzheimer's disease. *Scientific Reports*, 12, 22298.
4. Whyte, L. S., Hemsley, K. M., Lau, A. A., Hassiotis, S., Saito, T., Saido, T. C., Hopwood, J.J., Sargeant, T. J. (2018). Reduction in open field activity in the absence of memory deficits in the APP nl-g-f knock-in mouse model of alzheimer's disease. *Behavioural Brain Research*, 336, 177-181.
5. Hedges, L. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6(2), p.107.
6. Locci, A., Orellana, H., Rodriguez, G., Gottliebson, M., McClarty, B., Dominguez, S., Keszycki, R., Dong, H. (2021). Comparison of memory, affective behavior, and neuropathology in APP NL-G-F knock-in mice to 5xFAD and APP/PS1 mice. *Behavioural Brain Research*, 404, 113192.
7. Ni, J., Wu, Z., Meng, J., Saito, T., Saido, T.C., Qing. H., Nakanishi, H. (2019). An impaired intrinsic microglial clock system induces neuroinflammatory alterations in the early stage of amyloid precursor protein knock-in mouse brain. *J Neuroinflammation*, 16(1),173.
8. Tambaro, S., Mitra, S., Gera, R., Linderoth, B., Wahlberg, L.U., Darreh-Shori, T., Behbahani, H., Nilsson, P., Eriksdotter, M. (2023). Feasibility and therapeutical potential of local intracerebral encapsulated cell biodelivery of BDNF to AppNL-G-F knock-in Alzheimer mice. *Alzheimer's Research & Therapy*, 15(1), 137.
9. Mehla, J., Deibel, S.H., Karem, H., Hong, N.S., Hossain, S.R., Lacoursiere, S.G., Sutherland, R.J., Mohajerani, M.H., McDonald, R.J. (2023). Repeated multi-domain cognitive training prevents cognitive decline, anxiety and amyloid pathology found in a mouse model of Alzheimer disease. *Communications Biology*, 6(1), 1145.
10. Williams, E., Mutlu-Smith, M., Alex, A., Chin, X.W., Spires-Jones, T., Wang, S-H. (2023). Mid-Adulthood cognitive training improves performance in a spatial task but does not ameliorate hippocampal pathology in a mouse model of Alzheimer's disease. *Journal of Alzheimer's Disease*, 93(2), 683-704.

11. Latif-Hernandez, A., Shah, D., Craessaerts, K., Saido, T., Saito, T., De Strooper, B., Van der Linden, A., D'Hooge, R. (2019). Subtle behavioral changes and increased prefrontal-hippocampal network synchronicity in APPNL-G-F mice before prominent plaque deposition. . *Behavioural Brain Research*, 364, 431-441.
12. Salvetti, B., Morris, R.G.M., Wang, S-H., (2014). The role of rewarding and novel events in facilitating memory persistence in a separate spatial memory task. *Learning and Memory*, 21, 61-72.
13. Hughes, R. (2004). The value of spontaneous alternation behavior (SAB) as a test of retention in pharmacological investigations of memory. *Neuroscience & Biobehavioral Reviews*, 28(5), 497-505.

Behavioural characterisation of humanised APP knock-in mice

Loukia Katsouri [1], Stephen Burton [1,2], Angela Mišak [1], Jade Sangha [1], Cristina Mazuski [1], John O'Keefe [1,2]

1.Sainsbury Wellcome Centre, London, UK.

2.Cell and Developmental Biology Department, University College London, London, UK.

l.katsouri@ucl.ac.uk, s.burton@ucl.ac.uk, angela.miak.20@ucl.ac.uk, jade.sangha.18@ucl.ac.uk, c.mazuski@ucl.ac.uk, j.okeefe@ucl.ac.uk

Introduction

Transgenic models of Alzheimer's disease (AD) have long played a pivotal role in advancing our understanding of the disease [1]. Recently, humanized knock-in mouse models that express a human gene under the mouse gene promoter and regulatory elements [2] have emerged as invaluable tools for further exploration. Behavioural experiments assessing cognitive function in AD mouse models serve as critical phenotypic readouts for treatment assessments and genetic studies. This study emphasizes the need for standardized experimental design to ensure reproducibility and translational relevance [3].

Materials and Methods

Study design

The study focuses on humanized APP knock-in mice of three different age groups: 6 months (young), 12 months (middle aged), and 21-22 month (old), with wild-type littermates serving as controls. The rodents' innate exploratory behaviour is used as a platform to study cognitive processes without externally imposed rules or reinforcement.

In a series of behavioural tests, including the Open Field, Elevated Plus maze, Y-maze, Novel Arm Y-maze and Light/Dark box, we evaluated their motor performance, anxiety levels, and memory performance, including spatial, recognition, and episodic memory.

Furthermore, I will address crucial considerations essential when designing and conducting experiments such as the Novel object recognition [4] ensuring equal object exploration and ensuring that the mice explore without a side bias in the arena. I will also present how using machine learning techniques such as DeepLabCut [5] can greatly enhance the analysis and reliability of the experiments and lead to better-evaluated results. This can also remove the scorer's bias and subjectivity.

Ethical statement

All animal experiments were carried out in accordance with British Home Office Regulations (UK Animals Scientific Procedures Act 1986). Study protocols were in accordance with the terms of the Project License, which was reviewed by the Animal Welfare and Ethical Review Board at University College London.

Statistical analysis

The data were analysed with GraphPad Prism version 10 and SPSS version 29 (IBM) by using a two-tailed Student’s t-test or Mann–Whitney test and two-way ANOVA followed by Bonferroni post hoc analysis and correlation analysis. Power analysis was performed by using InVivoStat v4.7. Differences were considered significant for P < 0.5.

Conclusions

Our findings underscore the significance of employing rodent exploratory behaviour paradigms for advancing cognitive function understanding in AD. Rigorous experimental design and analysis are crucial for furthering our knowledge of Alzheimer's disease. This work contributes to the ongoing effort to standardise behavioural experiments in AD mouse models, facilitating reproducibility and translational impact.

References

1. J. Götz, L. G. Bodea, and M. Goedert, “Rodent models for Alzheimer disease,” *Nat. Rev. Neurosci. 2018 1910*, vol. 19, no. 10, pp. 583–598, Sep. 2018.
2. T. Saito *et al.*, “Single App knock-in mouse models of Alzheimer’s disease,” *Nat. Neurosci.*, vol. 17, no. April, pp. 1–26, Apr. 2014.
3. C. M. Loss, F. F. Melleu, K. Domingues, C. Lino-de-Oliveira, and G. G. Viola, “Combining Animal Welfare With Experimental Rigor to Improve Reproducibility in Behavioral Neuroscience,” *Front. Behav. Neurosci.*, vol. 15, p. 763428, Nov. 2021.
4. M. Antunes and G. Biala, “The novel object recognition memory: Neurobiology, test procedure, and its modifications,” *Cogn. Process.*, vol. 13, no. 2, pp. 93–110, May 2012.
5. A. Mathis *et al.*, “DeepLabCut: markerless pose estimation of user-defined body parts with deep learning,” *Nat. Neurosci.*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018.

Cognitive and Behavioural Phenotyping of APP-KI mice in the home cage enclosures

Julija Krupic[1, 2,*], Hinze Ho[1,2], Nejc Kejzar[2], Marius Bauza[3,*]

1 UK Dementia Research Institute at UCL, University College London, Great Britain, London, WC1E 6BT, UK (current affiliation)

2 Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK.

3 Sainsbury Wellcome Centre, University College London, London, UK

***Corresponding author: j.krupic@ucl.ac.uk; m.bauza@ucl.ac.uk**

Abstract

Home cage monitoring systems provide a powerful tool for the automated characterization of multiple cognitive and locomotion behaviours. Here we describe how a novel home cage monitoring system was used to phenotype App-mouse model. The identification accuracy reached 80%. Exploratory behaviours and quiescence were some of the most sensitive features observed. We argue that using multidimensional behavioural features is essential for identifying the pathology stages in individual mice required to elucidate the underlying neural mechanisms.

Introduction

Home cage monitoring systems present a powerful tool for automated characterization of an animal's cognitive function and locomotion activities [1]. Compared to standard tests conducted outside the home cage environments they are more animal friendly due to the absence of handling by the experimenter, they are less labour-intensive and offer improved standardization and increased reproducibility. They allow the collection of larger multidimensional datasets, which leads to improved statistical power of experiments and reduced number of animals required to demonstrate statistical effects. Moreover, continuous long-term monitoring of undisturbed animals in their habitats also allows studying a richer repertoire of observed behaviours [2]. These considerations are particularly important when conducting research using animal models of disease due to a large number of independent variables that are typically tested, such as different mutations, genders, ages, and treatments. Here, we used a newly developed home cage monitoring system known as smart-Kage [3] to characterize the cognitive and locomotion behaviours in amyloid-beta *App*$^{\text{NL-G-F}}$ KI mice [4] and their age-matched controls. In total, we used 32 behavioural features to describe a mouse's phenotype and showed that using multi-dimensional feature space we were able to differentiate between individual mice with 80% accuracy which was comparable or higher to the accuracy obtained using analogous standard tests. We argue that more attention should be paid to the characterization of individual mouse phenotypes for more accurate mouse stratification corresponding to different disease stages.

Methods

The data was previously described in (Ho et al., 2023). A summary is presented below.

Mice

Experimental procedures and animal use were performed in accordance with UK Home Office regulations of the UK Animals (Scientific Procedures) Act 1986, following ethical review by the University of Cambridge Animal Welfare and Ethical Review Body (AWERB). All animal procedures were authorized under Personal and Project licences held by the authors.

Two groups of mice were used in the study: 1) *App*$^{\text{NL-G-F}}$ [4] mice and their age matched C57BL/6J controls. All mice used in this study were males. Three AppNL-G-F KI mice and three controls were 22–24 weeks old when first tested in the smart-Kages. Their identities were unknown at the time of experiment. We also included two additional *App*$^{\text{NL-G-F}}$ KI positive males aged 39 weeks whose identity was known. The mice were continuously tested for ~4.5 months. The second testing period commenced ~8 months later and lasted for ~1.5 months. All eight mice were tested on the standard tests before and after they were tested in the smart-Kages. All mice were kept on a 12:12 h light: dark cycle (with lights on at 9:00 a.m. and off at 9:00 p.m.) at a controlled temperature (21–23°C) and humidity (50–60%).

Cognitive and Behavioural tasks

Smart-Kages were used to characterise mouse cogntive and locomotion function as previously described in [3]. In short, the mice were individually housed for many months and tested on T-maze task, novel-object and object-in-place recognition tests, time of moving, running on the wheel and quiscent states. 32 different outcome measures were used which include maximum performance on the T-maze, duration of working memory etc (see [3] for details).

Results

The data was previously described in [3]. We further found that behaviours related to quiescence states and exploration were the most important (i.e. resulted in the largest absolute eigenvector magnitudes) features in helping to differentiate between different phenotypes on an individual mouse level.

Discussion

Previously, other homecage environments known as Intellicages were used to characterise the cognitive locomotion function in *App*$^{\text{NL-G-F}}$ mice [5]. The authors demonstrated that these mice showed impairments in place preference learning and place reversal learning as well as place avoidance learning tasks. Unfortunately, the results are reported as an average group effect and it is not possible to deduce the accuracy of predicting individual mice group identity. We show that in smart-Kages the accuracy in identifying an individual mouse's phenotype is ~80% which is comparable to using analogous standard memory tests. Accurate identification of cognitive and behavioural impairments on an individual mouse basis (rather than a group effect) is potentially important for determining the stage of cognitive decline. In other words, individual mice may show a different rate of impairment at the hippocampal-entorhinal circuit level despite coming from the same strain, age, gender or sharing other important characteristics. Identifying individual variations in cognitive decline may help to better capture the underlying changes in the brain behind these differences.

References

1. Grieco, F., Bernstein, B.J., Biemans, B., Bikovski, L., Burnett, C.J., Cushman, J.D., van Dam, E.A., Fry, S.A., Richmond, Hacham, B., Homberg, J.R., Kas, M.J.H., Kessels, H.W., Koopmans, B., Krashes, M.J., Krishnan, V., Logan, S., Loos, M., McCann, K.E., Parduzi, Q., Pick, C.G., Prevot, T.D., Riedel, G., Robinson, L., Sadighi, M., Smit, A.B., Sonntag, W., Roelofs, R.F., Tegelenbosch, R.A.J., Noldus, L.P.J.J., 2021. Measuring Behavior in the Home Cage: Study Design, Applications, Challenges, and Perspectives. Front Behav Neurosci 15, 735387. https://doi.org/10.3389/fnbeh.2021.735387
2. Dennis, E.J., Hady, A.E., Michaiel, A., Clemens, A., Tervo, D.R.G., Voigts, J., Datta, S.R., 2021. Systems Neuroscience of Natural Behaviors in Rodents. J. Neurosci. 41, 911–919. https://doi.org/10.1523/JNEUROSCI.1877-20.2020
3. Ho, H., Kejzar, N., Sasaguri, H., Saito, T., Saido, T.C., De Strooper, B., Bauza, M., Krupic, J., 2023. A fully automated home cage for long-term continuous phenotyping of mouse cognition and behavior. Cell Rep Methods 3, 100532. https://doi.org/10.1016/j.crmeth.2023.100532
4. Saito, T., Matsuba, Y., Mihira, N., Takano, J., Nilsson, P., Itohara, S., Iwata, N., Saido, T.C., 2014. Single App knock-in mouse models of Alzheimer's disease. Nat Neurosci 17, 661–663. https://doi.org/10.1038/nn.3697
5. Masuda, A., Kobayashi, Y., Kogo, N., Saito, T., Saido, T.C., Itohara, S., 2016. Cognitive deficits in single App knock-in mouse models. Neurobiology of Learning and Memory, MCCS 2016 135, 73–82. https://doi.org/10.1016/j.nlm.2016.07.001

Symposium: TEA-TIME presents: Enhancing Reproducibility and Welfare through Home Cage Systems

Listening Carefully, the Challenges of Recording Home Cage Ultrasonic Vocalizations

R. Sonia Bains[1], Hamish Forrest[1] and Sara E. Wells[1]

[1] Mary Lyon Centre at Medical Research Council at Harwell, Oxfordshire, United Kingdom. r.bains@har.mrc.ac.uk

Introduction

Mice are known to use ultrasonic vocalizations (USVs) to communicate throughout their lives, these USV calls may evolve and change in frequency and complexity with age, but they are predominantly emitted in social contexts. In early life, pups emit these calls in order to elicit a care giving response from their mothers. As they grow older, these calls evolve to create social bonds and hierarchies in groups of mice [1]. As they grow and reach sexual maturity, males, as well as females emit courtship calls in order to attract suitable mates, and they emit warning or aggressive calls to warn off intruders in their territory [2]. Mouse social interactions are complex and ever evolving, until recently these were studied in standard out-of-cage short term testing environments with a single predefined parameter in mind, such as social dominance tube test [3], or to investigate the social interest of mice by using three chamber test [4]. With the arrival of longitudinal testing and the ability to use machine learning to automatically analyze vast amounts of data, efforts have been made to investigate non-evoked behavior in the home cage setting as being much more relevant to the biology and therefore clinical relevance of the models under investigation [5].

The Challenge

Processing USV data can be extremely time-consuming as many USVs can be emitted in a short period (USVs can be as short as 3ms [6], making it possible for mice to produce hundreds of calls in a very short timeframe and each USV would require manual measurement in specialized software. Recently though, software has been developed that uses deep-learning algorithms to automatically identify and measure USVs within recordings (e.g. DeepSqueak [6] and MUPET [7]. Whilst these software developments have aided in identifying USVs, it remains difficult to determine the context in which mice produce particular USV types. To do this, it is necessary to identify associations between USVs and behavior.

To this end, we have been developing the AVERT (Acoustic Vocalisation Early Response Technologies) system. The system is capable of simultaneously recording vocalizations as audio files and behavior as video files in the home cage environment, enabling reliable and accurate associations to be drawn between the two. AVERT system is designed to record from within an individually ventilated cage (IVC) placed in on a standard IVC rack. Ultimately, we aim to use this technology to be able to a build tool for investigating voluntary social behaviours in group housed mice. The ability to generate associations between behavior and USVs constitutes a crucial first step towards this aim.

Methods

All animals used in the study were randomly weaned into groups of 3 blocked by genotype and sex within individually ventilated cages (IVCs). Each cage was cleaned 72 hours prior to recording in AVERT, data was recorded for 24 hours and analyzed in DeepSqueak using the cage as the experimental unit.

Results and Discussion

The data acquired is complex and requires careful curation and the development of a bespoke analysis pipeline. Preliminary results indicate differences in age, sex and strain for various USV parameters.

Key Words: Phenotyping, Ultrasonic Vocalisations, Home Cage Monitoring

Ethical statement

All protocols used in this study were done in accordance with the Animals (Scientific Procedures) Act 1986 (UK) Amendment Regulations 2012 (SI 4 2012/3039).

References

1. Yao, K., Bergamasco, M., Scattoni, M.L.,Vogel, A.P., (2023). A review of ultrasonic vocalizations in mice and how they relate to human speech. *J. Acoust. Soc. Am.* **154** (2): 650–660. https://doi.org/10.1121/10.0020544

2. Hammerschmidt, K., Radyushkin, K., Ehrenreich, H., Fischer, J., (2012). The Structure and Usage of Female and Male Mouse Ultrasonic Vocalizations Reveal only Minor Differences. *PLoS ONE* **7**(7): e41133. https://doi.org/10.1371/journal.pone.0041133

3. Lindzey, G., Manosevitz, M. & Winston, H. (1966). Social dominance in the mouse. *Psychon Sci* **5**, 451–452. https://doi.org/10.3758/BF03331044

4. Moy, S.S., Nadler, J.J., Perez, A., Barbaro, R.P., Johns, J.M., Magnuson, T.R., Piven, J., Crawley, J.N. (2004). Sociability and preference for social novelty in five inbred strains: an approach to assess autistic-like behavior in mice. *Genes Brain Behav*. **3**(5):287-302. doi: 10.1111/j.1601-1848.2004.00076.x. PMID: 15344922

5. Bains, R.S., Forrest, H., Sillito, R.R, Armstrong, J.D., Stewart, M., Nolan, P.M. and Wells, S.E. (2023) Longitudinal home-cage automated assessment of climbing behavior shows sexual dimorphism and aging-related decrease in C57BL/6J healthy mice and allows early detection of motor impairment in the N171-82Q mouse model of Huntington's disease. *Front. Behav. Neurosci*. **17**:1148172. doi: 10.3389/fnbeh.2023.1148172

6. Coffey, K.R., Marx, R.E. & Neumaier, J.F. (2019). DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacol.* **44**, 859–868 (2019). https://doi.org/10.1038/s41386-018-0303-6 .

7. Kouzoupis S, Neocleous A, Athanassakis I. (2019). Categorization of Mouse Ultrasonic Vocalizations Using Machine Learning Techniques. *Acoustics.* **1**(4):837-846. https://doi.org/10.3390/acoustics1040050

Clever testing of smart mice with IntelliCage protocols avoiding water restrictions

I. Amrein[1] and D.P. Wolfer[2]

1 Division Functional Neuroanatomy, Institute of Anatomy, University of Zürich, Zürich, Switzerland. Department of Health Sciences and Technology, ETH, Zürich, Switzerland. irmgard.amrein@uzh.ch

2 Department of Health Sciences and Technology, ETH, Zürich, Switzerland. Division Functional Neuroanatomy, Institute of Anatomy, University of Zürich, Zürich, Switzerland. dwolfer@ethz.ch

Abstract

Water access is the motivation for learning while group housed in the IntelliCage system. To avoid potential water restriction in slow learners, we developed protocols providing free water, while encouraging learning with sweetened water reward. Our findings indicate that animal welfare can be improved with such protocols, without compromising the assessment of activity and learning in easy tasks. Complex tasks requires adjustments to the free water access to make it less attractive.

Introduction

The IntelliCage (TSE) is one of the many automated home cage monitoring systems collecting data of freely moving mice over 24 hours and days. In the IntelliCage testing environment, four learning corners are fitted into a large cage. Mice remain group housed in this cage for the entire testing period and are monitored individually for circadian activity, locomotion, and water intake. More importantly, mice can be subjected individually to a flexible battery of cognitive learning tasks, exploring e.g. working memory, spatial and spatio-temporal memory, motivation, impulsivity, anxiety-avoidance learning, taste preference, visual discrimination, and attention. The primary motivational driver behind the animal's performance is thirst. In the IntelliCage, mice can only drink if they make a correct choice, that is visiting the correct corner out of four corners and, within the corner, poking at the correct door out of two doors. When the correct corner and door is selected, the door will open and give access to water. If the choice was wrong, the door will not open, and the mouse has to leave the corner and start again. Thus, slow learners or mice repetitively expressing wrong choice patterns may risk dehydration. Water deprivation for extended periods of time can reduce animal welfare, which runs contrary to the animal-friendly global environment of the IntelliCage system. We therefore aimed to develop and test learning tasks in the IntelliCage that avoid possible water restrictions yet retain the ability to differentiate between good and poor learners.

Study design

Three rounds of reward-learning protocols were tested. In each study, the correct choice in a series of tests was rewarded with access to sweetened water (0.5% saccharin), while the conditions for access to free water were modified between studies. All animal experiments were conducted under permit ZH041/18, #29918 of the Canton Zurich Veterinary Office.

1. Pure reward learning: sweetened water reward with correct choice, unrestricted access to plain water.
2. Incentive-disincentive learning with quinine: sweetened water reward with correct choice, unrestricted access to bitter tasting water (0.3mM quinine).
3. Incentive-disincentive with probability: sweetened water reward with correct choice, access to plain water with a probability of 25%.

Results

The first study using a purely appetitive learning strategy [1] revealed that a sweet reward was sufficient to induce robust place learning. Improvements in time tasks or working memory tasks were observed, compared to controls however with lower success rate and gentler slope. A sweet reward was not sufficient to induce learning in complex tasks, mice simply skipped the tasks and switched to free water consumption. We also observed that mice found a work-around to consume both free water and sweetened water at the same visit in the impulsivity task. The studies using the incentive-disincentive paradigms [2] aimed to make the ‘free’ water less attractive and thus motivating the mice to engage in the task. Increased engagement in solving the task could indeed be observed for both incentive-disincentive paradigms compared to pure appetitive learning. Place learning tasks with increasing difficulties, such as place serial reversal, place x time or diagonal sequencing were solved better in the incentive-disincentive group than the appetitive learning group, however, performance in the most complex spatial working memory task was only marginally better than in the purely appetitively motivated group. Overall, the incentive-disincentive paradigm could improve performance, without a clear difference between the two disincentive stimuli.

Conclusion

The IntelliCage, one of the automated home cage monitoring systems for group housed mice, allows to assess both behavioral parameters and cognitive performance by using controlled water access as a motivational driver. Assessment of basic behavioral measures and performance in many cognitive tasks in the IntelliCage can easily be done if free plain water or either form of the free disincentive water paradigms is offered, and correct choice is stimulated with reward learning. With this approach, the risk of dehydration in slow learners can be reduced without compromising cognitive performance.

References

1. Bramati, G., Stauffer, P., Nigri, M., Wolfer, D.P., Amrein, I. (2023). Environmental enrichment improves hippocampus-dependent spatial learning in female C57BL/6 mice in novel sweet reward-based behavioral tests in the IntelliCage. *Front Behav Neurosci* **17**, 1256744.

2. Ma, X., Schildknecht, B., Steiner, A.C., Amrein, I., Nigri, M., Bramati, G., Wolfer, D.P. (2023). Refinement of IntelliCage protocols for complex cognitive tasks through replacement of drinking restrictions by incentive-disincentive paradigms. *Front Behav Neurosci* **17**, 1232546.

Home-cage based testing: How to bring the test to the animal and not the animal to the experiment

L. Lewejohann

Freie Universität Berlin, Berlin, Germany; German Center for the Protection of Laboratory Animals at the German Federal Institute for Risk Assessment, Berlin, Germany. Lars.Lewejohann@bfr.bund.de

Abstract

Home cage monitoring (HCM) has become increasingly important. We will elaborate on the possibilities of carrying out complex behavioral tests based on the animals' home cage. This is possible using sophisticated solutions within the cage as well as gates between the home cage and a test system. By this, the animals' home cage will be of greater importance and thus we will have to make greater efforts to provide them with behaviorally appropriate accommodation.

Introduction

In conventional animal studies for biomedical research, laboratory rodents are typically assessed individually in testing setups separate from their living enclosures at specific intervals. Yet, the results of such tests may be affected by various factors, including novelty induced anxiety, experimenter effects, or lack of motivation. Thereby, crucial data might be overlooked when animals are observed only for a short time outside their home cages. A solution to these challenges involves continuously monitoring mice and rats within their home cages, allowing for a more comprehensive understanding of their behavior and responses over an extended period. A recent systematic review conducted by members of the COST action "Teatime" (www.cost-teatime.org) outlined the development and application of home cage monitoring in laboratory mice and rats [1]. In brief, an increased usage of HCM systems and more automatization is foreseeable. On Teatime's website, an extensive catalog of currently existing HCM systems is listed and will be updated regularly (cost-teatime.org/about/technologies). We see this as a very good development that will allow researchers to obtain better animal models, collect better data and, last but not least, provide a great opportunity to improve animal welfare.

Various parameters within the home cage can be observed through either manual or automated methods, e.g., behavioral parameters such as activity, social behavior, feeding and drinking, and physiological parameters like heart rate or body temperature. Automating the techniques is a great way to collect long-term data, contributing to comprehensive phenotyping and monitoring of the animals' health status. In addition to the procedures for recording day-to-day behavior, elaborate tests can also be implemented without additional intervention by experimenters. This is possible with some of the commercially available systems, but also with self-developed devices or a combination of both. In this presentation we will introduce some of the approaches used in our lab. Our research is publicly funded, which is why we feel it is very important that the systems we develop ourselves are available to the public as open source electronics, software and building instructions (including 3D printing templates).

Another important consideration is that the barren home cage environment with the most common systems (e.g., type II long cage for mice) does not necessarily provide all possibilities to show natural behavior. This includes restricted social interactions, limited opportunities for physical activity, and even boredom [2]. This is not only an important animal welfare concern but also limits research quality by yielding data which do not utilize the full potential of the model and will therefore most likely be less translatable to humans. We will demonstrate that HCM can very well be conducted in larger and enriched cages systems up to large semi-naturalistic settings.

The IntelliCage system

The IntelliCage (IC) is a standardized HCM system, featuring four corner structures with operant conditioning panels that can be used for either reward-based conditioning (such as access to water) or aversion-based conditioning (using stimuli like air puffs) [3]. Individual behavior is detected using RFID transponders that in addition allow to record activity. The system has been developed to overcome shortcomings of conventional non-home-cage-based tests for learning and memory. As the system is well described already and successfully established in many labs, we will show a few somewhat more unusual studies that we have carried out with the system instead of just presenting the IC in all its facets. The example studies presented include long term monitoring over the life-time of male mice that were tested in the IC at various stages of their life [4].

Connecting the test cage to the home cage

Some tests require that the focal animal is separated from the group to observe undisturbed behavior. This might be a limitation of HCM when a group of animals is observed in their home environment. Especially if the resources they interact with for testing purpose are limited [5]. A way to avoid this is to connect the testing arena to a standard home cage via a gate that allow individual access to the test arena. We will present a few examples for such setups using either a commercial system (AnimalGate, TSE) or a home-made gate based on the Mouse Position Surveillance System (MoPSS) developed as open source software, electronic, and 3-D printed hardware in our lab [6, 7].

Semi-naturalistic setting

The third approach we are using is 24/7 observation in semi-naturalistic settings. The systems rely largely on RFID technique allowing to monitor the activity of large groups of animals. Additional data has so far been collected through live observation. The currently rapidly developing methods of tracking (e.g., DeepLabCut [8]) and the options for linking RFID identification with individual tracks (e.g., Live Mouse Tracker [9]) open up new avenues. We will briefly outline these in this talk.

Discussion

The possibilities for observing laboratory animals have developed rapidly. In particular, the latest possibilities of computer-assisted behavioral analysis and tracking are becoming a game changer. This offers excellent opportunities to combine better research with better animal welfare. At the same time, a wide variety of individual solutions are emerging worldwide, which unfortunately are not yet all compatible with each other. However, this is a problem that can be solved in principle if the scientific and industrial partners make joint efforts. Large networks, such as the Cost action Teatime, are a good place to start.

Ethical statement

All experiments were approved by the Berlin state authority, Landesamt für Gesundheit und Soziales (LAGeSo) and were in accordance with the German Animal Protection Law (TierSchG, TierSchVersV). In addition, all study protocols were pre-registered at www.animalstudyregistry.org since 2018.

References:

1. Kahnau, P; Mieske, P; Wilzopolski, J; Kalliokoski, O; Mandillo, S; Hölter, SM; Voikar, V; Amfim, A; Badurek, S; Bartelik, A; Caruso, A; Čater, M; Ey, E; Golini, E; Jaap, A; Hrncic, D; Kiryk, A; Lang, B; Loncarevic-Vasiljkovic1, N; Meziane, H; Radzevičienė, A; Rivalan, M; Scattoni, ML; Torquet, N; Trifkovic, J; Ulfhake, B; Thöne-Reineke, C; Diederich, K; Lewejohann, L; Hohlbaum, K (2023): A systematic review of the development and application of home cage monitoring in laboratory mice and rats. BMC Biol 21, 256.

2. Mieske, P; Hobbiesiefken, U; Fischer-Tenhagen, C; Heinl, C; Hohlbaum, K; Kahnau, P; Meier, J; Wilzopolski, J; Butzke, D; Rudeck, J; Lewejohann, L; Diederich, K (2022): Bored at home? - A systematic review on the effect of environmental enrichment on the welfare of laboratory rats and mice. Frontiers in Veterinary Science 9:899219.

3. Lipp, H-P; Krackow, S; Turkes, E; Benner, S; Endo, T; Russig, H (2023): IntelliCage – the development and perspectives of a mouse- and user-friendly automated behavioral test system. Front. Behav. Neurosci. 17. doi: 10.3389/fnbeh.2023.1270538.

4. Kahnau, P; Guenther, A; Boon, MN; Terzenbach, JD; Hanitzsch, E; Lewejohann, L; Brust, V (2021): Lifetime observation of cognition and physiological parameters in male mice. Frontiers in Behavioral Neuroscience 15:709775.

5. Lang, B; Kahnau, P; Hohlbaum, K; Mieske, P; Andresen, NP, Boon, MN; Thöne-Reineke, C; Lewejohann, L; Diederich, K (2023): Challenges and advanced concepts for the assessment of learning and memory function in mice. Frontiers in Behavioral Neuroscience.

6. Kahnau, P; Jaap, A, Urmersbach, B; Diederich, K; Lewejohann, L (2023): Development of an IntelliCage-based Cognitive Bias Test for Mice [version2; peer review: 2 approved]. Open Research Europe, 2:128.

7. Habedank, A; Urmersbach, B; Kahnau, P; Lewejohann, L (2022): O mouse, where art thou? The Mouse Position Surveillance System (MoPSS) - an RFID based tracking system. Behavior Research Methods. 54, 676–689.

8. Mathis, A.; Mamidanna, P; Cury, KM; Abe, T; Murthy, VN; Weygandt Mathis, M; Matthias Bethge (2018): DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 1281–1289.

9. de Chaumont, F; Ey, E; Torquet, N et al. (2019): Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. Nat Biomed Eng, 930–942.

TeaTime COST Action presents Enhancing Reproducibility and Animal Welfare through Home Cage Systems: A Comprehensive Assessment

Home cage: a uniquely sensitive tool for detecting hidden phenotypes

Lior Bikovski[1] [2]

[1]The Myers Neuro-Behavioral Core Facility, Sackler Faculty of Medicine, Tel Aviv University, Israel
[2]School of Behavioral Sciences, Netanya Academic College, Netanya, Israel

Conventional behavioral assays are often subject to confounding factors, such as human handling. Moreover, the behaviors observed are restricted to the methods used, including conditions like light intensity and stress, as well as protocol constraints such as short assessment durations and specific times of the day. These limitations allow only a glimpse into the behavior of a mouse within an artificial setting where specific stressors elicit the desired behavior (e.g., the novelty of an open field arena, for exploration behavior, or foot shock for freezing behavior). To address these shortcomings and improve animal welfare in the pre-clinical phenotyping process, the home-cage monitoring system (HCM) was developed. With over two decades of experience integrating different HCMs and methodologies into mouse-based preclinical studies, HCM seems to be not just a solution to the limitations of conventional assays, but also a method that, when combined with other, standard methods (such as open field and fear conditioning), provides comprehensive information, enriching our preclinical differential diagnosis process."

In our facility, we employ the Phenotyper home cage system provided by Noldus Information Technology for short-term (3-5 days) in-depth assessment, and the Low-Profile Wireless running wheel by Med-Associates as a long-term (>5 days) monitoring system. The data obtained from these systems across various experiments is unique and invaluable, and has significantly contributed to our understanding and interpretation of outcomes in the studies where they were implemented.

In some instances, data collected from HCM have filled gaps of information about novel mouse models, e.g., the work of Sloin et al [1] that demonstrated contrasting activity levels of mice in their home cages compared to those observed in standard tests. This discrepancy later helped to elucidate the differences in weight gain observed in the novel model mice. Another application of the HCM in our facility involved reevaluating a previous “known behavior” of mouse models. Tseitlin et al. [2] uncovered an abnormal stress reaction in mice with mild traumatic brain (mTBI) when a stressor was introduced in the home cage, a phenomenon not observed using standard methods like elevated plus maze and the open field tests. This finding sparked a discussion about the sensitivity of the mTBI model to the context of the stressor.

In my discussion, I will highlight the added value of employing HCM systems by presenting data obtained from varoius studies. This will be evaluated in the context of information gained using conventional behavioral assessment methods.

References

1. Sloin H.E, Bikovski* L., Levi A., Amber-Vitos O., Katz T., Spivak L., Someck S., Gattegno R., Sivroni S., Sjulson L., Stark E (2022). Hybrid offspring of C57BL/6J mice exhibit improved properties for neurobehavioral research. *ENEURO*.0221-22.
2. Tseitlin, L., Richmond-Hacham, B., Vita, A., Schreiber, S., Pick, C. G., & Bikovski, L. (2023). Measuring anxiety-like behavior in a mouse model of mTBI: Assessment in standard and home cage assays. *Frontiers in Behavioral Neuroscience*, 17, 1140724.

Using Digital Biomarkers to Measure Animal Behavior for Translational Research: The 3Rs Collaborative Initiative

Stefano Gaburro[1], Lucas Noldus[2], Megan R. LaFollette[3]

Tecniplast Spa, Buguggiate, Italy.

Noldus Information Technologies, The Netherlands.

The 3Rs Collaborative, USA.

Stefano.Gaburro@tecniplast.it

Abstract

Traditional measurements of animal behavior often occur in limited time points and sometimes stressful surroundings, which can negatively affect animal wellbeing and scientific quality. However, modern technologies can continuously collect translational digital biomarkers (TDB) in animal home environments which is beneficial for science, welfare, strategy, and operations. The 3Rs Collaborative's TDB initiative facilitates collaboration of key stakeholders to advance use of TDB including learning from case studies and current engagement.

Introduction

A shift from traditional measurements of animals to technologies that leverage TDB is an impactful innovation, that allows the scientific community to conduct more precise, humane, and efficient animal studies [1]. Traditional methods often involve removing animals from their home environments for invasive, intermittent assessments which can compromise both animal welfare and data quality due to stress and data alteration [2,3]. If instead, animals are kept in familiar environments and monitored remotely, these adverse effects can be mitigated with a strategy that embraces the principles of the 3Rs – Reduce, Refine, Replace – in animal research [4].

The concept of TDB is rooted in the utilization of digital technologies to capture and analyze data from animal models. This approach is not only non-invasive, ensuring minimal distress to the subjects, but also provides continuous, real-time monitoring. Such longitudinal data collection offers a more comprehensive understanding of biological processes and disease progression, leading to enhanced accuracy and reliability of research findings that could even enhance the speed and quality of scientific research [5]. The insight that measuring rodent behavior in an enriched home cage environment has distinct advantages over conventional testing in barren arenas or chambers is not new; it has been documented many years ago [6,7]. This triggered the development of various systems for continuous collection of TDB in instrumented home cages [8], which have been used successfully for a wide range of drug discovery, mutant phenotyping and neurobiological studies [9]. However, the scale at which these technologies are used has been limited by the relative complexity and cost of systems on the one hand, and lack of acceptance outside academic centers on the other hand. Only recently, data acquisition and processing technology has advanced to a level where large-scale deployment in the vivarium comes within reach [10]. Furthermore, there is now broader consensus among pharma companies and CROs that innovation of preclinical behavioral testing is needed in order to improve the predictive validity of animal models.

Despite the great potential of TDB, challenges persist in integrating these technologies as commonplace in scientific research. To address these challenges and accelerate widespread implementation of TDB, the 3Rs Collaborative launched a TDB initiative in 2020. This initiative is a non-competitive collaboration between pharmaceutical and biotechnology companies, technology providers, and other key subject matter experts. Its goal is to advance the understanding, adoption, and regulatory acceptance of scalable TDB. In this review, we will discuss the scientific impacts, 3Rs impacts, challenges, and solutions to TDB implementation while also discussing key resources from the 3Rs Collaborative's TDB Initiative.

Scientific and 3Rs impacts of translational digital biomarkers

TDB have several positive impacts on scientific research and the 3Rs [1]. While traditional assessments have contributed valuable scientific data, they also have key limitations and negative features such as short testing durations, invasive procedures such as telemetry, frequent handling, novel environments, or requiring single housing. These factors can significantly distort animal behavior and physiological measurements while also undermining data reproducibility [11].

Conversely, TDB play a crucial role in animal assessment that prioritizes both scientific rigor and animal welfare [11]. A key strength of TDBs is their ability to facilitate continuous, longitudinal, and non-invasive monitoring of animal models. This approach of real-time data collection – as opposed to episodic data - offers a richer, more dynamic understanding of physiological and behavioral changes over time which better reflect animal health and well-being. This depth of data not only enhances the accuracy of research findings but also minimizes the potential stress and discomfort to the animals involved.

The impact of TDB is also evident in its alignment with the 3Rs principles [1,12]. By enabling more detailed and accurate data collection from a single animal – and allowing for powerful within subject analysis – TDB can reduce the overall number of animals needed for research [5,13]. Additionally, its non-invasive nature refines the research process, causing less disruption to the animal's natural behaviors and habitats [13]. Furthermore, TDB can often allow researchers to detect disease progression much faster, which can reduce the amount of time animals must spend in the vivarium and allow for earlier humane intervention points [14,15].

Challenges and solutions to implementation

While TDB are clearly impactful technologies for science and the 3Rs, there are several challenges to implementation [1]. These challenges include operational, scientific, cultural, and technical aspects. Integrating these technologies into existing research frameworks requires careful planning, training, and resources. Researchers must be adept at not only the technical aspects of these tools but also in interpreting the vast amounts of data they generate. Furthermore, there is a need for standardization in TDB methodologies to ensure consistency and reliability across studies.

Collaboration and thoughtful implementation of TDB are keys to overcoming these challenges [5]. The 3Rs Collaborative’s TDB Initiative plays a crucial role in bringing together stakeholders from various sectors – academia, industry, technology, and regulatory bodies – to foster an environment of shared learning and progress. We also have created a framework and advice for addressing challenges that includes targeted value propositions for various stakeholders. We provide key examples of successful use cases of TDB for various disease models and have created a user-friendly technology hub to help connect end-users with technology providers. These resources can be accessed through the TBD Initiative's website (https://www.na3rsc.org/tdb/). Through workshops, publications, and collaborative projects, the initiative seeks to build a robust foundation for the widespread adoption of TDB in preclinical research. As of January 2024, the TDB Initiative includes 11 pharma and biotech companies, 8 technology providers, and 2 other organizations. We look forward to welcoming more members.

Conclusion

In conclusion, modern technology now allows us to measure animal behavior in a manner that is non-invasive, translational, and beneficial for scientific, operational, and welfare reasons. Although there are key challenges for adopting TDB, this domain of technology is well worth the effort. The 3Rs Collaborative’s TDB Initiative provides a venue for collaboration, education, and promotion of these technologies. In the future, this initiative will also forward efforts related to validation, verification, and regulatory acceptance. This initiative is a key part of our mission to advance better science – for both people and animals.

References

1. Baran, S. W., Bratcher, N., Dennis, J., Gaburro, S., Karlsson, E. M., Maguire, S., Makidon, P., Noldus, L. P. J. J., Potier, Y., Rosati, G., Ruiter, M., Schaevitz, L., Sweeney, P., & LaFollette, M. R. (2022). Emerging role of translational digital biomarkers within home cage monitoring technologies in preclinical drug discovery and development. *Frontiers in Behavioral Neuroscience*, **15**.
2. Balcombe, J. P., Barnard, N. D., & Sandusky, C. (2004). Laboratory routines cause animal stress. *Contemporary Topics in Laboratory Animal Science*, **43**(6), 42–51.
3. Gerdin, A.-K., Igosheva, N., Roberson, L.-A., Ismail, O., Karp, N., Sanderson, M., Cambridge, E., Shannon, C., Sunter, D., Ramirez-Solis, R., Bussell, J., & White, J. K. (2012). Experimental and husbandry procedures as potential modifiers of the results of phenotyping tests. *Physiology & Behavior*, **106**(5), 602–611.
4. Russell, W. M. S., & Burch, R. L. (1959). The principles of humane experimental technique. Methuen.
5. Baran, S. W., Gupta, A. D., Lim, M. A., Mathur, A., Rowlands, D. J., Schaevitz, L. R., Shanmukhappa, S. K., & Walker, D. B. (2020). Continuous, automated breathing rate and body motion monitoring of rats with paraquat-induced progressive lung injury. *Frontiers in Physiology*, **11**, 569001.
6. Wahlsten, D., Rustay, N.R., Metten, P. & Crabbe, J.C. (2003). In search of a better mouse test. *Trends in Neuroscience*, **26**, 132–136.
7. De Visser, L., van den Bos, R., Kuurman, W.W., Kas, M.J.H. & Spruijt, B.M. (2006). Novel approach to the behavioural characterization of inbred mice: automated home cage observations. *Genes, Brain and Behavior*, **5**, 458–466.
8. Spruijt, B.M. & De Visser, L. (2006). Advanced behavioural screening: automated home cage ethology. *Drug Discovery Today: Technologies,* **3** (2), 231-237.
9. Grieco, F.; Bernstein, B.J.; Biemans, B.; Bikovski, L.; Burnett, C.J.; Cushman, J.D.; van Dam, E.A.; Fry, S.A.; Hacham, B.R.; Homberg, J.R.; Kas, M.J.H.; Kessels, H.W.; Koopmans, B.; Krashes, M.J.; Krishnan, V.; Logan, S.; Loos, M.; McCann, K.E.; Parduzi, Q.; Pick, C.G.; Prevot, T.D.; Riedel, G.; Robinson, L.; Sadighi, M.; Smit, A.B.; Sonntag, W.; Roelofs, R.F.; Tegelenbosch, R.A.J.; Noldus, L.P.J.J. (2021). Measuring behavior in the home cage: Study design, applications, challenges, and perspectives. *Frontiers in Behavioral Neuroscience*, **15**, 735387.
10. Pernold, K., Iannello, F., Low, B. E., Rigamonti, M., Rosati, G., Scavizzi, F., Wang, J., Raspa, M., Wiles, M. V., & Ulfhake, B. (2019). Towards large scale automated cage monitoring—Diurnal rhythm and impact of interventions on in-cage activity of C57BL/6J mice recorded 24/7 with a non-disrupting capacitive-based technique. *PloS One*, **14**(2), e0211063.
11. Kahnau, P., Mieske, P., Wilzopolski, J., Kalliokoski, O., Mandillo, S., Hölter, S. M., Voikar V., Amfin, A., Badurek, S., Bartelik, A., Caruso, A., Čater, M., Ey, E., Golini, E., Jaap, A., Hrncic, D., Kiryk, A., Lang, B., Loncarevic-Vasiljkovic, N.,… & Hohlbaum, K. (2023). A systematic review of the development and application of home cage monitoring in laboratory mice and rats. *BMC biology*, **21**(1), 256.
12. Voikar, V., & Gaburro, S. (2020). Three Pillars of Automated Home-Cage Phenotyping of Mice: Novel findings, refinement, and reproducibility based on literature and experience. *Frontiers in Behavioral Neuroscience*, 14, 575434.
13. Steele, A. D., Jackson, W. S., King, O. D., & Lindquist, S. (2007). The power of automated high-resolution behavior analysis revealed by its application to mouse models of Huntington’s and prion diseases. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(6), 1983–1988.
14. Roughan, J. V., Wright-Williams, S. L., & Flecknell, P. A. (2009). Automated analysis of postoperative behaviour: Assessment of HomeCageScan as a novel method to rapidly identify pain and analgesic effects in mice. *Laboratory Animals*, **43**(1), 17–26.
15. Zentrich, E., Talbot, S. R., Bleich, A., & Häger, C. (2021). Automated home-cage monitoring during acute experimental colitis in mice. *Frontiers in Neuroscience*, **15**, 760606.

Symposium: AI Advances in pose estimation and behaviour recognition in laboratory animals

Multi-view triangulation-enabled annotation for multi-animal 3D pose in SLEAP

Liezl Maree[1], Shayan Afshar[1,2], Stefan Oline[2], Eric J. Leonardis[1], Annegret L. Falkner[2], and Talmo D. Pereira[1*]

[1]Salk Institute for Biological Studies, La Jolla, CA, USA.
2Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA.
***Address correspondence to: talmo@salk.edu**

Introduction

Deep learning-based markerless motion capture allows us to measure the kinematics of behavior at unprecedented resolution. SLEAP (Social LEAP Estimates Animal Poses) is a highly accessible open-source deep learning framework for markerless multi-animal 2D pose estimation that makes it particularly easy to track multiple interacting animals, making it useful for the study of social behaviors [1]. This type of behavior presents technical challenges, however, since social behaviors typically occur at close quarters, leading to frequent occlusions which result in loss of information and inferior tracking performance. To mitigate the effect of occlusions, we can use multiple cameras positioned at distinct viewpoints, affording practitioners the ability to track kinematics in 3D.

Multiple systems for multi-animal pose tracking are in development which project multi-view 2D pose into a 3D space [2]. Markerless deep learning based multi-animal 3D tracking systems have been developed for a variety of species, including rodents [3], monkeys [4], pigs [5], humans [6], and more [7]. A key challenge is generation of ground truth data for training and evaluating models. Here we present an extension to SLEAP, which enables the annotation of synchronized, multi-view, multi-animal pose, as well as 3D capabilities through integration with Anipose. Anipose is an open-source Python toolkit for camera calibration and markerless 3D pose triangulation. We propose practical solutions for the multi-view association problem and provide a usable pipeline for the multi-animal 3D pose tracking workflow. The utility of this approach is demonstrated through preliminary results on a large-scale multi-animal 3D dataset of freely-moving rodents in typical lab settings.

Methods

2D annotation pipeline

To understand the 3D annotation process, let us first review the existing 2D annotation pipeline. SLEAP currently ships with an easy-to-use annotation GUI which allows users to label body parts to track throughout a video. Users first create a “skeleton” to define which body parts to track and their connection to one another via “edges.” After loading a video into SLEAP, users can begin annotating on a frame by frame basis. Within each frame, the annotator should create an “instance” for each animal in the frame. Each instance uses the skeleton to create visual markers (or “nodes”) for the body parts to be labeled. The annotator then drags and drops each node to its corresponding body part location in the frame. The user should take care to annotate all animals in the frame or none at all, as leaving a frame half-labeled has the potential to teach the model to follow suit and also half-label frames.

Creating a few instances from scratch is the first step before entering the human-in-the-loop training cycle. The human-in-the-loop cycle starts by using a batch of frames annotated by the user to train a model that outputs predictions on unlabeled frames. The user then fixes these predictions and enters another round of this training cycle. Correcting the predictions output from a model is much faster than creating new annotations from scratch. Labeling and training the model in incremental batches also gives the user some feedback on which poses or lighting conditions may need more annotations.

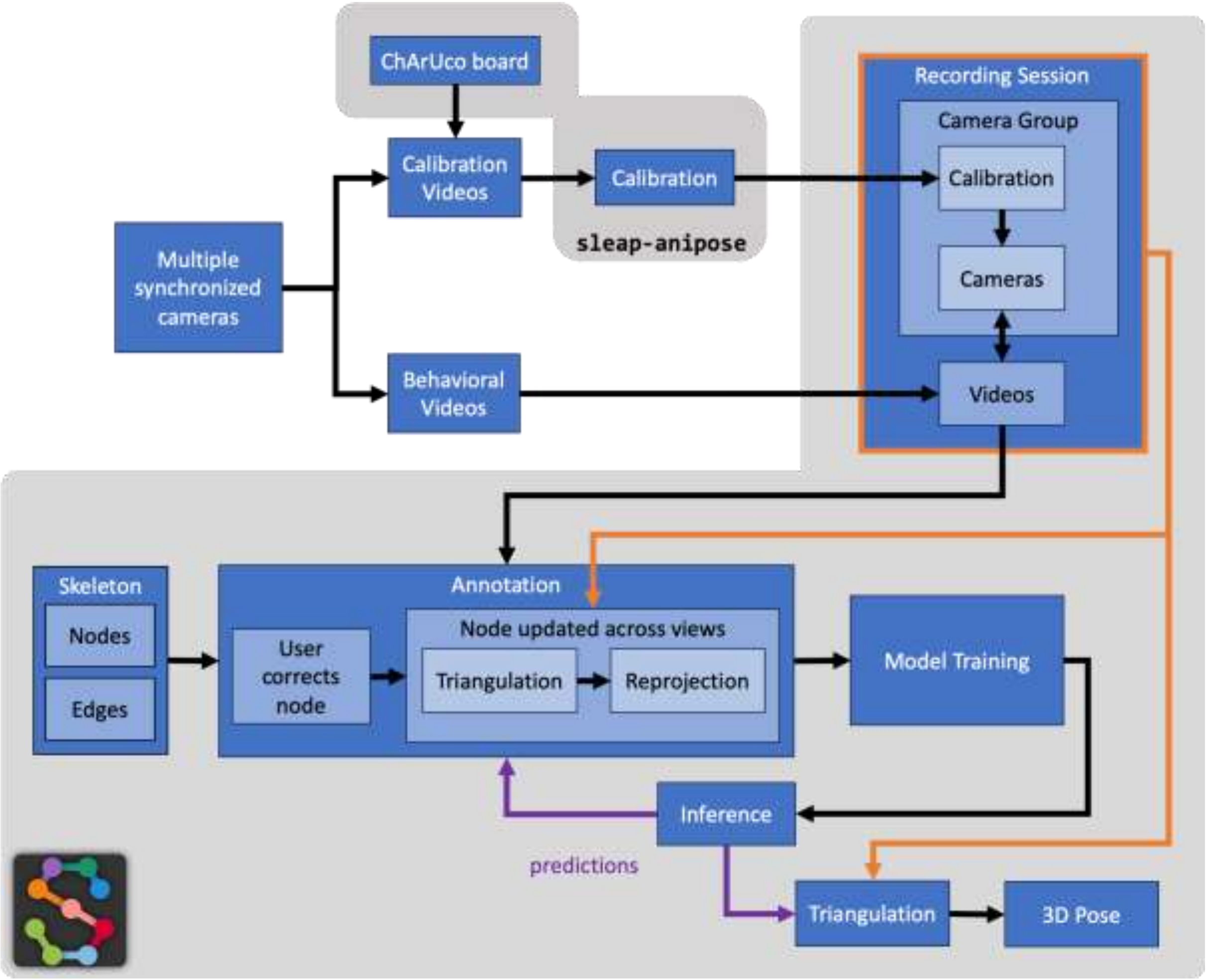


Figure 1. Flowchart of the 3D annotation pipeline using SLEAP and sleap-anipose.

3D annotation pipeline

The GUI for 3D annotation retains the usability of the original SLEAP GUI for 2D annotations with some additional improvements. When labeling, user's still only view and *directly* annotate one 2D frame at a time. However, after labeling the same temporal frame across at least two camera views, SLEAP is able to both correct and create new annotations on all camera views. The magic of 3D annotation starts with uploading a calibration file into SLEAP which parametrizes a multi-camera "recording session." The calibration file contains the extrinsic and intrinsic parameters of the cameras which allows SLEAP to take any points annotated across at least two views, triangulate them into 3D, and finally reproject the 3D points back onto all views as 2D points. The aforementioned calibration file can be created with a utility library sleap-anipose (version 0.1.7) described in further detail in the **Triangulation with anipose** section where we also outline triangulation and reprojection [8].

After uploading all related videos, the user can link videos to different cameras within a recording session. SLEAP assumes that these videos are synced, meaning that videos across different camera views start and end at the same time and have the same frame rate. Now, the user can start the multi-view annotation following a very similar user-in-the-loop training cycle with the only additional step being to annotate all frames across camera views before moving onto a new temporal frame. To easily switch between views, SLEAP has two shortcut keys to navigate forward and backwards between camera views at the same temporal frame index. After annotating the first two views, SLEAP uses these annotations to triangulate and reproject new annotations onto the remaining views. Each update made to a point in one view will also appear in related views. See Figure 1 for a visual depiction of the 3D annotation pipeline.

In the existing 2D SLEAP, annotations exist in two states. User-labeled annotations are "finalized" and ready for training, while model-generated annotations remain "pending" and require user validation before incorporation into training data. With the introduction of multi-view functionality, SLEAP now needs to account for an additional layer of annotation status. All annotations modified or created through reprojection will contribute to training, but users should have the ability to mark a point as “fixed” or “immutable,” preventing further adjustments during subsequent reprojections. SLEAP repurposes a feature initially designed for visual cues, changing a node's label color from green to red when a user manually positions a node, to indicate whether it should undergo updates during reprojection as noted in Figure 2 and Figure 3.

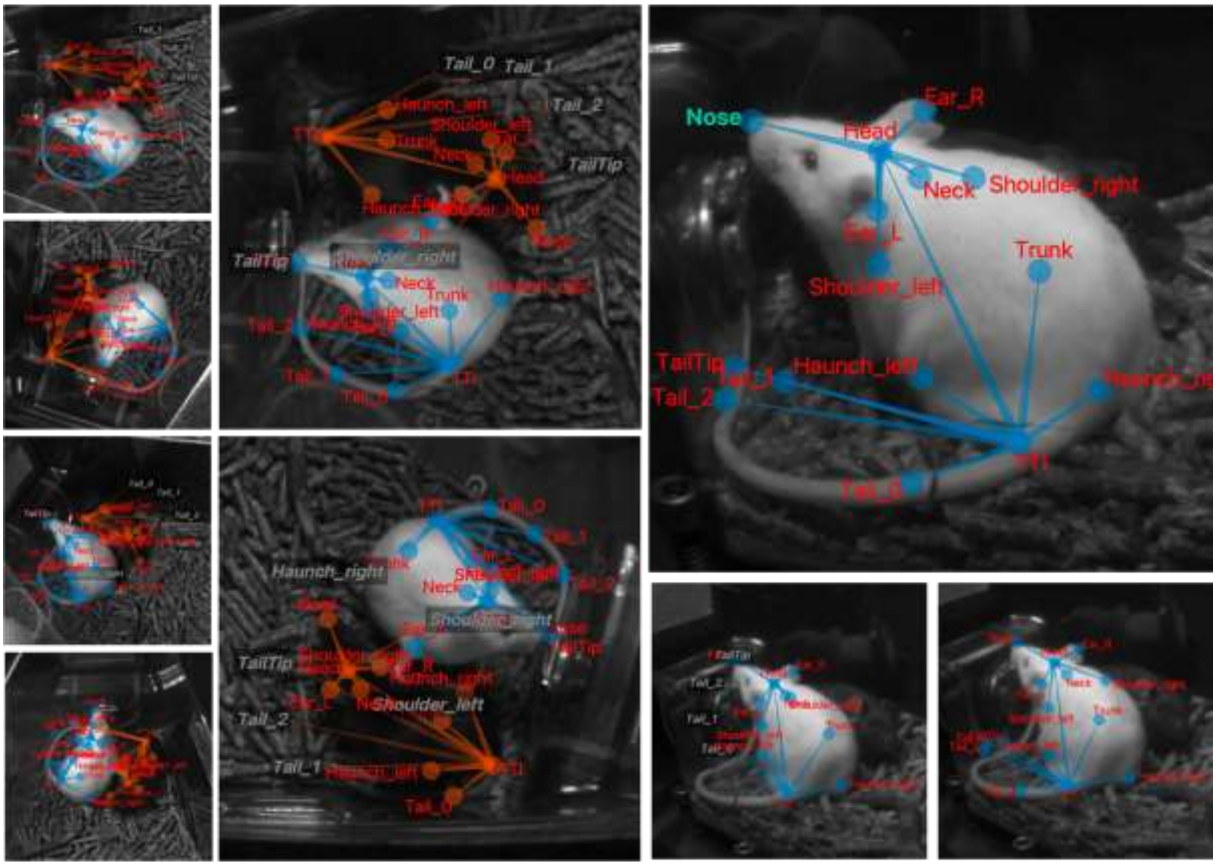


Figure 2. A collection of images from the same temporal frame index. All annotations seen above were created by double clicking predictions output from a trained model. Note that the “Nose” label in the upper right hand image is colored green to designate that the node should not be updated during reprojection.

Triangulation with anipose

Multi-view annotation within SLEAP is made possible through the utility library sleap-anipose. Extending off Aniposelib (version 0.4.3), sleap-anipose is designed for integration with multi-view SLEAP and exposes an API with convenient inputs and outputs for our needs [9]. SLEAP not only relies on the sleap-anipose library to create the calibration file needed for a recording session, but also depends on sleap-anipose to handle all aspects of triangulation and reprojection.

Calibration is performed by iteratively estimating camera parameters and 3D points then comparing them with some ground truth labels. The ground truth labels are usually estimated by using a highly recognizable stimulus, in this case, a checkerboard with Aruco patterns known as a ChArUco board. The bit encodings of the ChArUco patterns ensure that the board has a unique orientation. A calibration video is generated by pointing multiple synchronized camera views at the center of the arena, and the ChArUco board is placed in the center of the arena. Then an optimization procedure known as sparse bundle adjustment is performed to find the camera parameters and 3D points associated with the multiple views of the board. The calibration is stored as a toml file which stores the name, resolution, intrinsics, distortions, and extrinsics (rotation and translation) of each camera.

With the calibration file in hand, any point labeled across multiple overlapping views can be triangulated into 3D coordinates. Using just a single camera, we can draw a light ray from the camera’s focal point through a labeled node into 3D space. Without any other constraints on where that node is in 3D space, we can only guess that the node is somewhere along the light ray which unfortunately yields infinite possibilities. However, since the calibration file gives information on the relative positions of each camera to one another, we can narrow down where the 3D coordinate is located. If we draw light rays from a camera's focal point through a labeled node on the image plane for all cameras, we add a new constraint to the 3D point for each camera view considered. The most accurate 3D estimate of the node would be the coordinate where all light rays intersect. This process is known as triangulation, and although it sounds simple to implement, when taking into account lens distortion, which turns a light ray into something more akin to a light curve, the 3D estimate can easily lose accuracy without correct adjustment to the intrinsic camera parameters. Calibration effectively handles the intrinsic parameters, so we can confidently move onto reprojection.

Reprojection is essentially the node-update step for multi-view SLEAP. Once a node is moved, its 2D coordinates are triangulated to a 3D estimate, and finally the node coordinates are corrected to the reprojected 2D coordinates. In more detail, reprojection is when the estimations of the 3D points are multiplied by the camera parameters to transform the coordinates back into each 2D view. The reprojection error is calculated by taking the distance between the reprojected coordinates and the 2D ground truth labels. Reprojection provides a novel bootstrapping method where 3D estimates can be used to improve less reliable 2D camera views. This means that reprojection from 3D world coordinates can be used to recover poorly inferred keypoints and make pose estimation more robust. So, by adjusting a label in one view, we are able to correct the label across all views. This is particularly useful when a body part is difficult or impossible to see from a certain angle, but easily found in another view.

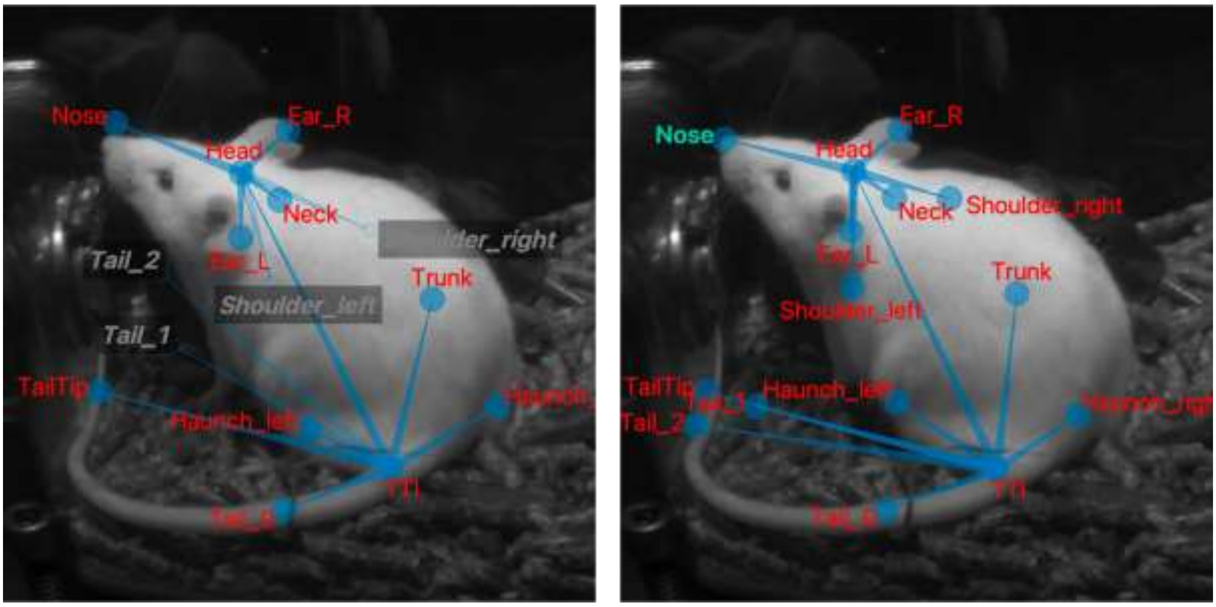


Figure 3. Two images from the same camera view and temporal frame showing a mouse before (left) and after (right) triangulation and reprojection. All nodes with green labels are considered "fixed" and are not updated during reprojection. In the above figure, the user updated the "Nose" node and the rest of the nodes were updated through reprojection. Most notably, the shoulders and "Tail_2" were updated to seemingly correct locations, while "Tail_1" still requires user adjustment.

Multi-view association

SLEAP is known for its ability to track *multiple* animals over time which can only be done by solving a temporal identity problem. SLEAP currently solves the temporal identity problem by either requiring the user to provide additional identity labels for training one of SLEAP's ID-based models, or by using a distance-based metric to determine identities across sequential frames.

Extending SLEAP to multiple views creates an additional across-view identity problem. To properly triangulate, SLEAP needs to be able to group together animals of a unique identity across camera views. While SLEAP could require that users manually label the identities of animals across views, to avoid additional strain to the annotator, SLEAP instead implements automated "hypothesis testing" to determine the correct identities for each animal as depicted in Figure 4 and Figure 5. Exhaustive hypothesis testing is a brute force approach for generating all possible groupings for animal identities, triangulating and reprojecting all groupings, and then measuring the reprojection error on every view for each grouping. This method of hypothesis testing finds the best grouping as the grouping with the lowest reprojection error, thereby solving the multi-view association problem. The major limitation of this approach is that the number of hypotheses generated is exponential in the number of cameras used (Figure 4). This limits the utility of this approach for realtime use depending on the scale of the setup, but is appropriate for most lab settings with relatively few animals and views.

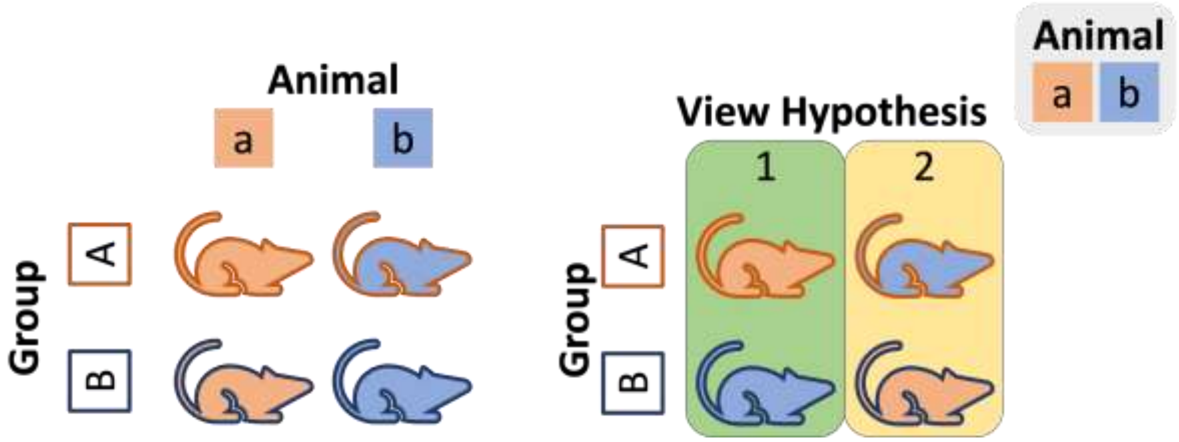


Figure 4. For a single view, hypotheses are generated by permuting the instances in the current view into each of the available instance groups. The number of available instance groups is equal to the number of unique animals in the entire video. This yields "the number of unique instances" factorial hypotheses for a single view. In this example there are 2 unique instances and, therefore, 2 grouping hypotheses with View Hypothesis 1 being the correct hypothesis.

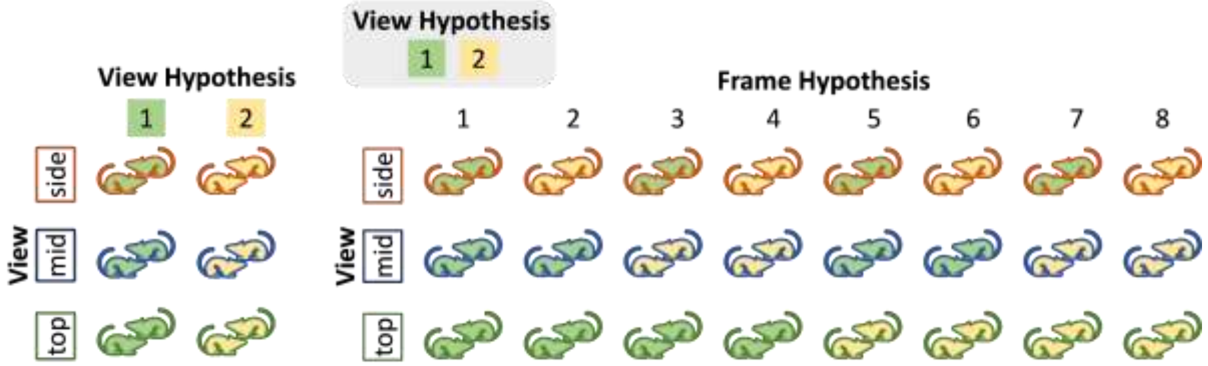


Figure 5. To generate frame-wide instance grouping hypotheses, we need to consider combinations of all possible view hypotheses. Thus, the number of frame-wide hypotheses created is "number view hypotheses" raised to the "number of views." From Figure 4, we know the correct view hypothesis is View Hypothesis 1, so the correct frame hypothesis would require all views to use View Hypothesis 1 (as is done in Frame Hypothesis 1).

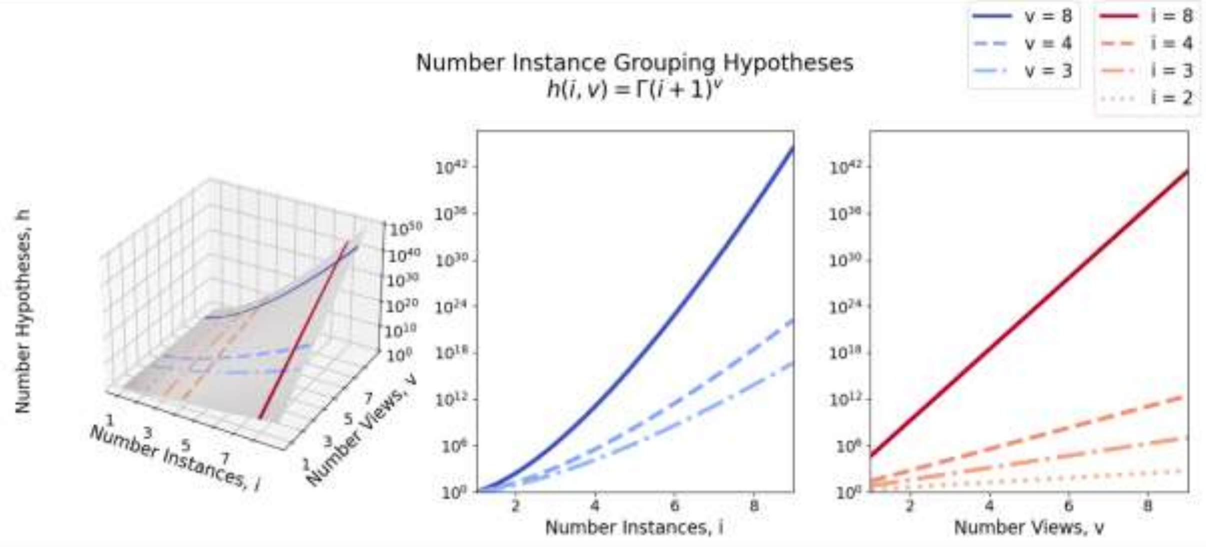


Figure 6. Plot showing the number of hypotheses generated is exponential in the number of camera views and factorial in the number of instances. The blue colored regions indicate values where the number of views are fixed while the red colored regions indicate values when the number of instances are fixed. These regions are plotted using a logarithmic scale to analyze the effects of views and instances on hypotheses separately. For visualization purposes, we use the gamma function *Γ* to interpolate the factorial function to non-integer values.

Future directions

Complete GUI integration

The GUI integration is the key piece that makes multi-view SLEAP a useful tool for its intended audience, i.e., everyone, regardless of programming ability. Multi-view SLEAP is still incomplete because users are unable to go through the entire 3D annotation pipeline via the GUI. So while the behind-the-scenes data structures are in place for a proof of concept, multi-view SLEAP still needs to add graphical elements for interacting with all adjustable aspects of these data structures.

Expose more functionality from sleap-anipose

While we aim to keep tasks modularized, i.e., sleap-anipose should handle all aspects related to calibration, triangulation and reprojection; it would be convenient if the user could access more of these utilities from within the SLEAP GUI. One direction would be to extend the 3D pipeline of SLEAP to include camera calibration. So, instead of expecting a calibration file as the starting point, the user could generate the calibration file from within SLEAP.

Add more functionality to sleap-anipose

Building on top of this idea of extending functionality, it would also be nice to take away some limiting factors such as requiring videos to be synchronized. While this requirement is fairly reasonable, providing a way to handle unsynchronized videos would make multi-view SLEAP easier to use in less constrained settings. One way to do this would be to permit manual or semi-automated resynchronization through the interface.

Faster multi-view association

More testing needs to be done on what assumptions can be made to lower the cost of automatic grouping across views. One idea is to allow users to manually set or verify a grouping for a unique instance across all views as a “hard assignment”, then perform hypothesis testing by only permuting unassigned instances into groups. Building off this idea of hard assignments, SLEAP could perform a greedy version of hypothesis testing as follows. Imagine a user moves to a completely unlabeled temporal frame. The user then labels all instances for a single camera view before moving to the next view. As soon as the user labels an instance in the second view, a round of hypothesis testing will take place in which the new instance will be placed into an instance grouping. However, in the greedy version of hypothesis testing, whichever group the instance is assigned to will be treated as a hard assignment, unable to be changed. This would effectively lower the number of hypotheses generated down to “the number of existing instance groups plus one" (the plus one is for the case when the new instance is not in any existing instance group).

Conclusions

In this paper, we have provided an overview of multi-animal 3D pose estimation methodologies and extended SLEAP to allow for rapid annotations of multiple 2D views using 3D reprojection. Having more camera views requires more annotations, these issues can be ameliorated by bootstrapping annotations from other camera views. This can be achieved by properly calibrating the cameras to get intrinsic, extrinsic and distortion parameters which allow for the projection from each 2D view to a 3D world view and back. These reprojections from 3D to 2D can also be useful for tracking identities for a small number of animals but becomes more challenging when the number of animals increases. It is common for many labs to examine dyadic and triadic interactions, so despite the limitations, this tool should still be useful for many experimental contexts. Integrating aspects of the Anipose package into the SLEAP GUI will lead to an improved user experience for annotating multi-view multi-animal data. We expect these changes to the GUI to be integrated in the near future and expect it to have a significant impact on the labeling process for 3D multi-animal pipelines.

Ethical Statement

The acquisition of video for experiments depicted in Figure 2 and Figure 3 was approved by Princeton's IACUC. Methods consisted of placing mice into a wedge cage for 5-60 minute intervals and observing behavior.

References

1. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., McKenzie-Smith, G.C., Mitelut, C.C., Castro, M.D., D'Uva, J., Kislin, M., Sanes, D.H., Kocher, S.D., S-H, S., Falkner, A.L., Shaevitz, J.W., Murthy, M. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature Methods* **19**(4).

2. Marshall, J. D., Li, T., Wu, J. H., & Dunn, T. W. (2022). Leaving flatland: Advances in 3D behavioral measurement. Current opinion in neurobiology, 73, 102522.

3. Marshall, J. D., Klibaite, U., Gellis, A., Aldarondo, D. E., Ölveczky, B. P., & Dunn, T. W. (2021). The PAIR-R24M Dataset for Multi-animal 3D Pose Estimation. *Proceedings of the 35th Neural Information Processing Systems Conference* (NeurIPS 2021).

4. Marks, M., Qiuhan, J., Sturman, O., von Ziegler, L., Kollmorgen, S., von der Behrens, W., Mante, V., Bohacek, J., & Yanik, M. F. (2022). Deep-learning based identification, tracking, pose estimation, and behavior classification of interacting primates and mice in complex environments. Nature machine intelligence, 4(4), 331–340.

5. An, L., Ren, J., Yu, T., Hai, T., Jia, Y., & Liu, Y. (2023). Three-dimensional surface motion capture of multiple freely moving pigs using MAMMAL. Nature Communications, 14(1), 7727.

6. Long, C., Ai, H., Chen, R., Zhuang, Z., & Liu, S. (2020). Cross-view tracking for multi-human 3D pose estimation at over 100 fps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3279-3288).

7. Kevin Luxem, Jennifer J Sun, Sean P Bradley, Keerthi Krishnan, Eric Yttri, Jan Zimmermann, Talmo D Pereira, Mark Laubach (2023) Open-source tools for behavioral video analysis: Setup, methods, and best practices eLife 12:e79305.

8. Afshar, S., Pereira, T.D. (2023). sleap-anipose: SLEAP to Anipose triangulation pipeline for 3D multi-animal pose tracking [Computer software]. <https://github.com/talmolab/sleap-anipose>. Accessed 19 December 2023.

9. Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., Tuthill, J.C. (2021). Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports* **36**(13).

Parsing the sub-second structure of animal behavior with Keypoint-MoSeq

Caleb Weinreb[1], Jonah E. Pearl[1], Sherry Lin[1], Mohammed Abdal Monium Osman[1], Libby Zhang[2,3], Sidharth Annapragada[1], Eli Conlin[1], Red Hoffmann[1], Sofia Makowska[1], Winthrop F. Gillis[1], Maya Jay[1], Shaokai Ye[4], Alexander Mathis[4], Mackenzie Weygandt Mathis[4], Talmo Pereira[5], Scott W. Linderman[3,6,*] and Sandeep Robert Datta[1,*]

1Department of Neurobiology, Harvard Medical School, Boston, MA, USA

2Department of Electrical Engineering, Stanford University, Stanford, CA, USA.

3Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA.

4Brain Mind and Neuro-X Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

5Salk Institute for Biological Studies, La Jolla, USA

6Department of Statistics, Stanford University, Stanford, CA, USA.

***email: scott.linderman@stanford.edu; srdatta@hms.harvard.edu**

Animals generate a vast array of behaviors by composing together a small set of stereotyped actions [1-3]. By classifying these action motifs (or "syllables") and identifying their occurrence in behavioral recordings, researchers can better understand the complex ways that behavior is altered by drugs, genes, neurons, and environmental affordances. Traditionally, such classifications were performed manually by trained observers. But over the last decade, machine learning has made it possible to automate the process, allowing unbiased analyses of large datasets with great sensitivity and consistency.

Machine-based behavioral analysis typically consists of two steps: pose tracking and action segmentation. Pose tracking methods (which include SLEAP [4], DeepLabCut [5] and others [6, 7]) enable users to specify and track keypoints corresponding to body parts in videos of behaving animals, and thereby quantify movement kinematics. These methods are simple to implement, applicable to a wide range of video data, and have been widely adopted by the neuroscience community.

However, action segmentation based on keypoint tracking – in which continuous pose dynamics are divided into intervals representing distinct movement motifs, such as turns, grooms, or darts – remains a difficult and unsolved problem[8-10]. While several segmentation methods exist [11-16], their underlying logic and assumptions differ, with different methods giving distinct descriptions of the same behavior [12, 15]. Furthermore, pervasive noise in keypoint tracking frequently contaminates behavior inferences [12, 17]. Thus, there is an important gap between our access to movement kinematics and our ability to understand how kinematics are organized to impart structure upon behavior.

To address this gap, we developed keypoint-MoSeq, an unsupervised machine learning algorithm that transforms pose tracking data into a set of behavioral motifs (like rears, turns and pauses) called syllables. Keypoint-MoSeq differs from prior methods of action segmentation by (1) employing an explicit Bayesian model that articulates clear assumptions about the timescale and structure of syllables and how they are composed into sequences during ongoing behavior [16, 18-21]; (2) simultaneously inferring correct pose dynamics from noisy or missing data based on the learned behavioral repertoire (Fig 1).

We benchmarked keypoint-MoSeq by comparing it to alternative behavioral clustering methods (including B-SOiD [11], VAME [12] and MotionMapper [22]). We find that keypoint-MoSeq preserves important information about behavioral timing and outperforms alternative methods at recognizing behavioral transitions in kinematic data, capturing systematic fluctuations in neural activity [23], and identifying complex features of solitary and social behavior highlighted by expert observers [24, 25] (Figure 2a-j). Furthermore, we show that keypoint-MoSeq generalizes beyond mouse syllables to capture behaviors at multiple timescales and in several species, including the fly [22, 26].

Given that keypoint-MoSeq can be applied to 2D or 3D keypoint tracking data, naturally accommodates noise and missing observations, and functions across species and timescales, we anticipate that it will serve as a general tool for understanding the structure of behavior. To facilitate broad adoption of this approach, we have built keypoint-MoSeq to be directly integrated with widely-used keypoint tracking methods (including SLEAP and DeepLabCut), and have made keypoint-MoSeq code freely accessible for academic users at www.MoSeq4all.org; the modular codebase includes novice-friendly Jupyter notebooks to enable users without extensive computational experience to use keypoint-MoSeq, methods for motif visualization in 2D and 3D, a pipeline for post-hoc analysis of the outputs of keypoint-MoSeq, and a hardware-accelerated and parallelization-enabled version of the code for analysis of large datasets.

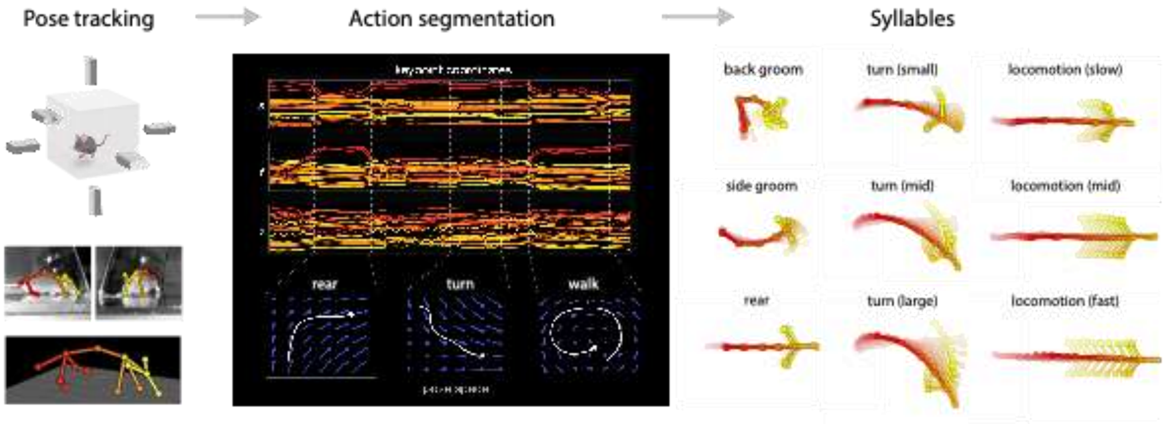


Figure 1: Keypoint-MoSeq pipeline. Left: animal behavior is recorded with cameras either in 2D (single camera) or 3D (multiple cameras, as shown in the figure). The locations of body parts (referred to as “keypoints”) are detected in each frame. Center: the animal’s pose trajectory is then segmented into action motifs or “syllables” using keypoint-MoSeq. Each syllable encompasses a stereotyped pattern in the dynamics of the keypoint coordinates (top), which is defined mathematically as dynamical system in pose space (bottom). Right: the syllables – which are discovered algorithmically without human input – can then be labeled post hoc by a human observer, e.g., as grooms, rears, turns etc.

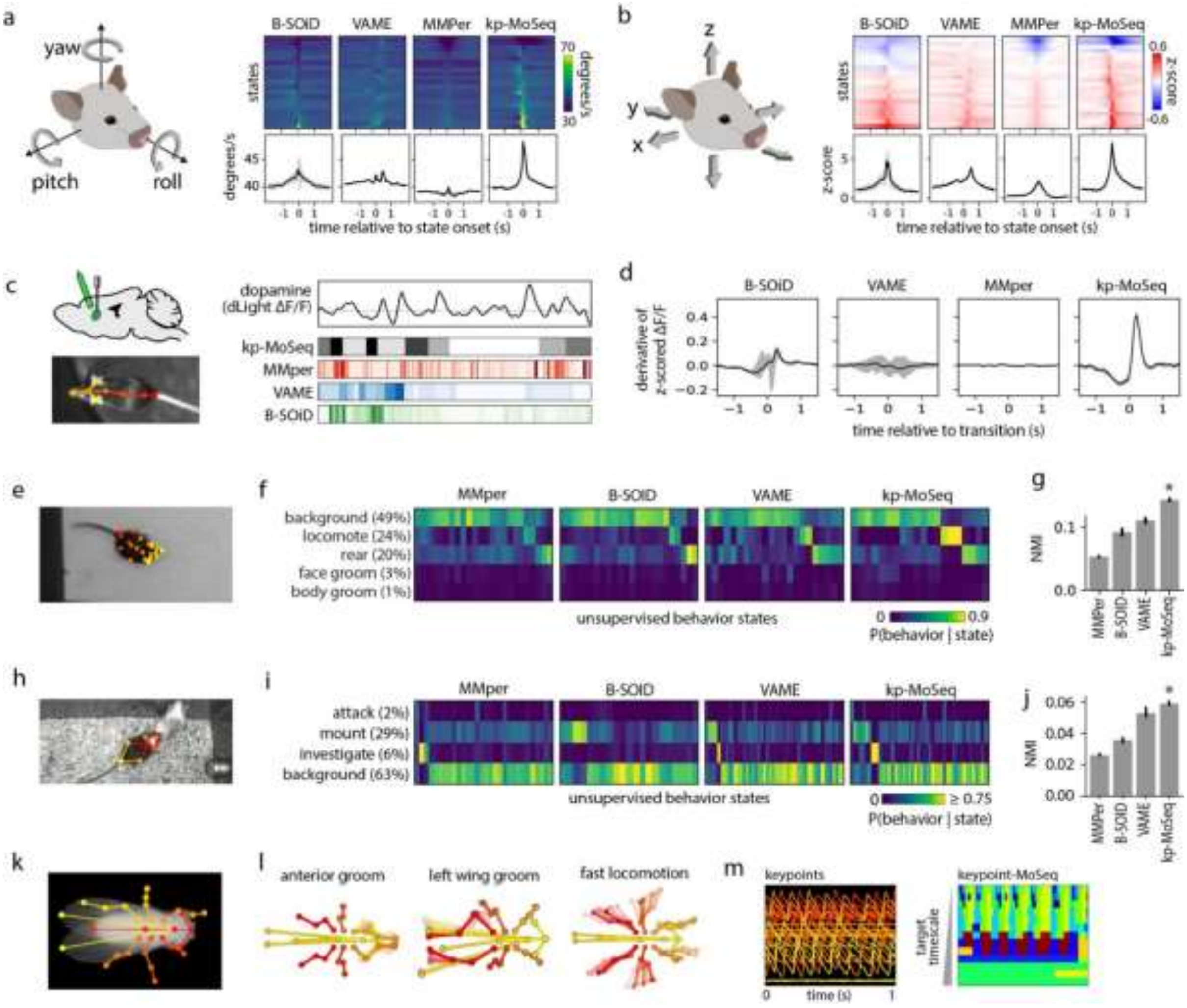


Figure 2: Validation of keypoint-MoSeq. (a) Angular velocity was captured using a head-mounted inertial measurement unit (IMU) while mice explored an open field. Average angular velocity at motif onset is shown in the heatmaps above, with the median across all motifs plotted below is plotted below. (b) As (a), but now showing jerk (derivative of acceleration). Both angular velocity and jerk display more pronounced peaks at the onset of keypoint-MoSeq-derived motifs, indicating a greater alignment of these motifs with sudden changes in the animal's kinematics ($P < 0.0005$, N=10). (c) Dopamine transients in dorsolateral striatum – measured using the fluorescent indicator dLight – were aligned to behavior motifs from several segmentation methods. (d) Keypoint-moseq motifs – but not those of other methods – tend to align with sudden increases in dopamine release. (e) Keypoint tracking was applied to a benchmark dataset in which human labelers annotated a set of behaviors while mice explored an open field. (f) Overlap between human annotations and unsupervised behavior motifs from four segmentation methods. (g) Normalized mutual information (NMI) between unsupervised behavior motifs and human labels, showing the median (gray bars) and inter-quartile interval (black lines) across N=20 independent model fits. Keypoint-MoSeq consistently has higher NMI ($P < 10-6$). (h-j) As (e-g), but now showing a benchmark dataset of human-annotated social interactions. (k-m) Results of keypoint-MoSeq applied to videos of flies on a 2D substrate. (l) Three example motifs identified by keypoint-MoSeq. (m) Example of keypoint-MoSeq outputs for 1-second interval while the fly is walking. Keypoint coordinates are plotted on the left, and motif sequences for a range of target timescales are plotted on the right.

Ethical compliance

All experimental procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (Protocol Number 04930) and were performed in compliance with the ethical regulations of Harvard University as well as the Guide for Animal Care and Use of Laboratory Animals.

References

1. Tinbergen, N. The study of instinct. Oxford,: Clarendon Press; 1951. 228 p. p.
2. Dawkins, R. Hierarchical organisation: A candidate principle for ethology. Growing points in ethology. Oxford, England: Cambridge U Press; 1976.
3. Baerends, GP (1976). The functional organization of behaviour. Animal Behaviour The functional organization of behaviour 24: pages.
4. Pereira, TD, Tabris, N, Matsliah, A, Turner, DM, Li, J, Ravindranath, S, et al. (2022). Sleap: A deep learning system for multi-animal pose tracking. Nature Methods Sleap: A deep learning system for multi-animal pose tracking 19: pages.
5. Mathis, A, Mamidanna, P, Cury, KM, Abe, T, Murthy, VN, Mathis, MW, et al. (2018). Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. Nature Publishing Group Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning 21: pages.
6. Graving, JM, Chae, D, Naik, H, Li, L, bioRxiv, BK (2019). Fast and robust animal pose estimation. biorxivorg Fast and robust animal pose estimation pages.
7. Sun, JJ, Ryou, S, Goldshmid, RH, Weissbourd, B, Dabiri, JO, Anderson, DJ, et al. (2022). Self-supervised keypoint discovery in behavioral videos. 2022 Ieee Cvf Conf Comput Vis Pattern Recognit Cvpr Self-supervised keypoint discovery in behavioral videos 00: pages.
8. Datta, SR, Anderson, DJ, Branson, K, Perona, P, Leifer, A (2019). Computational neuroethology: A call to action. Neuron Computational neuroethology: A call to action 104: pages.
9. Anderson, DJ, Perona, P (2014). Toward a science of computational ethology. Neuron Toward a science of computational ethology. 84: pages.
10. Pereira, TD, Shaevitz, JW, Murthy, M (2020). Quantifying behavior to understand the brain. Nature Neuroscience Quantifying behavior to understand the brain 23: pages.
11. Hsu, AI, Yttri, EA (2021). B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. Nature Communications B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors 12: pages.
12. Luxem, K, Mocellin, P, Fuhrmann, F, Kursch, J, Miller, SR, Palop, JJ, et al. (2022). Identifying behavioral structure from deep variational embeddings of animal motion. Commun Biol Identifying behavioral structure from deep variational embeddings of animal motion 5: pages.
13. Berman, GJ, Choi, DM, Bialek, W, Shaevitz, JW (2013). Mapping the structure of drosophilid behavior. Mapping the structure of drosophilid behavior pages.
14. Marques, JC, Lackner, S, Félix, R, Orger, MB (2018). Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering. Current Biology Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering 28: pages.
15. Todd, JG, Kain, JS, de Bivort, BL (2017). Systematic exploration of unsupervised methods for mapping behavior. Physical Biology Systematic exploration of unsupervised methods for mapping behavior 14: pages.
16. Wiltschko, AB, Johnson, MJ, Iurilli, G, Peterson, RE, Katon, JM, Pashkovski, SL, et al. (2015). Mapping sub-second structure in mouse behavior. Neuron Mapping sub-second structure in mouse behavior. 88: pages.
17. Wu, A, Buchanan, E, Whiteway, M, Schartner, M, Meijer, G, Noel, J-P, et al. Deep graph pose: A semi-supervised deep graphical model for improved animal pose tracking2020.
18. Batty, E, Whiteway, M, Saxena, S, Biderman, D, Abe, T, Musall, S, et al. Behavenet: Nonlinear embedding and bayesian neural decoding of behavioral videos. In: H Wallach HLABFdAeBEFRG, editor. Advances in neural information processing systems 32: Curran Associates, Inc.; 2019. p. 15706 - 17.
19. Costacurta, JC, Duncker, L, Sheffer, B, Gillis, W, Weinreb, C, Markowitz, JE, et al. (2022). Distinguishing discrete and continuous behavioral variability using warped autoregressive hmms. NeurIPS Distinguishing discrete and continuous behavioral variability using warped autoregressive hmms pages.

20. Jia, Y, Li, S, Guo, X, Lei, B, Hu, J, Xu, X-H, et al. (2022). Selfee, self-supervised features extraction of animal behaviors. eLife Selfee, self-supervised features extraction of animal behaviors 11: pages.
21. Findley, TM, Wyrick, DG, Cramer, JL, Brown, MA, Holcomb, B, Attey, R, et al. (2021). Sniff-synchronized, gradient-guided olfactory search by freely moving mice. eLife Sniff-synchronized, gradient-guided olfactory search by freely moving mice 10: pages.
22. Berman, GJ, Choi, DM, Bialek, W, Shaevitz, JW (2014). Mapping the stereotyped behaviour of freely moving fruit flies. Journal of the Royal Society, Interface / the Royal Society Mapping the stereotyped behaviour of freely moving fruit flies 11: pages.
23. Markowitz, JE, Gillis, WF, Jay, M, Wood, J, Harris, RW, Cieszkowski, R, et al. (2023). Spontaneous behaviour is structured by reinforcement without explicit reward. Nature Spontaneous behaviour is structured by reinforcement without explicit reward 614: pages.
24. Bohnslav, JP, Wimalasena, NK, Clausing, KJ, Dai, YY, Yarmolinsky, DA, Cruz, T, et al. (2021). Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. eLife Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels 10: pages.
25. Sun, JJ, Karigo, T, Anderson, DJ, Perona, P, Yue, Y, Kennedy, A (2021). Caltech mouse social interactions (calms21) dataset. Caltech mouse social interactions (calms21) dataset pages.
26. Pereira, TD, Aldarondo, DE, Willmore, L, Kislin, M, Wang, SSH, Murthy, M, et al. (2019). Fast animal pose estimation using deep neural networks. Nature Methods Fast animal pose estimation using deep neural networks 16: pages.

Fast Annotation of Rodent Behaviors with AI Assistance: Human Observer and SmartAnnotator Collaborate through Active Learning

E.A. van Dam[1,2], T.J. Daniels[1], L. Ottink[1], M.A.J. van Gerven[2], and L.P.J.J. Noldus[1,2]

[1]Noldus Information Technology BV, Wageningen, The Netherlands. elsbeth.vandam@noldus.com

[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Abstract

AI-assisted behavior annotation saves time compared to manual annotation. Although automated systems for rodent behavior annotation exist, specific behaviors are still scored by hand, as results are not equally accurate across behaviors. With active learning, we can reduce the annotation effort and tailor the result towards the needs within a research experiment. We present the benefit of active learning on two particular and ambiguous behaviors: 'stretched attend' and 'unsupported rearing'.

Introduction

Automated annotation of rodent behavior from video recordings is an essential tool for behavioral research. It speeds up and improves rodent behavior analysis with more readily available and consistent behavior annotations. Since early 2013, commercially available solutions as well as open-source projects exist for a specific set of behaviors [1,2,3,4]. Although great strides have been made and the most common rodent behaviors can be detected in video streams under specific recording conditions, many ambiguous or rare behaviors that are also relevant in behavioral research (such as epileptic seizures, stereotypic variants of behaviors, whisking), or for which a different definition is used, are still scored by hand.

Developing generic, robust automatic solutions is costly since it requires a large set of precise and consistently labeled video footage that contains the same variation as the variation in the footage that the solution will be applied to [5]. This refers not only to the appearance of the animals and environment (fur color, cage, lighting), but also to the way the behaviors are conducted. The speed of walking, the length of a grooming session, the height of a rearing and angle to the camera, the behaviors before and after a scratching event etc. vary with for instance age, time of day, mood, motor skills and drug treatment. An automated classifier can only reliably recognize what has been seen in the training data, so all variations in the deployment data set must occur in the training set as well. If this is not the case, the classifier will suffer from selection bias. To collect and label a sufficient amount of such training data is difficult to achieve, especially for rare and subtle behaviors, or for behaviors that are very specific to the research question at hand.

To meet the demand for faster annotation of behaviors for which there is no generic solution available, and to tackle the challenges in behavior classification, we developed a novel AI-assisted annotation tool, SmartAnnotator (Figure 1). This tool helps the researcher to annotate behaviors, by training a classifier through active learning. Here, active learning refers to the ability of an AI system to interactively query a human user to label new data points to maximally improve learning performance. Instead of playing a video from start to end and scoring behavior by hand, the researcher is presented short video clips to annotate. Simultaneously, a classifier is trained in the background on these annotations, and infers behaviors on unlabeled video clips, until the entire video is annotated. Importantly, the SmartAnnotator selects video clips that were given an uncertain label by the model and asks the user to label these clips. In other words, the SmartAnnotator selects events whose annotation will maximally improve behavior classification. This exploits both human expertise and AI to increase annotation accuracy. This interactive approach is much more efficient than labeling all data from start to end and avoids observer drift, as well as unavoidable decision delay in manual scoring. Hence, it reduces annotation time while increasing the quality of the labels. Furthermore, unlike previously described tools for interactive behavior annotation [6,7], SmartAnnotator is cloud-based, so annotation can be done in any web browser and resources are scalable.

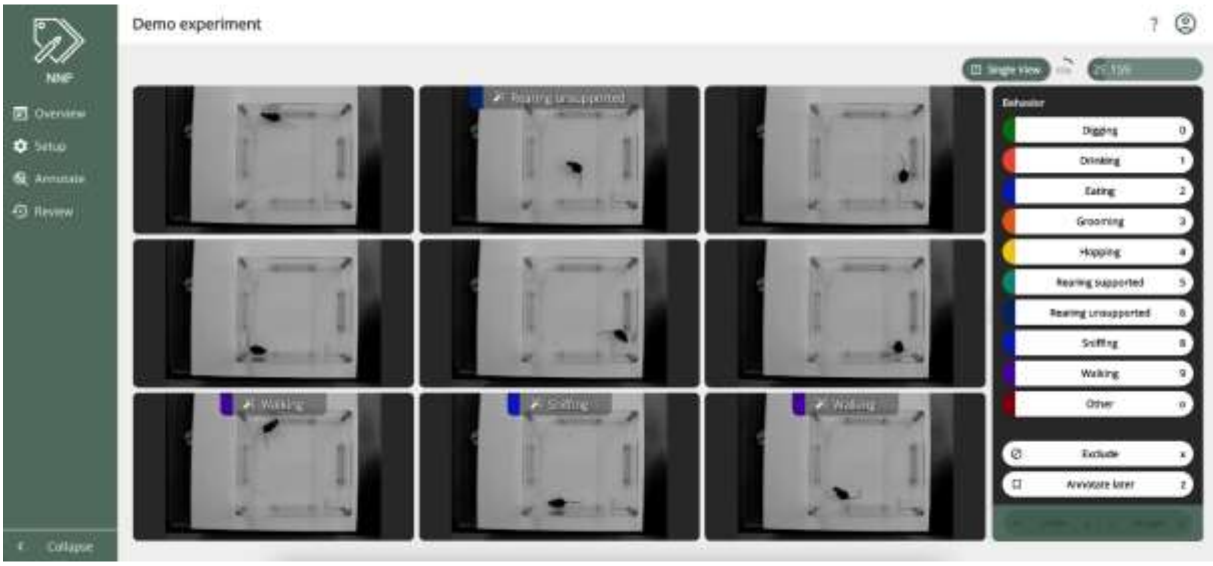


Figure 6. A screen capture of the SmartAnnotator tool. Multiple selected clips are presented to the user for labeling.

AI-assisted annotation, nevertheless, also has its challenges. It relies on the relevance and quality of the features that it can extract from the videos, the segmentation of the data into clips, and on how reliably the tool can detect similarities in the data. Furthermore, the learning and processing needs to be fast, since it is an interactive process that must be user-friendly. Yet the advantage over generically trained classifiers that are deployed out-of-the-box is that with AI-assisted annotation, a dedicated classifier can be trained on a specific set of features with a specific interpretation of the behavior by the user. Also, the algorithm does not need to account for unseen variance since it works on the entire set of videos from one behavioral experiment at once.

Recent work [8,9] on automated behavior detection furthermore highlights the use of behavioral clusters, i.e. short pieces of behavior of a certain type that are learned by clustering a latent representation (embedding) of the data, learned by a self-supervised auto-encoder. It has also been shown [10] that there is not always a straightforward one-to-one relation between such clusters and specific behaviors of interest. Therefore, we have not applied classification on the embedding at this moment but used the behavioral clusters only for segmentation of our data.

We have used SmartAnnotator to annotate specific target behaviors in a set of videos: ‘stretched attend’ in a mouse dataset, and ‘unsupported rearing’ in a rat data set. Especially ‘unsupported rearing’ is difficult to detect automatically from videos since event durations are typically short and contain a relatively large amount of transitional postures that are equal to postures of other behaviors [10]. This can easily lead to a large number of false positive detections. We show that our method improves annotation accuracy and reduces the amount of data that needs manual labeling compared to a supervised classification of these behaviors. This demonstrates the benefit of active learning and our SmartAnnotator tool in finding particular or rare behaviors in a larger dataset, that would otherwise easily have been misclassified.

Methods

Annotation through active learning using SmartAnnotator

Using SmartAnnotator, instead of playing a video from start to end and simultaneously scoring behavior, the researcher is presented short video clips to annotate, based on their similarity in the videos. The similarity is derived from low-level behavior features that are precalculated from the experiment videos by EthoVision XT[1]. While the researcher is labeling clips, a classifier is trained on these annotations, infers the labels of similar clips, and clips with low certainty are presented to the researcher to ensure high annotation accuracy. Once all clips are labeled, the researcher can inspect the clips and edit the labels if necessary.

[1] www.noldus.com/ethovision

The low-level behavior features were designed for automatic mouse and rat behavior recognition. They are described in [1] and combine spatial body shape features, movement features, multi-scale temporal window features, and environment proximity features based on location. Together these features form a low-level behavior profile over time, independent of species, gender, age and appearance. The features have proven their richness and robustness in the Mouse and Rat Behavior Recognition modules in EthoVision XT. Also, these features were successfully used as input for a scratch behavior classifier [2].

In order to create an AI-assisted annotation, the following steps are performed: 1) Cluster experiment data (low-level behavior features) and use these for data segmentation to create events in the data, 2) Let the human label *n* events, and 3) Train a behavior classifier on all the labeled data available. We first transformed the data using MiniRocket [11], and then used a linear classification model with one fully connected layer. Steps 2 and 3 are repeated until a user-defined accuracy threshold is reached. 4) Apply the best classifier resulting from tuning to all unlabeled data, and label the certain events, 5) Pick *n* uncertain events and present them to the human for labeling. Steps 3-5 are repeated until all data are labeled. For the experiments in this study, we made use of a look-up table (so called oracle) with access to ground truth labels, as a replacement of the human labeler.

Annotation using supervised classification

To address the benefits of an active learning approach using our SmartAnnotator tool in annotation of ‘stretched attend’ and ‘unsupported rearing’, we compared it to supervised classification of those behaviors. To this end, we trained a classifier consisting of two main parts: first, a variational autoencoder of three 1D-CNN layers to learn an embedding of the features that can be used to reconstruct the input, and second a classification head of two linear layers to estimate the behaviors from the embedding.

For supervised classification, we used the same datasets as for the active learning approach. For the mouse dataset we made a training split including 4 out of 5 videos and a validation split with the remaining data. For the rat dataset we made a training split including 12 out of 14 videos and a validation split including data of the remaining 2 videos.

In both the active learning scenario and the supervised scenario, we optimized the classification towards high recall at the cost of low precision, because in post-processing it is easier to correct for false positives than for false negatives.

Datasets and behaviors

The mouse dataset that we used for annotation of ‘stretched attend’ consists of 5 × 5 minutes of video with annotated behaviors ‘stretched attend’ (121 events), ‘walk’ (110) and ‘other’ (172). Since we focused on the annotation of ‘stretched attend’ (Figure 2A), the ‘walk’ behavior events were also considered as ‘other’ (resulting in 282 events for ‘other’). The recordings were made for other purposes by Utrecht University and were given to us with permission to use for our research.

The rat dataset that we used for annotation of ‘unsupported rearing’ consists of 14 × 5 minutes of video, with annotated behaviors ‘unsupported rearing’, ‘drink’, ‘eat’, ‘groom’, ‘jump’, ‘rearing supported’, ‘rest’, ‘sniff’, ‘walk’, and ‘other’. Since we focused on the annotation of ‘unsupported rearing’ (Figure 2B), the rest of the behavior events were also considered as ‘other’, resulting in 31 events for ‘unsupported rearing’ and 912 events for ‘other’. The dataset was reused from previous work and described in [1].

With respect to ethical permissions, we remark that no animals were handled for his work.

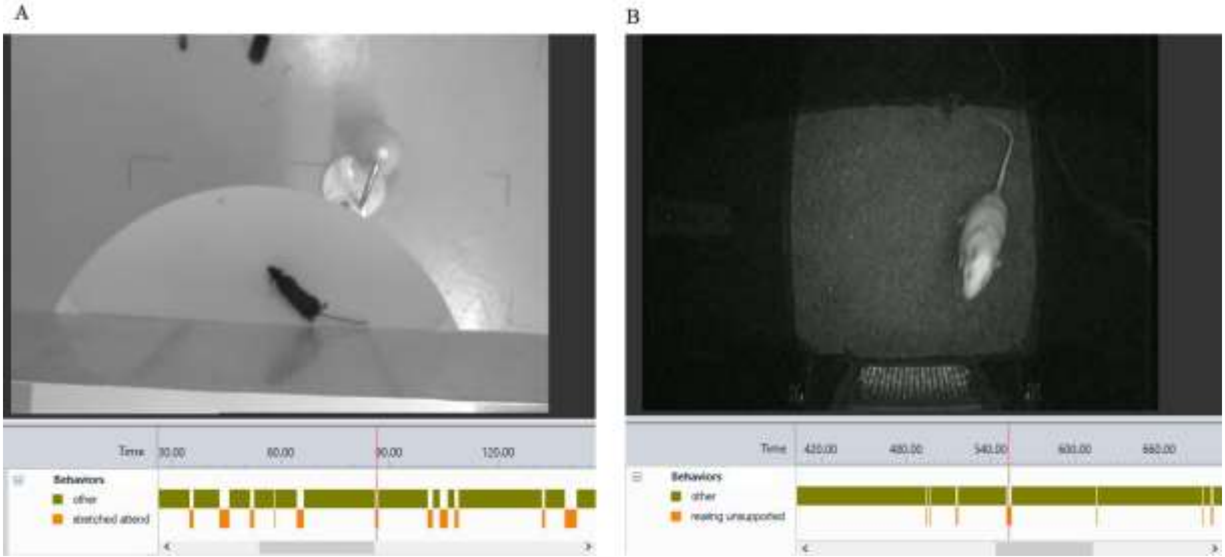


Figure 7. A screen capture of one of the used ‘stretched attend’ videos with event logs. **B**. A screen capture of one of the used ‘unsupported rearing’ videos.

Results

We evaluated the benefit of active learning using SmartAnnotator with two examples of specific behaviors that are easily misclassified by generic automatic tools: ‘stretched attend’ and ‘unsupported rearing’. We analyzed precision, recall and F1-scores of the annotation using the active learning approach as well as the supervised classification approach (Table 1). We performed a Wilcoxon rank sum test to test for differences between the active learning and supervised approach.

The active learning approach results in similar recall compared to supervised classification, for both ‘stretched attend’ ($p = 0.104$) and ‘unsupported rearing’ ($p = 0.762$; Table 1). The precision, however, is higher in the active learning approach ($p < 0.001$ for both behaviors), and with that also the F1-scores ($p < 0.001$ for both behaviors; Table 1). The standard deviation, however, of precision is quite high, especially for ‘unsupported rearing’. One explanation might be the different training data between runs because of the setup of the active learning process. Additionally, the active learning approach requires much less manually labeled data compared to supervised classification. For ‘stretched attend’, 18.3% manual labels was required for the active learning method (75 out of 403 events), as opposed to the 85.1% (343 out of 403 events) we used for the supervised method, and for ‘unsupported rearing’, 25.4% (239 out of 943 events) manual labels was required for active learning while we used 75.1% (708 out of 943 events) manual labels for the supervised method (Table 1). Overall, these results indicate that the active learning method using SmartAnnotator recalls a high number of specific behavior instances, while also reducing the number of false positives (reflected in the higher precision) compared to supervised classification of these two behaviors.

This is also reflected in Figure 3, where we plotted the annotation of the target behaviors by the active learning approach across the dataset, compared to ground truth. Via visual inspection we can see that the pattern over time is recovered well for both ‘stretched attend’ and ‘unsupported rearing’, and that most of the events have been retrieved.

Table 1. Results of the active learning and supervised approach in classifying ‘stretched attend’ and ‘unsupported rearing’. Reported are the true number of events of the behavior, the total number of events in the dataset, the mean percentage of manual labels (events labeled by the human) required, and precision, recall and F1-scores. The results are the mean of 10 runs for each behavior and for both methods. * $p < 0.001$ for the comparison between the active learning and the supervised approach.

	classification method	*n* events behavior	*n* events total	% manual labels	precision (mean ± std)	recall (mean ± std)	f1-score (mean ± std)
Stretched attend	Active learning	121	403	**18.3**	**0.73 (± 0.13)***	**0.88 (± 0.08)**	**0.79 (± 0.09)***
	Supervised	121	403	85.1	0.44 (± 0.04)	0.84 (± 0.03)	0.57 (± 0.04)
Unsupported rearing	Active learning	31	943	**25.4**	**0.41 (± 0.32)***	0.75 (± 0.05)	**0.46 (± 0.27)***
	Supervised	31	943	75.1	0.04 (± 0.01)	0.75 (± 0.10)	0.08 (± 0.01)

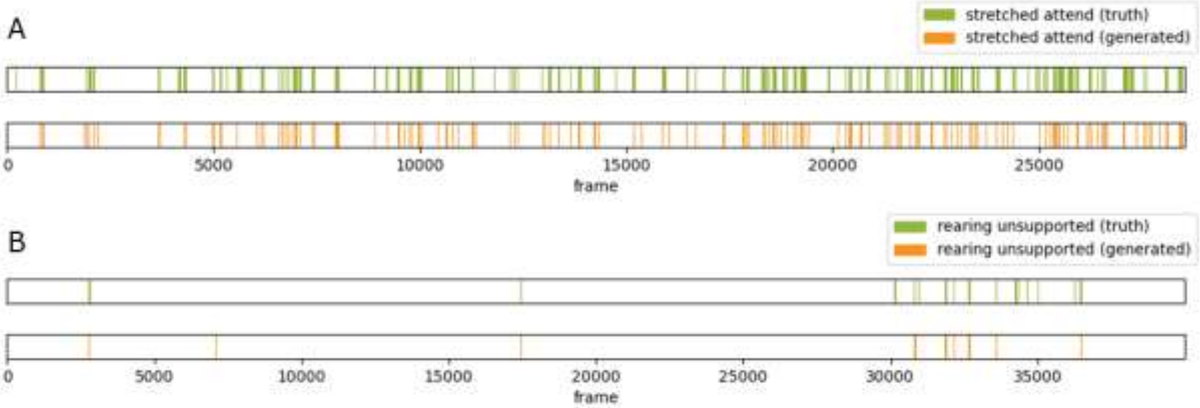


Figure 3. The generated (predicted) annotations using the active learning approach, compared to ground truth, of the run that resulted in statistics closest to the mean (Table 1). Annotations of videos in the dataset are appended. **A**. Generated annotation of ‘stretched attend’ (precision = 0.74, recall = 0.9, f1 = 0.81). **B**. Generated annotation of ‘unsupported rearing’ (precision = 0.44, recall = 0.77, f1 = 0.56). We plotted the first 40000 frames (out of 95750), as in that portion most of the ‘unsupported rearing’ events occur.

Discussion

In this study, we demonstrate the benefit of an active learning approach of our SmartAnnotator tool in annotating specific or rare behaviors that are otherwise easily misclassified: ‘stretched attend’ and ‘unsupported rearing’. The results in this study indicate that SmartAnnotator increases annotation accuracy and reduces the number of required manual annotations which are otherwise labor-intensive and thereby decreases annotation time.

While automatic annotation was already available for the most commonly observed rodent behaviors, more ambiguous or rare behaviors are still scored by hand. We demonstrate the advantage of active learning in annotating such behaviors. These results are also promising considering annotation of specific behaviors that are not commonly observed and have a clear definition. By letting the model pick uncertain events and ask the human for a label of these events, the model can be specifically trained to recognize the target behavior, even if the behavior only rarely occurs in the dataset. In such situations, we argue that the interactive way of annotating behaviors using SmartAnnotator is useful, as the algorithm can automatically annotate a large portion of the data but ask feedback from the human about parts where it is uncertain, and thereby ask for examples of a difficult behavior.

In the current study, we present results for ‘stretched attend’ and ‘unsupported rearing’. Considering that these two behaviors are difficult to automatically classify, our annotations using the active learning approach are quite accurate. They might be, however, not as precise as could be desired, especially for less difficult behavior types. For instance, for a behavior like ‘unsupported rearing’, which is difficult to recognize automatically, the model needs a relatively high number of manual labels, as is reflected in the percentage of manual labels (Table 1). We expect that for less difficult behavior types, the tool will need fewer manual labels to reach an accurate annotation, and furthermore yield higher precision and recall. Besides, the tool can be used to annotate multiple behaviors in the same dataset.

In our current active learning approach, event segmentation is based on framewise clustering of the temporal features from EthoVision XT. This leads to far more events than would be annotated by a human observer. One of the potential improvements that is to be explored is to combine cluster traversals into coarser events, to investigate whether this could increase annotation accuracy even further.

In short, the results in the current study demonstrate the benefit of applying an active learning approach to classify ambiguous or rare rodent behaviors. These are easily misclassified by conventional automatic classification and are currently still often scored by hand. Tools like SmartAnnotator increase annotation accuracy of such difficult behaviors and reduce annotation time. Through the development of this AI-assisted approach we contribute a hybrid AI solution, where combining the skills of humans and machines yields better performance than using one of both.

Acknowledgements

This publication is part of the project Dutch Brain Interface Initiative (DBI2) with project number 024.005.022 of the research program Gravitation which is (partly) financed by the Dutch Research Council (NWO).

References

1. Dam, E.A. van, Harst, J.E. van der, Braak, C.J.F. ter, Tegelenbosch, R.A.J., Spruijt, B.M., & Noldus, L.P.J.J. (2013). An automated system for the recognition of various specific rat behaviours. *Journal of Neuroscience Methods*, **218**, 214–224.
2. Dam, E.A. van, Roosken, M.H., & Noldus, L.P.J.J. (2022). Robust scratching behavior detection in mice from generic features and a lightweight neural network in 100 fps videos. *Volume 2 of the Proceedings of the Joint Meeting of Measuring Behavior 2022, the 12th International Conference on Methods and Techniques in Behavioral Research, and the 6th Seminar on Behavioral Methods* (Held Online,18-20 May 2022), 5.
3. Isik, S., & Unal, G. (2023). Open-source software for automated rodent behavioral analysis. *Frontiers in Neuroscience*, 17, 1149027.
4. Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., & Kennedy, A. (2020). The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *ELife* **10**:e63720.
5. Dam, E.A. van, Noldus, L.P.J.J., & Gerven, M.A.J. van. (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of Neuroscience Methods*, **332**, 108536.
6. Kabra, M., Robie, A.A., Rivera-Alba, M., Branson, S., & Branson, K. (2013). JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature Methods*, **10,** 64-67.
7. Lorbach, M., Poppe, R., & Veltkamp, R.C. (2019). Interactive rodent behavior annotation in video using active learning. *Multimedia Tools and Applications*, **78**, 19787–19806.
8. Luxem, K., Mocellin, P., Fuhrmann, F., Kürsch, J., Miller, S.R., Palop, J.J., Remy, S., & Bauer, P. (2022). Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology*, **5**(1), 1267.
9. Weinreb, C., Osman, M.A.M., Zhang, L., Lin, S., Pearl, J., Annapragada, S., Conlin, E., Gillis, W.F., Jay, M., Shaokai, Y., Mathis, A., Mathis, M.W., Pereira, T., Linderman, S.W., & Datta, S.R. (2023). Keypoint-MoSeq: Parsing behavior by linking point tracking to pose dynamics (p. 2023.03.16.532307). *BioRxiv preprint*.
10. Dam, E.A. van, Noldus, L.P.J.J., & Gerven, M.A.J. van (2023). Disentangling rodent behaviors to improve automated behavior recognition. *Frontiers in Neuroscience*, **17,** 1198209.
11. Dempster, A., Schmidt, D.F., & Webb, G.I. (2021). MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (New York, USA, 2021), 248–257.

End-to-end behavior annotation pipeline for mouse behavior annotation

Glen Beane, Brian Q. Geuther, Thomas J. Sproule, Anshul Choudhary, Jarek Trapszo, Leinani Hession, Vivek Kohar, Vivek Kumar

The Jackson Laboratory, 600 Main Street, Bar Harbor ME 04609.

vivek.kumar@jax.org

Abstract

Automated detection of complex animal behavior remains a challenge in neuroscience. Developments in computer-vision have greatly advanced automated behavior detection and allow high-throughput pre-clinical studies. An integrated hardware and software solution is necessary to facilitate the adoption of these advances in the field of behavioral neurogenetics, particularly for non-computational labs. We have published a series of papers using an open field arena to annotate complex behaviors such as grooming, posture, and gait as well as higher-level constructs such as biological age and pain. Here, we present our, integrated rodent phenotyping platform, JAX Animal Behavior System (JABS), to the community for data acquisition, machine learning-based behavior annotation and classification, classifier sharing, and genetic analysis. The JABS Data Acquisition Module (JABS-DA) enables uniform data collection with its combination of 3D hardware designs and software for real-time monitoring and video data collection. JABS-Active Learning Module (JABS-AL) allows behavior annotation, classifier training, and validation. It supports the execution of both inference and downstream genetic analyses (heritability and genetic correlation) on a curated, publicly accessible dataset across 160 mouse strains. This enables the use of genetics as a guide to proper behavior classifier selection. This open-source tool is an ecosystem that allows the neuroscience and genetics community for shared advanced behavior analysis and reduces the barrier to entry into this new field.

Results and Discussion

Although major advances have been made in behavior annotation using computer vision, a major challenge exists in the democratization of this technology [1], [2], [3], [4]. Pose estimation through keypoint tracking has been an important advance in the field of animal behavior annotation. However, keypoint tracking is one of the first steps in automated behavior annotation. Currently, a high level of expertise is needed for efficient use of these methods. For instance, large amounts of training data are needed to train a pose estimation or grooming detection network. Many labs collect data in disparate ways which lack standardized visual appearance. In order to apply existing models for detecting pose or behavior, each individual lab must train their own model at high cost of labeled data. This is a challenge to the field - many useful models exist, but they cannot be directly applied across data originating from different labs. If visual input is standardized across labs, trained models can be adopted across labs, decreasing barriers to entry, and increasing reproducibility of data.

With this in mind, we describe our data acquisition and behavior annotation system here. We have developed an integrated mouse phenotyping platform called JAX Animal Behavior System (JABS), which consists of video collection hardware and software, a behavior labeling and active learning app, and an online database for sharing classifiers. Furthermore, JABS is populated with data from over 150 mouse strains and once a behavior classifier is deposited, heritability and GWAS results can be generated. Thus, genetics is integrated into JABS as a core feature. I will describe each component of JABS in my talk in detail.

Briefly, I describe these below.

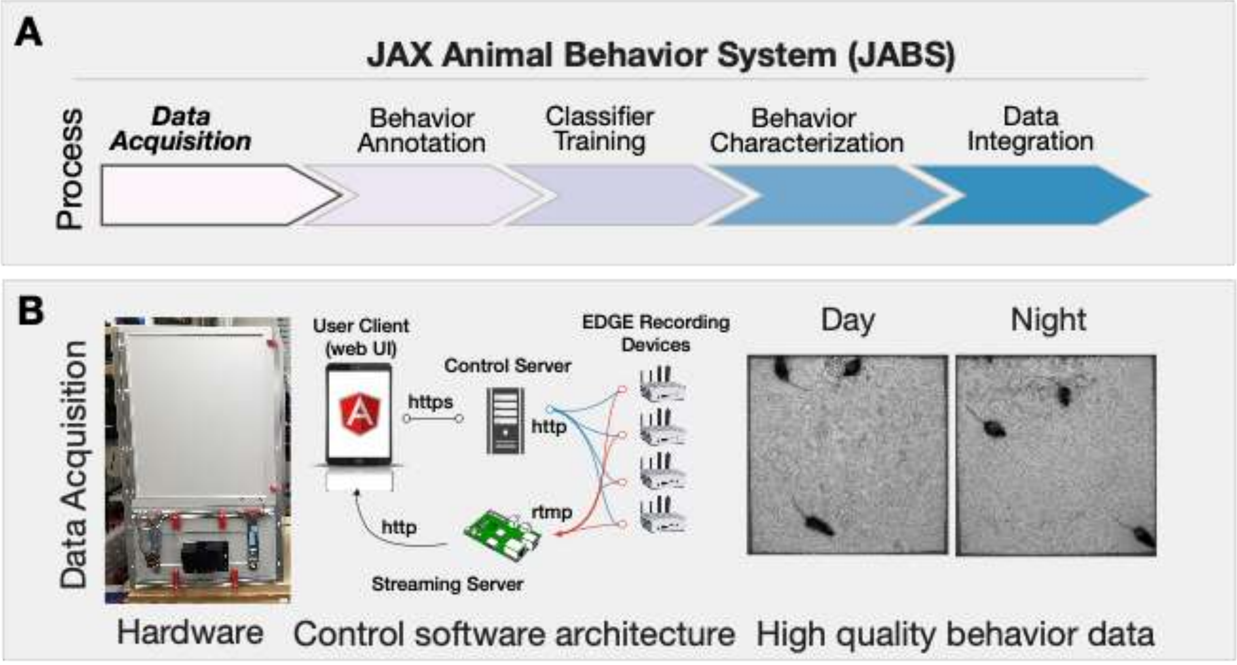


Figure 8: JABS overview and hardware

Figure 1A describes the various components of JABS. The data acquisition hardware is fully open source for academic use (Figure 1B). It includes hardware design files and software for data acquisition. Sample data are shown in B. The hardware has met strict JAX criteria for animal husbandry and is fully IACUC approved.

JABS-AL

In order to annotate behavior we provide models for keypoint tracking that has been validated and works across 62 mouse strains [5]. This alleviates the need for any labeling or model training by the user as long as the video data is collected in accordance with JABS specs.

After tracking, we have built an active learning system that allows behaviorists to label and build behavior classifiers (Figure 2). The system is easy to use for non-computational researchers and provides precision, recall, and F1 statistics to gauge classifier performance. We have validated performance of JABS-AL using a published hand annotated grooming dataset [6]. We find that JABS-AL achieves comparable performance using features from keypoints as classification from raw videos. Importantly, it achieves this performance with much less training data. We also provide guidance on how to postprocess results for improved performance.

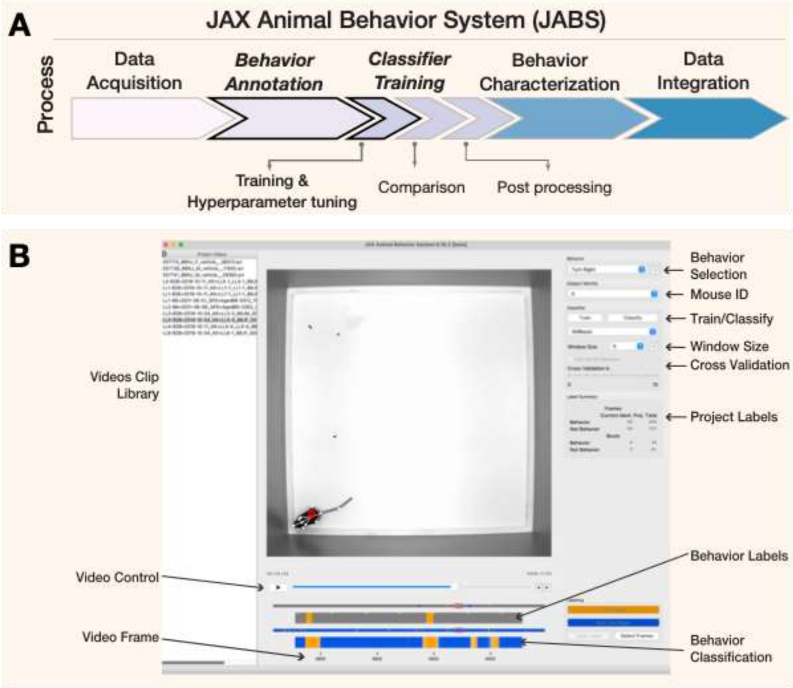


Figure 9: JABS-AL, an active learning module for behavior classification.

JABS-DB

Finally, we provide a JABS-Database, a webapp for sharing classifiers and carrying out genetic studies (Figure 3). Users can upload their behavior classifiers or download an existing classifier. This allows the community to start build upon each other's work and create replicable results. Underlying JABS-DB are features from over 150 mouse strains that enables the user to determine genetic correlation of their classifier with existing classifiers deposited into JABS-DB. GWAS can also be carried out for the behavior of interest.

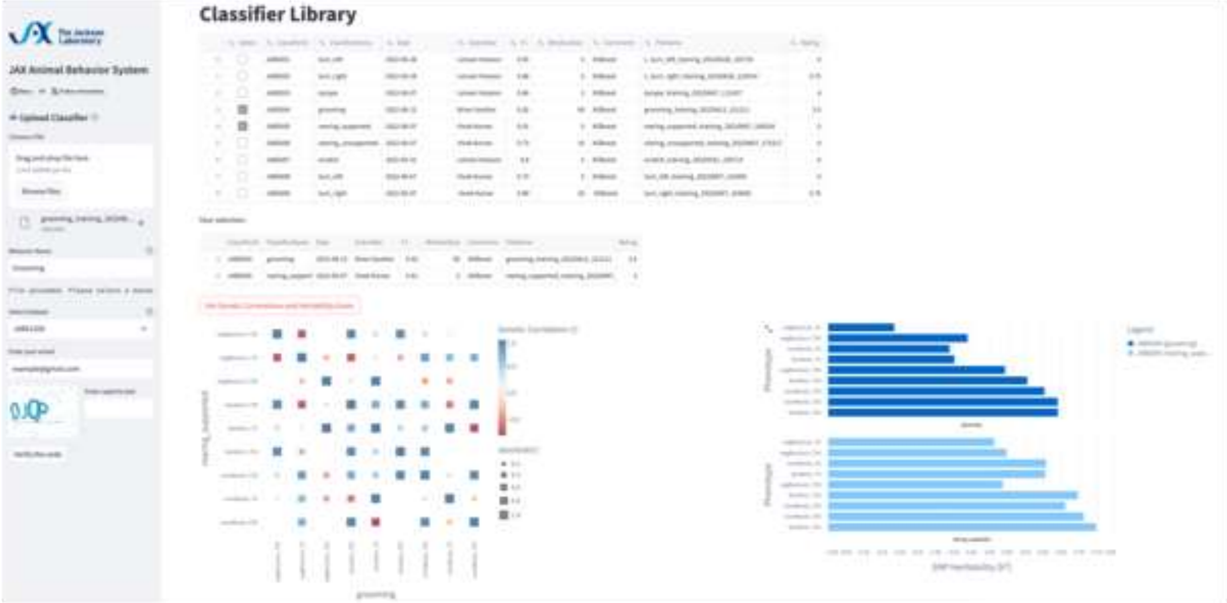


Figure 10: JABS-DB a database for sharing classifiers and for carrying out genetic analysis.

Conclusion

We present an integrated end-to-end system for behavior annotation. By adopting a common hardware and data acquisition system, the downstream ML tasks become easier with no need for keypoint tracking or feature generation. More importantly, classifiers across labs can be shared allowing labs to share expertise.

Internally, we have used JABS for tracking [7], action detection [6], gait and posture [5], frailty [8], nociception [9], and even mouse mass assessment [10]. All these behavior annotation models become accessible for labs that use JABS.

References

1. T. D. Pereira, J. W. Shaevitz, and M. Murthy, "Quantifying behavior to understand the brain," *Nat. Neurosci.*, vol. 23, no. 12, Art. no. 12, Dec. 2020, doi: 10.1038/s41593-020-00734-z.
2. M. Raghu and E. Schmidt, "A Survey of Deep Learning for Scientific Discovery," *ArXiv200311755 Cs Stat*, Mar. 2020, Accessed: Mar. 13, 2021. [Online]. Available: http://arxiv.org/abs/2003.11755
3. M. W. Mathis and A. Mathis, "Deep learning tools for the measurement of animal behavior in neuroscience," *Curr. Opin. Neurobiol.*, vol. 60, pp. 1–11, Feb. 2020, doi: 10.1016/j.conb.2019.10.008.
4. J. D. Choi and V. Kumar, "A new era in quantification of animal social behaviors," *Neurosci. Biobehav. Rev.*, vol. 157, p. 105528, Dec. 2023, doi: 10.1016/j.neubiorev.2023.105528.
5. K. Sheppard *et al.*, "Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation," *Cell Rep.*, vol. 38, no. 2, p. 110231, Jan. 2022, doi: 10.1016/j.celrep.2021.110231.
6. B. Q. Geuther, A. Peer, H. He, G. Sabnis, V. M. Philip, and V. Kumar, "Action detection using a neural network elucidates the genetics of mouse grooming behavior," *eLife*, vol. 10, p. e63207, Mar. 2021, doi: 10.7554/eLife.63207.
7. B. Q. Geuther *et al.*, "Robust mouse tracking in complex environments using neural networks," *Commun. Biol.*, vol. 2, p. 124, 2019, doi: 10.1038/s42003-019-0362-1.
8. L. E. Hession, G. S. Sabnis, G. A. Churchill, and V. Kumar, "A machine-vision-based frailty index for mice," *Nat. Aging*, vol. 2, no. 8, pp. 756–766, Aug. 2022, doi: 10.1038/s43587-022-00266-0.
9. G. S. Sabnis, L. E. Hession, K. Kim, J. A. Beierle, and V. Kumar, "A high-throughput machine vision-based univariate scale for pain and analgesia in mice," *bioRxiv*, p. 2022.12.29.522204, Jan. 2023, doi: 10.1101/2022.12.29.522204.
10. M. Guzman, B. Geuther, G. Sabnis, and V. Kumar, "Highly Accurate and Precise Determination of Mouse Mass Using Computer Vision." bioRxiv, p. 2023.12.30.573718, Dec. 30, 2023. doi: 10.1101/2023.12.30.573718.

DeepRod: A human-in-the-loop system for automatic rodent behavior analysis

A. Loy[1], M. Garafolj[1], H. Schauerte[2], H. Behnke[1], C. Charnier[2], P. Schwarz[3], G. Rast[2] and T. Wollmann[1]

[1] Merantix Momentum GmbH

[2] Boehringer Ingelheim GmbH & Co. KG, Drug Discovery Sciences

[3] BI X GmbH

Introduction

In drug discovery and development, systematic assessment of drug safety using highly regulated preclinical studies prior to first-in-human clinical trials are mandatory [1] to ensure safety for volunteers and patients. These assessments allow detailed views on risk, benefit and the therapeutic index of potential future therapeutics. The evaluations include, among others, standardized functional behavioral studies in rodents [2, 3, 4]. During the research phase, automated video-based systems (e.g., PhenoTyper) are used to assess continuous quantitative and qualitative motor behavior during the active phase of the rodents. Infrared video cameras located in the top unit of each observation arena populated with one rodent per arena record 14 h of video material per animal. During each study, groups of animals exposed to an active ingredient at various doses or receiving a placebo are recorded from the top, generating large video datasets. Events of interest can be very rare and require in depth analysis of the footage. Manual analysis is not feasible in an appropriate time and with acceptable effort.

Automated analysis of distance moved and animal velocity (e.g. EthoVision XT) provides very sensitive measures for central nervous effects and general tolerability. However, these features are not discriminative enough to detect complex events.

In this work, we propose an UX-optimized platform for behavior labeling and analysis integrated into the workflow, AI-based complex behavior prediction, active learning to find rare events and propose candidates for new behavioral categories.

Methods

We propose a novel system for analyzing rodent behavior at scale, which combines a user-centered interface with AI-based behavior prediction, novel behavior recognition and active learning. The system should support rodent behavioral analysis by automatizing the behavior classification of rodents. To achieve this, the system needs to enable users in annotating the rodent behavior in video snippets to gather training data, support users in detecting novel, previously unseen behaviors and finally to automatically classify the behavior of rodents.

System overview

Each routine study performed to profile a research substance includes 35 million frames from typically 28 individual 14h long video clips with 25 fps, recorded by a PhenoTyper camera [5, 6]. To tackle this amount of data, an efficient, parallelized, and cloud-native data processing pipeline processes these raw video files [7]. First, the pipeline re-encodes the input video for efficiency and storage and registers the video's metadata into the system. Second, visual information is extracted using a deep learning approach [8] and meaningful features are further generated using that information. Lastly, a classifier predicts the rodent behavior for each frame based on those features.

Behavior recognition

The core component of our system is the behavior recognition component. This component uses a two-staged machine learning pipeline to classify the behavior of rodents. The first stage extracts visual information from the video stream by localizing nine anatomical landmarks ("keypoints" in the following) of the rodent. Similar to

MARS [9], keypoints correspond to the nose, ears, body center, hips, tail base, tail center and tail end. Our keypoint extraction method is based on DeepLabCut [8].

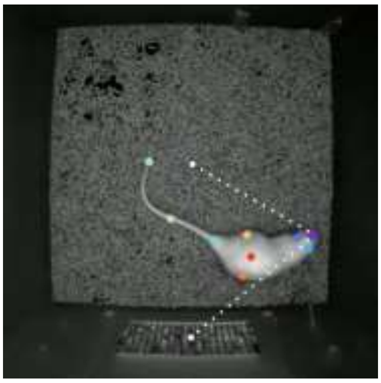


Figure 1.1. Features based on the distances between keypoints are relevant to detect interactions.

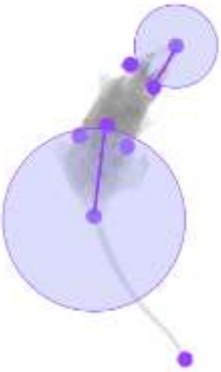


Figure 1.2. Features based on [10] use the relative keypoint positions are relevant to identify typical postures.

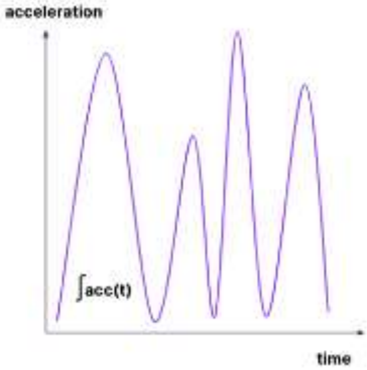


Figure 1.3. Features based on [11] use the keypoint movements to identify temporal patterns like directed or undirected motion.

The second stage is a classifier that uses features based on the keypoints. We identified that features capturing the position, pose, and movement are discriminative for characterizing the rodent's behavior. Per each feature category, we engineered a range of features based on the keypoints capturing the rodent's position, pose and movement (Figure 1). Some features are aggregated within sliding windows of various sizes. We frame the behavior detection problem as a multi-class classification problem, containing all known behaviors and an extra class representing any unknown behavior, which is explicitly labeled to not conform to any of the known behavior. Our method is leveraging XGBoost [12], which is recognized as a good classifier under skewed data and noise [13]. The tree-based model also enables interpretability like computing feature importance, which is favorable in life sciences [14]. The system trains new models automatically based on user request. Users receive a report after training that offers an intuitive overview of model improvements.

Detection and labeling of rare behavior

To enable automatic rodent behavior classification, collection of annotated data is necessary to train the machine-learning-based classifier. As there are thousands of hours of video material that can be used to create the annotated training data set, the choice of which video sections to annotate is not trivial and is subjected to time constraints of the labeling force.

For efficient use of human labeling resources, the system implements a labeling assistant shown in Figure 2. The labeling assistant leverages an active learning method based on Meal [15] to suggest areas to label across the whole video material. The active learning methodology can be formulated as an SQL query, which selects a fixed amount k of currently unlabeled model predictions from the database. We refer to the result of this query for a given video as the "labeling queue". The active learning query consists of multiple subqueries, which each have a capped contribution to the overall labeling queue. The subqueries select samples with the following properties: Sample is likely to belong to an underrepresented class, has high prediction uncertainty, has high novelty score, is selected randomly. To find rare behaviors, we select high likelihood samples showing rare behavior, despite the rare behavior might not be currently predicted. Prediction uncertainty is estimated by observing high variance class probabilities over time. Including random samples avoids selection bias.

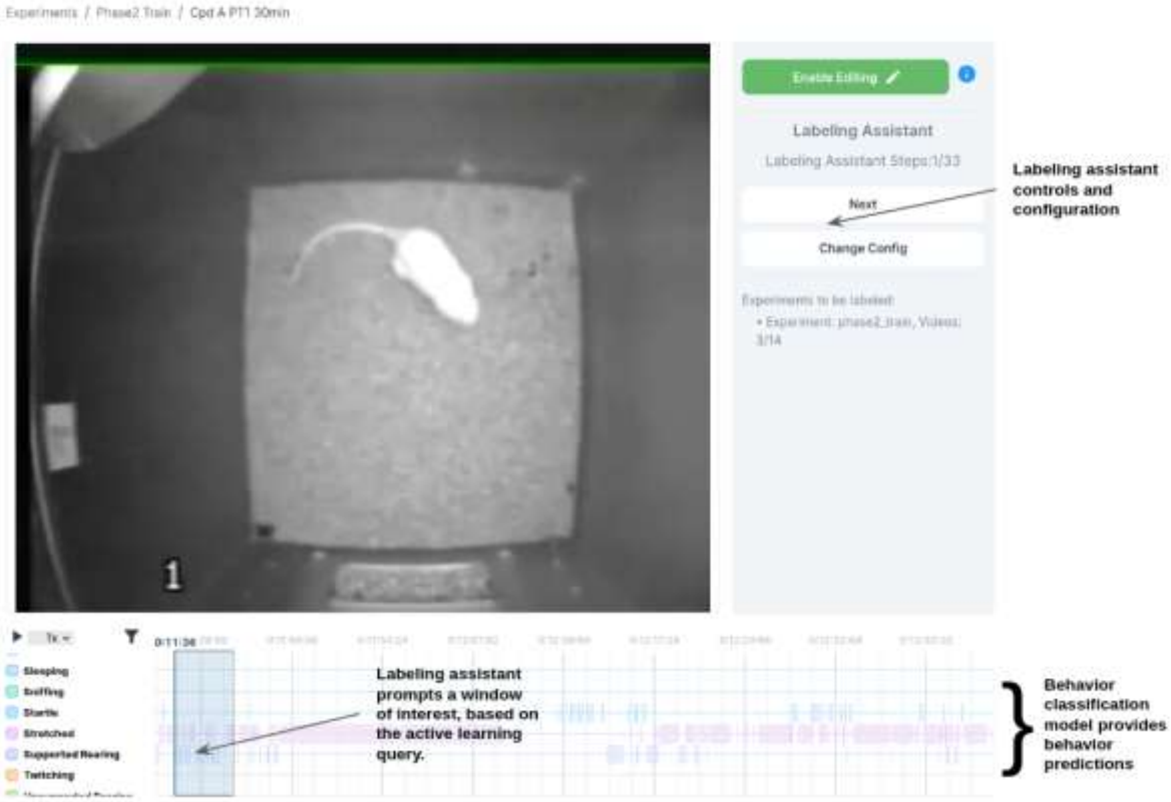

Experiments / Phase2 Train / Cpd A PT1 30min
Enable Editing
Labeling Assistant
Labeling Assistant Steps:1/33
Next
Change Config
Experiments to be labeled:
• Experiment: phase2_train, Videos: 3/14
1
Labeling assistant controls and configuration
1x
0:11:36
Sleeping
Sniffing
Startle
Stretched
Supported Rearing
Twitching
Labeling assistant prompts a window of interest, based on the active learning query.
Behavior classification model provides behavior predictions

Figure 2. Main view of annotation view with the labeling assistant. Labelers are presented with the video footage and the timeline of model-predicted behaviors as well as annotations that are already set. Annotations can be created with a single click to allow for an efficient process.

Each item in the labeling queue corresponds to a window of interest, which users are expected to annotate. The order of the queue is randomized to prevent bias of which sections in the video get labeled. Once the queue is exhausted, labelers are directed to a new video with a fresh queue and therefore iteratively cycle through all videos. This setup prevents annotations concentrated in only a small subset of videos and ensures that each video acquires labels for the top *k* most relevant sections.

Novel behavior recognition

The rodents in the experiments might demonstrate unusual or novel behaviors due to the effect of the compounds that they are exposed to. Therefore, the system needs to support users in observing unusual behavior to enable them to possibly categorize it as a new behavior class. We refer to this problem as “novel behavior recognition” and formulate it as an outlier detection problem. Each frame gets embedded into a low-dimensional feature space using principal component analysis (PCA) to reduce the dimensionality and the correlation between features. Then, mean and variance across all labeled samples for each class (i.e. each defined and annotated behavior type) are computed. With that, we define the *novelty score* of a frame as the *Mahalanobi's distance* [16] to the closest known distribution.

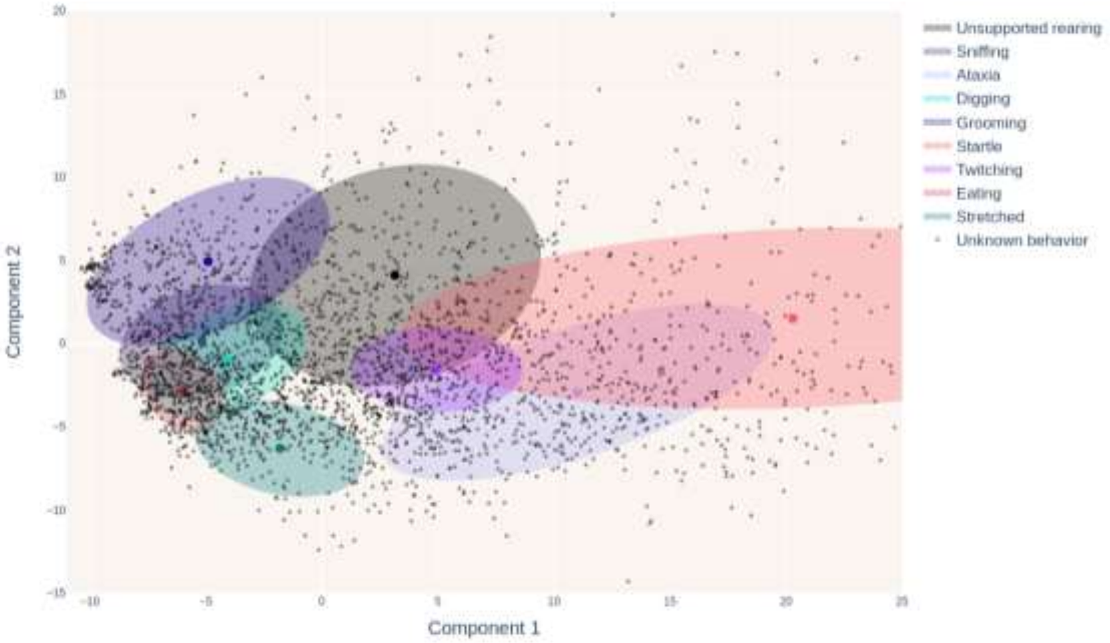


Figure 3. Visualization of the lower dimension feature embedding space used for novelty scoring of a sample. We visualize the center and variance of the estimated class distributions. Samples are scored according to their distance to the distributions of labeled samples. Note that the figure contains a subset of randomly sampled 3000 unlabelled behavior points.

Results

The system was evaluated in a pilot with three distinct expert annotators who created 13.862 new annotations. In this context, one annotation refers to an identified behavior type with a start and end frame. These annotations are distributed across 226 individual rodents from 16 distinct experiments.

Active learning results

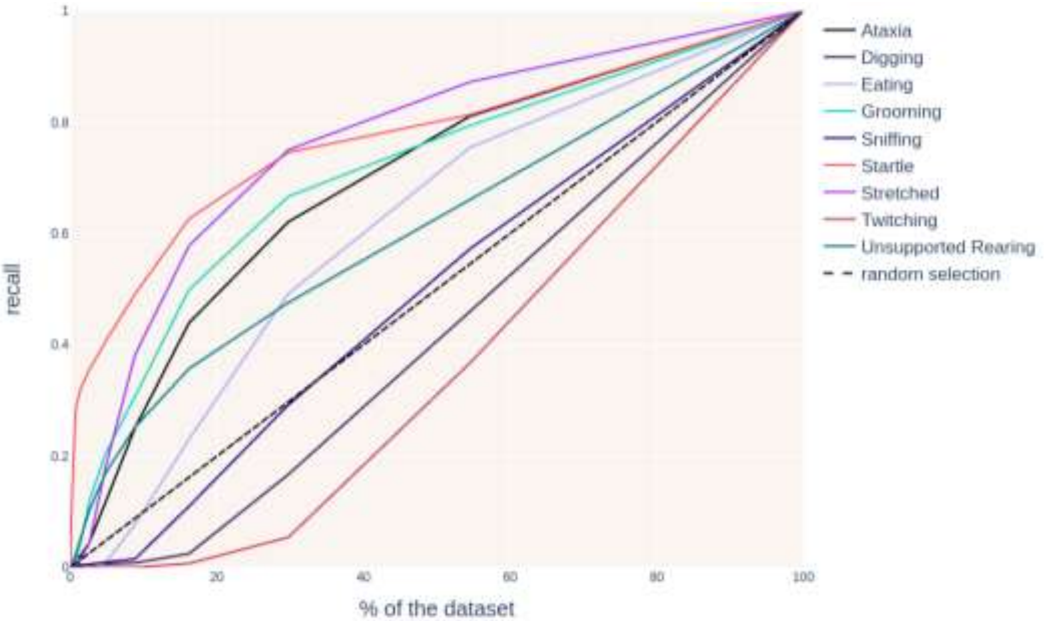


Figure 4. Recall of our system for rare behavior types with different subsets of the dataset. The percentage of coverage of a novel behavior when annotating a certain percentage of the overall dataset is shown.

With the help of the active learning component highlighting areas of high interest, the experts identified and added several new behavior types. Figure 4 demonstrates the benefit of using the system to extend the training dataset.

The novelty behavior detection model was evaluated through a leave-one-out assessment due to the absence of explicit labels for training and evaluation. This approach involves iteratively treating each known rodent behavior as novel, allowing us to gauge the proficiency of the method in identifying these established behaviors as potentially new instances. Further, it gives insights into the ability to expedite the discovery of behaviors by ranking them higher in the novelty queue. This is crucial for labelers using the labeling assistant algorithm, as it aids in identifying behaviors promptly rather than randomly later in the process. We observed a significant improvement in 6 out of 9 behavior types. Some of the behavior types would not have been prioritized by the novelty ranking, largely due to their intermediate position in the reduced feature space (Figure 3). Thus, we extended the active learning component to be composed of multiple strategies additionally to novelty ranking.

Classification results

As the amount of behavior types was heavily extended during the project, a direct comparison of model results at the beginning and the current state of the project is not feasible. However, a strong improvement can be seen in behavior types for which little training data was available at the start of the project due to the rare occurrence of these behaviors as shown in Table 1. With the data collection using our system and active learning, the amount of labels for some of these behavior types could be increased substantially. This enabled the training of a model, which can detect these behaviors more reliably. Examples for this are “Grooming”, where the true positive (TP) rate improved from 6% to 73%, and “Twitching” with an improvement from 3% to 29%. Moreover, the present model is able to detect a larger number of distinct behavior types.

Behavior Type	TP Initial Model	TP Current Model	Label Data Increased
Ataxia	0.79	0.83	× 1.1
Digging	0.02	0.06	× 4.3
Eating	0.38	0.88	× 2
Grooming	0.17	0.79	× 3.88
Sniffing	0.63	0.54	× 5.64
Startled	0.08	0.09	× 5.67
Twitching	0.03	0.29	× 4.8
Unsupported Rearing	0.72	0.72	× 1.7
Catalepsy	-	0.59	
Drinking	-	0.91	
Gnawing	-	0.29	
Interrupted Sleeping	-	0.45	
Jumping	-	0.85	
Supported Rearing	-	0.70	
Stretched		0.57	
Walking	-	0.78	
Writhing	-	0.11	

Table 1. Comparison of the performance between the initial model and the current model, as well as the difference in data set size. Each row refers to a class of behavior type that the models aim to detect. TP refers to the true positive rate of the predictions, evaluated on a per-frame basis. Label Data Increased lists the factor by how much the amount of annotated data for that class was increased from initial to current training data set. For simplicity, not all classes that the model was trained on are included.

Conclusion

Our study has demonstrated that the integration of a human-in-the-loop approach, combined with advanced AI technologies, significantly enhances the efficiency and accuracy of rodent behavior analysis in the context of drug development.

The active learning component has proven instrumental in discovering and annotating rare behavior types, as evidenced by the substantial increase in annotations and the identification of several new behavior classes. This enhancement in data richness not only improves the model's accuracy but also broadens the spectrum of behavior types that can be reliably detected. The increase in the amount of labeled data for rare behaviors has notably improved the model's performance, as highlighted by the substantial improvements in TP rates for behaviors such as Grooming and Twitching. The system has been proven to accelerate the discovery process and aid labelers in prioritizing behaviors for annotation.

DeepRod represents a significant advancement in the field of automated rodent behavior analysis. Its ability to efficiently process large datasets, coupled with its enhanced detection and classification capabilities, makes it a powerful tool for drug discovery and development. The collected user feedback confirmed that the speed of annotating the experiment data and the discovery of such a large amount of new behavior types would not have been possible without the provided system. As the system continues to evolve, it holds great potential for further improving the understanding of rodent behavior, contributing to more effective and efficient drug development processes and safe clinical trials.

Acknowledgements and contributions

Our deepest gratitude goes to Beatrice Kley, Cindy Janin Jung, and Michael Winter (Boehringer Ingelheim GmbH & Co. KG, Drug Discovery Sciences) for annotation of the data and for their feedback, wishes, ideas and immediate testing of these. We want to thank Malte Janssen, Moisei Shkil, Ziyad Sheebaelhamd, Lev Udaltsov, Atacan Korkmaz (Merantix Momentum GmbH) who developed software components used for DeepRod. DeepRod would not be possible without Florian Montel, Menorca Chaturvedi, and Markus Köster from opnMe.com [17] (F.M. Boehringer Ingelheim GmbH & Co. KG; M.C., M.K. Boehringer Ingelheim International GmbH). Fabian Heinemann and Peter Gross contributed the opnMe initiative at various stages and provided input on the requirements of the MVP. Kathrin Eschmann had a substantial impact on the late stage of the project and contributed to the manuscript (all Boehringer Ingelheim GmbH & Co. KG, Drug Discovery Sciences).

Adrian Loy, Miha Garafolj, Thomas Wollmann developed the machine learning method. Adrian Loy and Miha Garafolj did the algorithm implementation and performed the machine learning experiments. Hanna Behnke did the usability engineering and orchestrated the project. Cyrille Charnier and Philipp Schwarz evaluated the methodology. Georg Rast, Heike Schauerte contributed the initial concept and provided domain knowledge. Georg Rast and Thomas Wollmann supervised the project. Thomas Wollmann, Miha Garafolj, Adrian Loy, Heike Schauerte, Georg Rast, and Hanna Behnke wrote and proofread the manuscript.

Ethical statement

Maintenance and handling of animals are carried out in compliance with (i) the ethical guidelines established by German National Animal Welfare Laws within the framework of the European Union Directive 2010/63/EU and (ii) the Guide for the Care and Use of Laboratory Animals produced by the National Research Council and the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC). The study protocol was approved by the responsible German authority (Regierungspräsidium Tübingen).

References:

1. European Parliament and Council (2001). Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use. <https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32001L0083>. Accessed 21 December 2013.

2. FDA & EMA, 2001. ICH S7A Safety Pharmacology Studies for Human Pharmaceuticals (2001). <https://www.fda.gov/media/72033/download>. Accessed 21 December 2013.

3. Gad SC (2019). Safety Pharmacology in Pharmaceutical Development: Approval and Post Marketing Surveillance, Second Edition. *CRC Press*, Chapter 2.1, 19-24 and Chapter 5.1-5.2, 69-77.

4. Hamdam, J., Sethu, S., Smith, T., Alfirevic, A., Alhaidari, M., Atkinson, J., Ayala, M., … & Goldring C (2013). Safety pharmacology - current and emerging concepts. *Toxicol Appl Pharmacol*. 273(2), 229-41.

5. Spink, A.J., Buma, M.O.S., Tegelenbosch, R.A.J. (2000). EthoVision color identification: a new method for color tracking using both hue and saturation. *Proceedings of Measuring Behavior 2000*, 295-297.

6. Spink, A.J., Tegelenbosch, R.A.J., Buma, M.O.S., Noldus, L.P.J.J. (2000). The EthoVision video tracking system: a tool for behavioral phenotyping of transgenic mice. *Physiology & Behavior 73*, 731-744.

7. Otterbach, J., & Wollmann, T. (2021). Chameleon: A Semi-AutoML framework targeting quick and scalable development and deployment of production-ready ML systems for SMEs. *arXiv preprint arXiv:2105.03669*.

8. Mathis, A., Mamidanna, P., Cury, K. M., … & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21, 1281-1289.

9. Segalin, C., Williams, J., Karigo, T., … & Ann Kennedy (2021). The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* 10:e63720.

10. Föll, M. C., Moritz, L., Wollmann, T., Stillger, M. N., Vockert, N., Werner, M., ... & Schilling, O. (2019). Accessible and reproducible mass spectrometry imaging data analysis in Galaxy. *Gigascience*, 8(12)

11. Ritter, C., Wollmann, T., Lee, J. Y., Imle, A., Müller, B., ... & Rohr, K. (2021). Data fusion and smoothing for probabilistic tracking of viral structures in fluorescence microscopy images. *Medical Image Analysis*, 73.

12. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of ACM SIGKDD 2016*, 785-794.

13. McElfresh, D., Khandagale, S., Valverde, J., Ramakrishnan, G., Goldblum, M., & White, C. (2023). When Do Neural Nets Outperform Boosted Trees on Tabular Data?. *arXiv:2305.02997*.

14. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.

15. Sreenivasaiah, D., Otterbach, J., & Wollmann, T. (2021). Meal: Manifold embedding-based active learning. *Proceedings of IEEE ICCV 2021*. 1029-1037.

16. Bitterwolf, J., Müller, M., Hein, M. (2023). In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. *Proceedings of ICML 2023*.

17. Boehringer Ingelheim (2017). opnMe - Boehringer Ingelheim Open Innovation Portal. <https://opnMe.com>. Accessed 21 December 2013.

Symposium: Optimising Analysis of Longitudinal, high resolution behavioural data

Methods to assess variation of movement within and across laying hens within a commercial system over extended periods of time

M.J. Toscano

Center for Proper Housing: Poultry and Rabbits (ZTHZ), Division of Animal Welfare, VPH Institute, University of Bern, Burgerweg 22, 3052, Zollikofen, Switzerland. Michael.Toscano@unibe.ch

Introduction

Observations of agricultural animals using various sensor technologies is becoming more common with applications to identify early stages of adverse health conditions and other critical events. Beyond the ability to monitor different types of behaviors, we are also able to capture responses with a high degree of resolution and over extended periods of time that can enable detection of differences and changes across or within animals. Our research group has been focusing on positional data of poultry (mainly laying hens) using different tracking systems installed within several quasi-commercial, cage-free housing environments. Assessment of poultry within commercial cage free housing is particularly problematic relative to other agricultural species and housing types given the small size of the animals and density. As a consequence of the animal size and density, behavioral observation is difficult which intensifies the benefits that sensor technology can provide. Within Europe, but also increasingly across North America and globally, laying hens are maintained in cage-free housing (vs. small cages with 5-20 animals) creating a growing need for accurate and reliable methods to observe behavior.

The combination of high-resolution data measured over extended periods of time allows for visualization of patterns as they develop and change (over time) or maintain key features. Most interestingly, the integration of high-resolution positional data with concepts used to assess and understand animal behavior such as personality and behavioral syndromes is an exciting research theme. Whereas concepts such as personality where traditionally done with a limited number of assessments under controlled conditions with limited relevance to actual field conditions[1], [2], sensor technology allows such assessments to be done continuously within the home pen.

Although a powerful tool, the visualization of collected data and the creation of usable, meaningful metrics remains a challenge if the data is to be linked with established or novel indicators of health, welfare, and productivity. Our data is generally aggregated into daily variables of the original location as part or in entirety of a 24-hour day cycle, although we have also combined streams of location data (typically also within a 24-hour day cycle) into a composite variable for examination. The present abstract describes several approaches (including that of previously published efforts [3]–[8]) we have taken to capture the positional data, different metrics to describe that data, and then how those metrics were related to health and possible social indicators.

Animals and tracking environment

All data is collected within quasi-commercial barns containing approximately 4’500 laying hens reared on site under controlled conditions. The barn contains a Bolegg Terrace aviary split into 20 identical pens (described here [4] with each pen containing 225 hens/pen) including an outside covered winter garden (WG) accessible through pop holes. All barns are located at the Aviforum facility in Switzerland where standard animal husbandry practices are used and ethical approval for animal use was sought and approved.

Tracking data to date is only collected during the laying phase, i.e, from 17 weeks of age, though typically will span over forty weeks. Our experiments have used different tracking system [9]–[11] although the generated data was similar in nature where recordings are of animals being registered and transitioning between five specific zones within the barn – the upper, middle, and lower tiers; the floor, and the exterior wintergarden (Figure 1). Although we are not able to identify behaviors occurring within a zone, each contains specific resource for which the animals are known to be highly motivated to use. For instance, the upper and lower tiers contain feed and water available *ad libitum* whereas the floor contains litter which the birds use for dustbathing. The middle tier contains nestboxes for egg-laying. From the data confirming the animals' location and associated time-date stamps, we can extract the duration in each zone and transitions between zones, as well as when animals move together or enter/exit a zone within specified time windows.

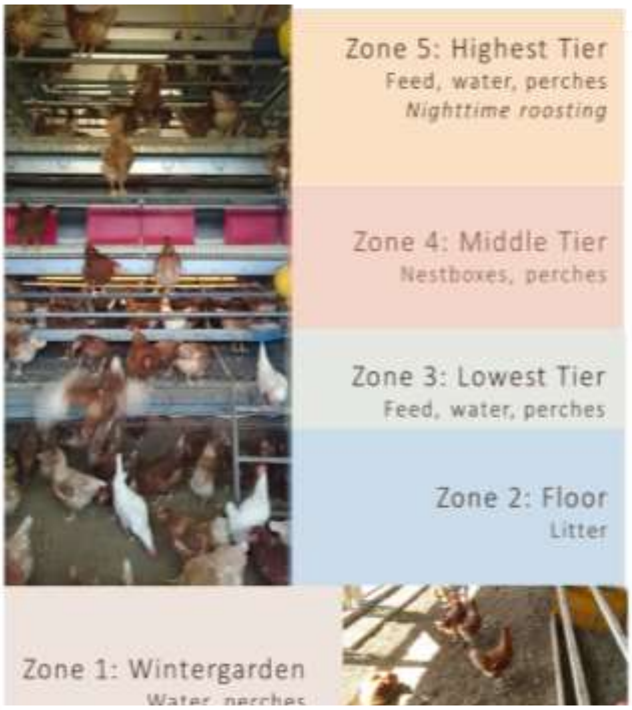


Figure 11. Representation of the animal pen divided into five distinct areas with associated resources.

Methods of analysis

We have used several measures that aggregate the data into specific metrics over part or the entirety of a 24-hour daily cycle. For example, examined patterns considered elements of the location data (number of entries and duration in zones, and total transitions) that were then reduced into a principal component that appeared to reflect general movement. Examined over a 54-day period, the principal component was demonstrated to show consistent inter individual differences in the overall average between animals that explained 44% of variation as well as variation in predictability and change over time [8]. More interestingly, hens with greater predictability in daily movements also demonstrated more severe bone fractures at the end of production. We are continuing to examine how particular behavioral phenotypes are associated with the development of keel bone fractures.

We have also been able to examine consistency of behavioral traits in relation to daily routines which can be extracted from the tracking data [7]. For instance, although we are not able to identify specific behaviors, there are features of the data which can be used to guide inference. In one example, egg laying will normally involve presence of 30 minutes or more within the middle tier that contains the nestbox (used for egg-laying) during the first seven hours of the lights coming on. Similarly, although food is available *ad libitum*, the animals prefer the feed that is delivered freshly five to seven times a day via automated feeders in the upper and lower tiers at specific times. Examination of these different behaviors across different commercially relevant contexts (e.g., vaccination or extreme cold) demonstrated them to be repeatable over time and across contexts with consistent differences between individuals explaining 23% to 66% of variation. We believe the consistency of these behaviors within individuals, especially during challenging episodes such as vaccination, can be used as indicators of behavioral resilience within commercial breeding programs.

Positional data has also been used to demonstrate coordinated movement between individuals that could reflect social connections in an effort using social network analysis. By using a Guassian Mixture Models method that clusters bursts of gathering events [12], we were able to extract associations between hens in terms of their proximity in time and space. Although all individuals were related to all other individuals in the pen, our system was able to identify a certain portion of hens (approximately 9% of the 225 hens per pen) that maintained particularly high strength ties over a contiguous 30-week period. Interestingly, an association index calculated for each week over the course of the study observed a weakened association over time, although the individuals identified in the relatively high strength tie group remained stable. Future work will need to better characterize these networks, including the role of individuals in maintaining the stability of high strength subgroups and possible benefits of connections via social buffering (e.g., improved health).

Conclusion

Sensor technology offers many opportunities to monitor animal health, welfare, and productivity within modern agriculture. Commercial laying hens and the ongoing transition away from cage to cage-free housing systems stands to greatly benefit from such technology given the difficulty to monitor the animals using traditional methods. We have used positional data of commercial poultry in cage-free housing to develop metrics that can then be linked with measures of health, welfare, and productivity. The strength in extended periods of data collection characterized by relatively high resolution resides in the ability to detect meaningful and useful patterns that can reveal contrasts not apparent with isolated or brief observational periods. When integrated with traditional themes of representing variation in animal behavior, sensor technology allows for a more accurate assessment with greater relevance to the animal's actual living conditions, breeding programs, and general improvements in quality of life.

References

1. A. M. Bell, S. J. Hankison, and K. L. Laskowski. (2009). The repeatability of behaviour: a meta-analysis. *Animal Behavior* **77**, 771–783.

2. G. G. Mittelbach, N. G. Ballew, and M. K. Kjelvik. (2014). Fish behavioral types and their ecological consequences. *Canadian Journal of Fisheries and Aquatic Sciences* **71**, 927–944.

3. C. Rufener, J. Berezowski, F. Maximiano Sousa, Y. Abreu, L. Asher, and M. J. Toscano. (2018). Finding hens in a haystack: Consistency of movement patterns within and across individual laying hens maintained in large groups. *Scientific Reports* **8**, 12303.

4. C. Rufener, Y. Abreu, L. Asher, J. Berezowski, F. Maximiano Sousa, A. Stratmann, and M.J. Toscano. (2019). Keel bone fractures are associated with individual mobility of laying hens in an aviary system. *Applied Animal Behaviour Science* **217**, 48–56.

5. Y. Gómez, J. Berezowski, Y. Jorge, S. Gebhardt-Henrich, S. Vögeli, A. Stratmann, M.J. Toscano, and B. Voelkl. (2022). Similarity in Temporal Movement Patterns in Laying Hens Increases with Time and Social Association. *Animals* **12**, 555.

6. C. Guerrero-Bosagna, F. Pértille, Y. Gómez, S. Rezaei, S. Gebhardt-Henrich, S. Vögeli, A Stratmann, B Voelkl, and M.J. Toscano. (2020). DNA methylation variation in the brain of laying hens in relation to differential behavioral patterns. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **35**, 100700.

7. C. M. Montalcini, M. B. Petelle, and M. J. Toscano. (2023). Commercial laying hens exhibit long-term consistent individual differences and behavioural syndromes in spatial traits. *Royal Society Open Science* **10**, 10.1098/rsos.230043.

8. C. M. Montalcini, M. J. Toscano, S. G. Gebhardt-Henrich, and M. B. Petelle. (2023). Intra-individual variation of hen movements is associated with later keel bone fractures in a quasi-commercial aviary. *Scientific Reports* **13**, 2377.

9. L. Candelotto, K. J. Grethen, C. M. Montalcini, M. J. Toscano, and Y. Gómez. (2022). Tracking performance in poultry is affected by data cleaning method and housing system. *Applied Animal Behaviour Science* **249**, 105597.

10. C. M. Montalcini, B. Voelkl, Y. Gómez, M. Gantner, and M. J. Toscano. (2022). Evaluation of an Active LF Tracking System and Data Processing Methods for Livestock Precision Farming in the Poultry Sector. *Sensors* **22**, 10.3390/s22020659.

11. S. G. Gebhardt-Henrich, A. Kashev, M. B. Petelle, and M. J. Toscano. (2023). Validation of a Radio frequency identification system for tracking location of laying hens in a quasi-commercial aviary system. *Peer Community Journal* **3**, 10.24072/pcjournal.324.

12. I. Psorakis, S. J. Roberts, I. Rezek, and B. C. Sheldon. (2012). Inferring social network structure in ecological systems from spatio-temporal data streams. Journal of the Royal Society Interface 9, 3055–3066.

Dyadic linear models for genetic analysis of behavioral interactions

Juan Steibel[1] and Janice Siegford[2]

[1] Department of Animal Science, Iowa State University, Ames, IA, USA. jsteibel@iastate.edu

[2] Department of Animal Science, Michigan State University, East Lansing, MI, USA, siegford@msu.edu

Abstract

In animal breeding, it is widely recognized that behaviors are subject to genetic as well as environmental control. Moreover, prediction of social genetic effects has been proposed to indirectly select for animals that are behaviorally better adapted to groups rearing conditions. However, a more efficient way to improve a herd's social genetic effects is through direct selection against or in favor of the behaviors in question. Dyadic models have been introduced for the phenotypic analysis of behavioral interactions between animals in a group. In this paper we expand those models to include genetic effects and we illustrate its application to a well-known pig behavior dataset. For the data in question, we show that the heritability of the probability of delivering aggression is around 0.3. Moreover, we also show that the probability of receiving aggression is not equal to zero but is one order magnitude smaller than the probably of delivering aggression. The presented models are a useful statistical tool for implementing selection of behavioral traits in group-housed domestic animals.

Background

Behavior interactions as dyadic data and linear models

In our previous work [1], we introduced dyadic linear models for phenotypic analysis of behavioral interactions occurring between animals in a group. The observational unit of dyadic linear models is defined at the level of a pair of animals (dyad). In this context, a dyadic trait or phenotype is a quantitative observation of a behavioral interaction between these two specific animals. For instance, the phenotype could be a binary outcome such as the occurrence (or not) of a defined interaction (e.g., play, aggression, etc.) between animals, or it could consist of a continuous or discrete response such as intensity of interactions using a pre-defined scale, duration of interactions in seconds, or the count of the number of interactions in a given period of time. We demonstrate that when modeling these dyadic data, it is important to include appropriate fixed and random effects to account for sources of variations affecting the expression of the dyadic phenotype. Moreover, our first analysis was purely phenotypic, in other words, we did not model any genetic source of variation.

Direct and Social genetic effects and behavioral interactions

It is well known that behavioral traits are subject to genetic and environmental control and it is possible to implement artificial selection to improve social behavior of farm animals [2]. Moreover, behavioral interactions between animals often result in variation in other phenotypes. For instance, when pigs are housed in groups, animals showing behaviors that enable them to compete for feeder space may grow at a faster rate compared to other animals in the same social group. Animals breeders recognize this and have proposed the estimation of social genetic effects [3]. Social effects are expressed as observed variation in the phenotypes of the social group mates of the individual carrying those genetic effects. From the point of view of the prediction of social genetic effects, they are considered indirect effects, as the breeding value of a focal animal will be expressed in the phenotype of (behaviorally) connected animals and not in the focal animal itself. Our group pioneered the use of behavioral interaction data for improving the identifiability of those effects under standard mating designs in pig breeding [4].

Here, we revisit the estimation of direct and indirect genetic effects, but in the context of dyadic linear models. We expand our previously published models to include phenotypic and genetic effects and illustrate their use with a previously published dataset.

Methods

Basic Dyadic model

As presented before, the dyadic model is fit to data observed in pairs of individuals. Thus, each observation is indexed with three subscripts: the index of each animal involved in the interaction and the index of the social group in which they coexist. Furthermore, the modeling varies slightly depending on the directional nature of the behavioral interaction. In the case of directional behavioral interactions, one of the animals delivers a behavior and the other animal receives it. Thus, for directional dyadic data the subindexes identifying each animal are not exchangeable and there are two observations per each dyad (see equation 1). Contrarily, for non-directional interactions, there is no clear deliverer or receiver and only one interaction can be recorded for each pair of animals (not shown in this paper).

For directional interactions, we can write the following general linear model:

$$y_{ijk} = b_0 + FE_{ijk} + g_i + r_j + sg_k + e_{ijk} \ , \quad (1)$$

where, y_{ijk} is the observed phenotype of the interaction where the *i*th animal in the *k*th group delivers a behavior to the *j*th animal in the same group. In the case of the reciprocal interaction (same dyad in opposite direction), this would be represented as y_{jik}. FE_{ijk} are fixed effects that can be modeled at the individual or dyadic level as presented elsewhere [1]. $g_i \sim N(0, \sigma_g^2)$ is the random effect of the deliverer or giver of the behavior, $r_i \sim N(0, \sigma_r^2)$ is the random effect of the receiver of the behavior. $sg_k \sim N(0, \sigma_s^2)$, is the random effect of the social group, and $e_{ijk} \sim N(0, \sigma_e^2)$ is a residual. Note that in this example we are assuming a trait that can be modeled with a fully Gaussian model, but in our previous paper [1] we showed the use of a mixture Bernoulli-Poisson model and in the application section of this presentation we show a Bernoulli distribution. Thus, the model is general enough to incorporate various distributional assumptions depending on the phenotypes of interest.

For conciseness we do not show the non-directional model, which is a simplified version of equation 1.

Expanding the model to incorporate genetic effects

In Equation 1, animal effects g_i and r_i can be extended to incorporate genetic effects. For instance, given a suitable genetic relationship matrix **G**, the vector of giver random effects can be assumed to be $g \sim N(0, G\sigma_g^2)$. Similarly, the vector of receiver random effects can be assumed $r \sim N(0, G\sigma_r^2)$. Many other assumptions can be incorporated such as inclusion of temporary environmental effects, correlated giver and receiver effects, and dyadic random effects, which are beyond the scope of this paper.

Once an appropriate model is elicited for a suitable dataset, standard frequentist or Bayesian methods can be used to estimate all unknown parameters and to predict breeding values for each animal delivering and receiving the behavior in question.

Illustration with pig aggressive interactions

To illustrate the use of these models with real data, we reanalyzed our previously published dyadic data [1] using a genomic relationship matrix for the animals in question [5].

Dataset

The data consisted of 10,032 directional dyadic of single-sided attacks in 797 purebred Yorkshire barrows and gilts pigs housed in 59 single-sex social groups. The behavioral observations occurred in the 6 hours that immediately followed the mixing of pigs into new social groups upon moving from nursery to finishing pens. The finishing social groups had between 12 and 14 animals that came from four to six different nursery social groups. For present work, we modeled the probability of aggression, thus the data consisted in binary 0/1 observations depending on if a particular animal attacked a social group mate. The relationship matrix was derived from over 40,000 SNP genotypes available from a previous study.

Model elicitation and model fit details

The model was similar to the one described in equation 1,but a Bernoulli-logit model (rather than a Gaussian model) was used for the binary data. Fixed effects included in the model were sex, weight of the receiver relative to the average social group weight, weight of the deliverer relative to the average social group weight, an indicator of previous litter mates (1 if the animals were litter mates and 0 of they did not share the same litter) and an indicator of previous nursery pen mates (1 if the animals had been nursery pen mates and 0 of they did not share the same nursery social group).

Model fitting was performed using a fully Bayesian approach with flat priors and a Gibbs sampler to estimate model parameters. Convergence of the Gibbs algorithm was monitored using standard MCMC convergence diagnostic tools described in our previous paper and its supplemental materials [1].

For presenting results, we characterize the posterior distribution of the fixed effects coefficients and variance covariance parameters using a three-number summary. The summary statistics included the 2.5 percentile, the 97.5 percentile, and the 50 percentile (median) of the posterior distribution. These summary quantities can be easily interpreted as the central value (median) and a posterior credibility interval around the central value defined by the extreme percentiles. We also derived and summarized the posterior distribution of the heritability of the giver and receiver effects (Table 1).

Table 1. Posterior percentiles of the posterior distribution of variance components and genetic parameters.

	2.5%	50%	97.5%
Group Variance	0.108	0.168	0.269
Receiver Genetic Variance	0.030	0.047	0.072
Giver Genetic Variance	0.543	0.670	0.823
Residual Variance	0.097	0.165	0.242
Heritability of Giver Effect	0.281	0.327	0.372
Heritability of Receiver Effect	0.015	0.023	0.035

The giver effect explained the most variance and its heritability was approximately 0.32. This indicates that the probability of delivering aggression is inheritable and that there is potential for observing a response to selecting against it. The heritability of the receiver effect was much smaller, estimated as 0.023. We fit a reduced model and found that this small effect was statistically significant. Thus, the variance of the receiver effect was small but not equal to zero, hinting at the existence of a heritable “victim effect” or propensity to receive attacks from group mates.

Posterior distributions for fixed effects were close to those obtained with a purely phenotypic model [1] and they are omitted from this paper.

Conclusions

- Dyadic models can be used to estimate direct and indirect phenotypic and genetic effects from behavioral interactions occurring between pairs of animals in a group.
- These models represent a parameter-rich alternative to more traditional models for the analysis of individual-specific behaviors because they allow the modeling of the probability or intensities of delivering and receiving a specific behavior while accounting for genetic similarities as well as past life history (e.g., shared social groups, etc.).
- The inclusion of genetic effects in dyadic behavioral models is the first step in enabling the direct genetic selection of animals for or against specific social behaviors in farm animals.

References

1. J. Han, J. Siegford, G. de los Campos, R.J. Tempelman, C. Gondro, J.P. Steibel, Analysis of social interactions in group-housed animals using dyadic linear models, Appl Anim Behav Sci 256 (2022) 105747. https://doi.org/10.1016/j.applanim.2022.105747.

2. S.P. Turner, Breeding against harmful social behaviours in pigs and chickens: State of the art and the way forward, Appl Anim Behav Sci (2011). https://doi.org/10.1016/j.applanim.2011.06.001.

3. P. Bijma, The quantitative genetics of indirect genetic effects: A selective review of modelling issues, Heredity (Edinb) (2014). https://doi.org/10.1038/hdy.2013.15.

4. B.K. Angarita, R.J.C. Cantet, K.E. Wurtz, C.I. O'Malley, J.M. Siegford, C.W. Ernst, S.P. Turner, J.P. Steibel, Estimation of indirect social genetic effects for skin lesion count in group-housed pigs by quantifying behavioral interactions1, J Anim Sci 97 (2019) 3658–3668. https://doi.org/10.1093/jas/skz244.

5. K.E. Wurtz, J.M. Siegford, R.O. Bates, C.W. Ernst, J.P. Steibel, Estimation of genetic parameters for lesion scores and growth traits in group-housed pigs, J Anim Sci 95 (2017) 4310–4317. https://doi.org/10.2527/jas2017.1757.

European Network on Livestock Phenomics (EU-LI-PHE): Mission on Big Data Focused on All Types of Animal-Related Phenotypes, Including Behavior

M.J. Toscano[1], T. Norton[2], L. Fontanesi[3] and the EU-LI-PHE Consortium[4]

[1] Center for Proper Housing: Poultry and Rabbits (ZTHZ), University of Bern, Switzerland. michael.toscano@unibe.ch
[2] M3-BIORES Research Group, Division of Animal and Human Health Engineering, Department of Biosystems, KU Leuven, Belgium. tomas.norton@kuleuven.be
[3] Animal and Food Genomics Group, Division of Animal Sciences, Department of Agricultural and Food Sciences, University of Bologna, Italy. luca.fontanesi@unibo.it
[4] COST Action CA22112 Working Group Members, 40 Countries

Background

Phenomics is emerging as a major new technical discipline in applied biology, including animal husbandry. Phenomics is focused on one major aim: to systematically describe the phenome, referred to as the physical and molecular traits of an organism, that in this context is an animal. This discipline can be defined as the ensemble of methodologies and technologies for the acquisition, analysis, and exploitation of high-dimensional phenotypic data on an animal-wide scale [1].

In the animal production sector, the availability of accurate and specific phenotype data can inform new breeding objectives and related breeding and selection programs and provide novel essential information for the daily activities and choices of the farmers needed to optimise reproduction strategies, disease control and welfare of the animals. Therefore, phenomics applied to animal breeding and husbandry can be considered an essential innovation to support the sustainability of all animal production systems.

Animal phenotypes can be classified in different ways according to the level in which they are measured, the type of information that is recorded, the temporal acquisition of the information and the objective of the collected parameters. For example, targeted phenotypes can be dynamic (changing rapidly within short periods of time) or stable (minimal change across a predetermined time window). Depending on the level of analysis, phenotypes can also be classified i) as external or final phenotypes and ii) as internal or molecular phenotypes (endophenotypes). Examples of final phenotypes are performance, morphological, disease resistance and behavioral traits. Examples of molecular phenotypes are the level or the presence/absence of different types of biomolecules (and their modifications) in animal biofluids and tissues and so on. Final phenotypes are determined by the contribution and interplay of many molecular phenotypes (with multi-level relationships) and their interaction with environmental factors. As a consequence of the broad heterogeneity in phenotype classes, a wide array of scientific approaches and technologies can be used to capture and manage phenotypic information. For example, phenomics can benefit from the development and application of automatic sampling or non-invasive methods to obtain repeated sampling and images, records or continuous data collection (including photographs, videos, sounds, movement traces, and so on) from a part of an animal, the whole individual or a population at different stages, or on the final animal products to describe final external phenotypes with high resolution and in real-time. To capture internal phenotypes, phenomics can also use sequence-based and functional omics technologies to detect and quantify molecular phenotypes (e.g. DNA methylation, RNA transcripts, proteins, metabolites, microbiota, glycomics, etc.).

The broad spectrum of phenotypes and the multiplicity of ways that they can be captured will inevitably produce very large quantities of heterogeneous and complex data outputs, placing phenomics firmly in the realm of data science and "big data". In this regard, phenomics is becoming increasingly important and attracting great scientific interest in livestock. It may be predicted that phenomics will be either on a par with genomics or will be the most demanding biological discipline in terms of data acquisition, storage, distribution, and analysis.

Animal behavior is a very complex trait that can be better described and then understood with appropriate phenomics approaches that can capture high density, multi-dimensional and continuous phenotypic data, producing large amounts of data that should be properly mined and elaborated to extract meaningful features and descriptors.

European Network on Livestock Phenomics

The development and the application of phenomics in livestock clearly requires multi-disciplinary and multi-actor approaches to bring together different expertise, resources, and expectations. As livestock phenomics requires experts in many fields, the critical mass of knowledge and expertise, we developed a European-based network that has been funded by the European Union. The network is a COST Action (**European Network on Livestock Phenomics – EU-LI-PHE** [2]) that has been constructed over four main challenges, which have been used to structure the activities in four working groups (WG):

WG1) Phenotyping technologies whose aims are i) to provide an overview of current phenotyping technologies and infrastructures that can be used for applications in livestock phenomics and ii) to define a roadmap of the research needs to capture high-dimensional phenotypic information on an animal-wide scale;

WG2) Genome to phenome integration whose aims are i) to provide an overview of the links between genome/epigenome variation and phenotypic variation at multiple levels in the main livestock species, ii) to identify synergies with related initiatives on functional analyses of livestock genomes and to iii) identify knowledge gaps and research needs and provide a road map with a clear trajectory to new applications.

WG3) Computational resources and methodologies for data analyses whose aims are i) to provide an overview of the computational models, methods and tools available and current and future needs for development of applications in the context of livestock phenomics, ii) to identify the needed synergies and developments in terms of cyberinfrastructures and computational capabilities;

WG4) Economic impact, regulations, policies, and society whose aims are i) to provide an overview on the potential technological and economic impact of livestock phenomics, ii) to summarize the regulatory frameworks around this discipline and evaluate access to information and data generated and iii) to analyze societal perceptions of livestock phenomics.

Another WG, **WG5 Stakeholder engagement, communication, and dissemination** links all the activities carried out in WG1-4, i) to ensure a continuous engagement of the stakeholders, ii) to ensure overall communication and iii) to ensure publication of reviews, reports, surveys and establishment of a website and social medias.

EU-LI-PHE is funded for four years (2023-2027) and at present gather more than 200 WG members from more than 40 countries. One of the most relevant missions of EU-LI-PHE is on big data focused on all types of animal-related phenotypes, including behavior.

Here we will present the structure of EU-LI-PHE and the expected deliverables.

References

1. Houle, D., Govindaraju, D.R., Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics* **11**, 855-866.

2. COST – European Cooperation in Science & Technology (2023). CA22112 – European Network on Livestock Phenomics (EU-LI-PHE). <https://www.cost.eu/actions/CA22112/>. Accessed 10 January 2024.

Oral General Sessions

Behavioural Tests

Measuring motivational switching in mice using open-design: the Switchmaze

C. Hartmann[1], A. Mahajan[1], V. Borges[1], L. Razenberg[1], Y. Thönnes[1] and M.M. Karnani[1,2]

1 Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, The Netherlands
2 Institute for Neuroscience and Cardiovascular Research, Centre for Discovery Brain Sciences, University of Edinburgh, UK. mkarnani@ed.ac.uk

Abstract

Animals need to switch between motivated behaviours, like drinking, feeding or social interaction, to meet environmental availability and internal needs. However, motivational switching is rarely studied, partly due to lack measurement systems. We designed an automated extended home-cage for measuring motivational switching in mice, the Switchmaze, using open source hardware and software. As proof-of-concept, we show environmental manipulation and targeted brain manipulation experiments which altered motivation switching without effect on traditional parameters.

Introduction

Motivated behaviours such as feeding, drinking and social interaction form a scaffold upon which we build our daily lives. These are generated by neural mechanisms, here termed drives[1,2]. In health, these drives alternate to meet internal needs, while also being controlled by availability of goal objects[3] and controllable through conscious effort. Therefore, the neural mechanisms behind motivation selection are likely to be complicated and probabilistic, requiring measurements of repeatable discretized behavioural epochs. Here, we set up an apparatus (Figure 1) to measure motivational switching in mice on the timescale of milliseconds.

Rapid motivational switching is not studied often. Instead, typical assays measure initiation and overall duration of motivated behaviours, often indicated via operant levers/nose-pokes[4–7]. On a slower timescale, a probabilistic behavioural satiety sequence of feeding-grooming-resting, occurring over the course of an hour has been described in rodents and crayfish[8,9]. Transitions between sequential motivated behaviours in the home-cage environment are seldom measured at a high temporal resolution[10–12]. This may be because motivational switching can be highly disordered in the home-cage where the goal objects are constantly accessible. This makes quantification difficult and does not reflect typical conditions outside the laboratory. Therefore, we set up a naturalistic sequential foraging task which discretizes motivational switching between competing behavioural cycles (Figure 1A).

The Switchmaze is an automated, liveable maze where feeding and drinking are discretized spatially. Previous automated measurement habitats have focused on recording/training for a head-fixed task[13–15] or tracking individualistic traits, like sociability, across a long time course[16–19]. Our goal was to capture spontaneous switching between motivations for feeding, drinking and returning to the nest for social interaction, enabling studies of the underlying neural mechanisms, i.e., drive switching. The Switchmaze consists of a home-cage coupled through a single entry module to a foraging environment where an animal can serially retrieve a quantum of food or water from an isometrically placed decision point (Figure 1B). This discretizes the phases of feeding and drinking cycles, including the switch epoch, making them readily quantifiable.

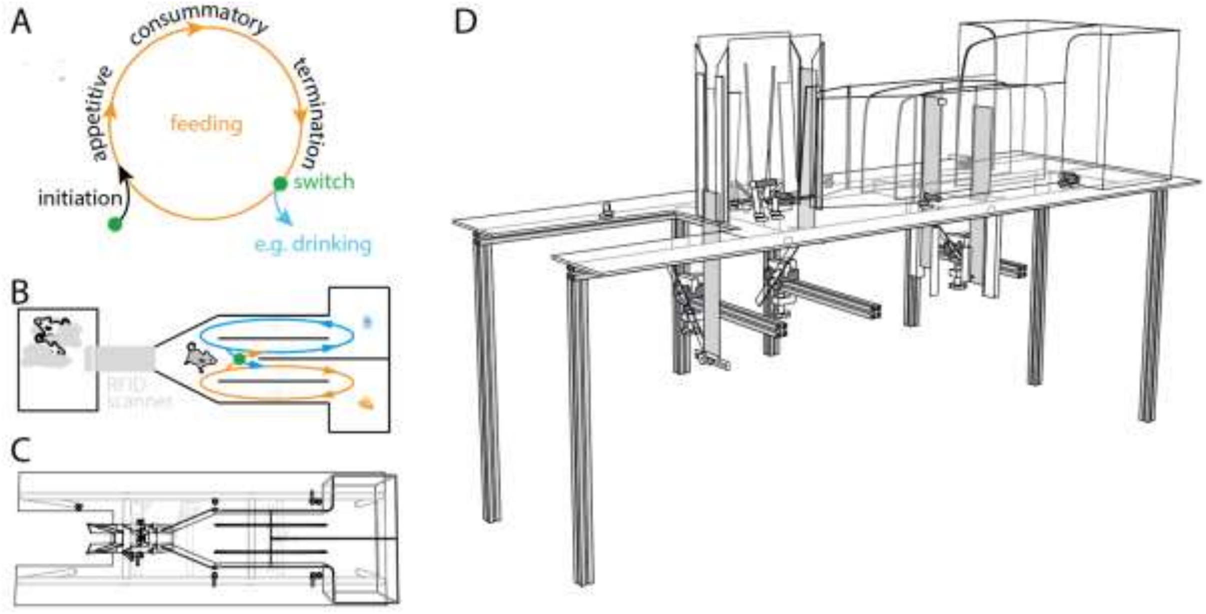


Figure 12, Measuring behavioural cycles spatially using the Switchmaze. A, Concept of feeding as a behavioural cycle, after[20,21]; B, Schematic of feeding and drinking cycles separated spatially in the Switchmaze; C, CAD model from above; D, CAD model.

Methods

Switchmaze

Detailed build instructions for the Switchmaze, a bill of materials and code are available in our public document and repository[22]. The apparatus (Figure 1, C,D) discretizes behavioural cycles of feeding and drinking, which animals perform one at a time in a foraging environment separated from the home cage by a single-entry module. Once inside the foraging environment, upon each entry to a goal area (a trial), food and drink were only available in a 'quantum' of either one 14 mg pellet (Bio-Serv™ Dustless Precision Pellets™ for Rodents, F05684) or one ~10-20 µl drop of water. After goal entry, the animal has one 'return' path available to the start position, which differs from the entry path. Once the animal is in the start position, the availability resets for both food and drink, and the animal also has the option to return to the home-cage. Thus, switching between three fundamental motivated behaviours can be followed when the animal is in the start position. Mice live in the apparatus for several days on a 12/12 light dark cycle (lights off at 9 am). The Switchmaze records entries to the foraging environment, start point, and feeding and drinking areas, as well as the consummatory actions, pellet retrieval and drinking. Health and welfare monitoring and maze cleaning was done at least once daily and safe operation was monitored on an overhead camera. The apparatus was in an isolated procedure room for all recordings.

Stereotaxic surgery

30 wild-type C57BL6 male mice were used in this study. All experimental procedures were approved by the Netherlands Central Committee for Animal Experiments and the Animal Ethical Care Committee of the Vrije Universiteit Amsterdam (AVD11200202114477). Of the 30 mice, 9 were controls (ctrl), 10 expressed inhibitory designer receptors exclusively activated by designer drugs (DREADDs) in the hypothalamus (H-hM4Di), and 11 expressed inhibitory DREADDs in PFC→hypothalamus projection neurons (PFC-hM4Di). Mice were anesthetized with 'sleep mix' (i.p., fentanyl 0.05 mg/kg, medetomidine 0.5 mg/kg and midazolam 5 mg/kg in saline), the scalp was injected subcutaneously with lidocaine, opened, and 0.2 mm craniotomies were drilled bilaterally at 0.9 mm lateral, 1.4 mm posterior from Bregma. For medial PFC injections of AAV8-syn-DIO-hM4Di-mCitrine, additional craniotomies were drilled bilaterally at 0.4 mm lateral, 1.8 mm anterior from Bregma, and 0.4 mm lateral, 2.3 mm anterior from Bregma. A pulled glass injection needle was used to inject the below doses of virus at a rate of 10-50 nl/min. All H-hM4Di and PFC-hM4Di animals received hypothalamic injections bilaterally 5.4 mm deep in the brain. For PFC-hM4Di animals, the hypothalamic injections contained 30 nl of AAVrg-hSyn-Cre-P2A-tdTomato (1.5*10^{13} GC/ml). For 4 out of 9 H-hM4Di animals, the injections contained 30-150 nl of 1:5 mixture of AAV9-CMV-Cre-tdTomato (10^{12} GC/ml) and AAV8-hSyn-DIO-HA-hM4Di-

mCitrine (10^{13} GC/ml), and the other 5 out of 9 animals received 30 nl of AAV8-hSyn-hM4Di-mCherry ($2*10^{13}$ GC/ml). All PFC-hM4Di additionally received medial PFC (mPFC) injections bilaterally. These contained three injections for each hemisphere, of 150 nl AAV8-hSyn-DIO-HA-hM4Di-mCitrine (10^{13} GC/ml). Two 150 nl doses were injected at 1.8 mm anterior from Bregma, 0.4 mm lateral at depths 2.0 mm and 1.5 mm, and one dose at 2.3 mm anterior from Bregma, 0.4 mm lateral at depth 1.75 mm. Injection needles were kept in place for 20 min in hypothalamus and 3-5 min in mPFC before withdrawing. After the injections, an RFID chip (Sparkfun SEN-09416) was implanted under the chest skin, the wounds were closed with tissue glue, anaesthesia was antagonized with wake mix (i.p., flumazenil 0.1 mg/ml and atipamezole 5 mg/ml in saline) and animals received 0.05 mg/ml carprofen in drinking water for 2-4 days as post-operative pain medication.

DREADD manipulation experiments

Animals were injected with saline on the first experiment day and 5 mg/kg of compound-21 (C21, the agonist of hM4Di) dissolved in saline on the second day. Motivation switching was measured over a 6-h window as the effect of C21 is likely to last at least that long [23,24].

Statistics

Data were analyzed using Matlab R2019a. For the chemogenetic experiments we used two-way repeated measures ANOVA with time (vehicle or C21) as the within-subjects factor, and cohort (ctrl, H-hM4Di or PFC-hM4Di) as the between-subjects factor. This was done by fitting a ‘WithinDesign’ repeated measures model (fitrm) in Matlab with Cohort as the predictor variable, followed by repeated measures analysis of variance (ranova). When a significant cohort-time interaction was found, three follow-up paired t-tests were used with Bonferroni-corrected significance threshold of 0.0167. Paired t-tests (Bonferroni-corrected significance threshold in Figure 3, 0.01) and Wilcoxon rank sum tests were also performed in Matlab.

Results and discussion

To study motivational switching, C57BL6 mice were housed in the Switchmaze for up to a month in cohorts of 2-4 animals. The mice used the foraging area purposefully for drinking and feeding: A pellet was consumed upon 93.9 ± 10.1 % of food goal entries, and drinking occurred upon 98.5 ± 2.3 % of entries into the water goal.

To satisfy a need such as hunger, an animal would be expected to enter the food area repeatedly, as only one quantum of consumption (see Methods) was available at each trial. These repeated entries, termed runs, occurred to both food and drink areas (Figure 2A, inset). However, the most common run length was one trial. Similar single trials are common in rodent behaviour, in particular during trained performance of simple tasks by highly skilled animals[7,25,26]. This seemingly counterproductive behavioural stochasticity may be part of a behavioural camouflage mechanism evolved to elude competitors[27]. In order to capture these prevalent ‘singles’ in a metric of motivational switching rate, we analysed the ratio of singles to runs (from here, we define a run as a sequence of repeated food or drink trials longer than one). This motivation switching rate was on average 1.4 ± 0.9 (Figure 2B,C) after habituation (>7 days). To test if motivation switching was different from random, we randomly shuffled the trial sequence and recalculated the switch rate 1000 times for each animal (Figure 2C,D). The arising distribution encompassed the actual switch rate in most cases (only 4/30 animals had a motivation switching rate higher than the 99th percentile of the shuffled data). This suggests that spontaneous motivational switching is optimized to appear random, which could conceivably function to decrease the predictive information available to competitors and predators.

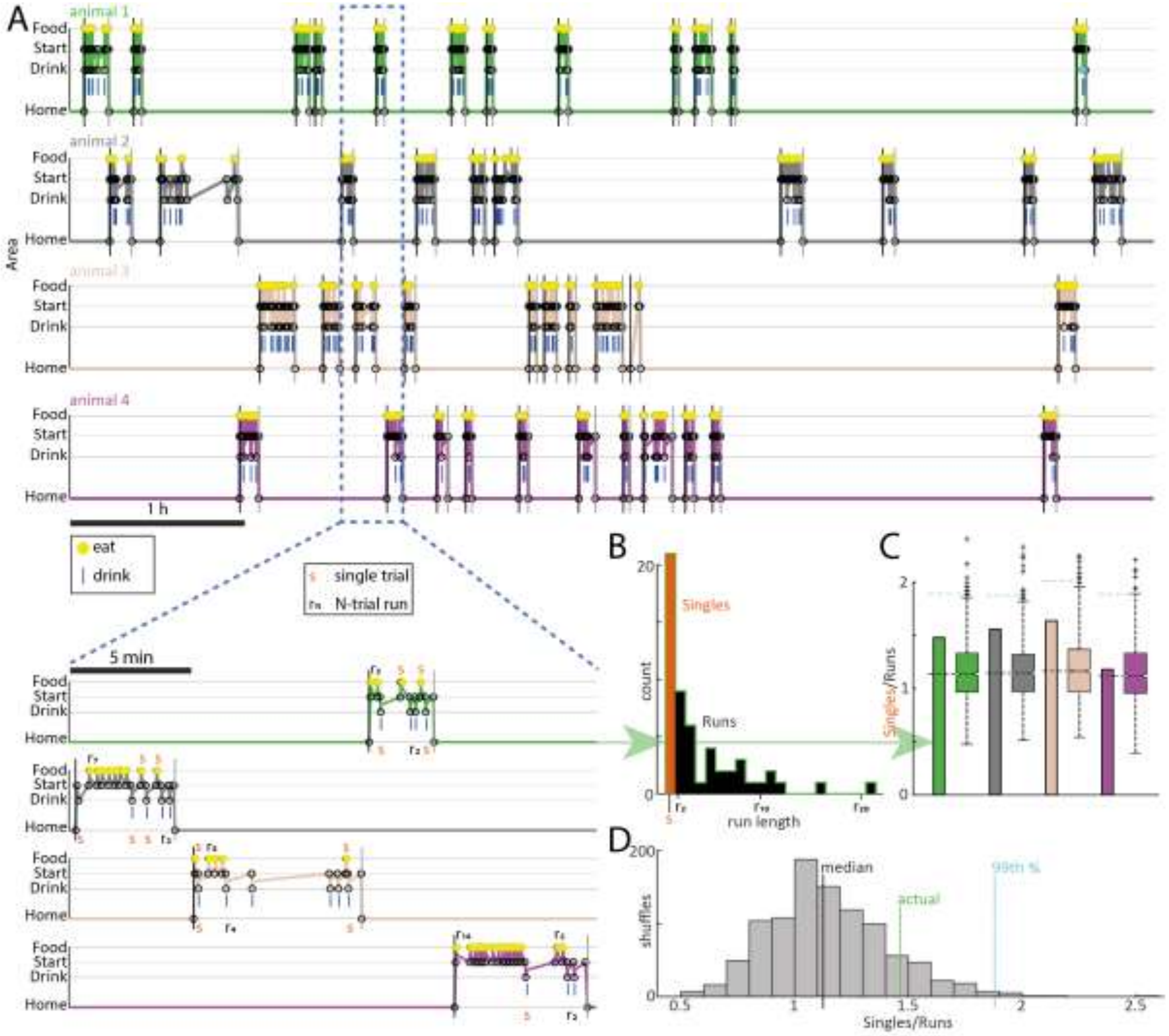


Figure 2, Example data from four mice in the Switchmaze showing discrete switching between drinking and eating. A, Ethograms for four mice entering the foraging environment one at a time for open-ended blocks. One block for each animal shown in detail in the expanded time window (dashed box). ‘Singles’ and runs labelled for clarity. B, For animal 1 only, distribution of singles and runs by length. C, Motivation switching for each animal (bars, number of singles divided by number of runs of any length) and motivation switching expected by random chance (box plots are the singles/runs metric from 1000 random permuted behaviour sequences for each animal; dashed horizontal lines denote median (black) and 99th percentile (cyan)). D, Distribution of switching rates from shuffled behaviour sequences for animal 1.

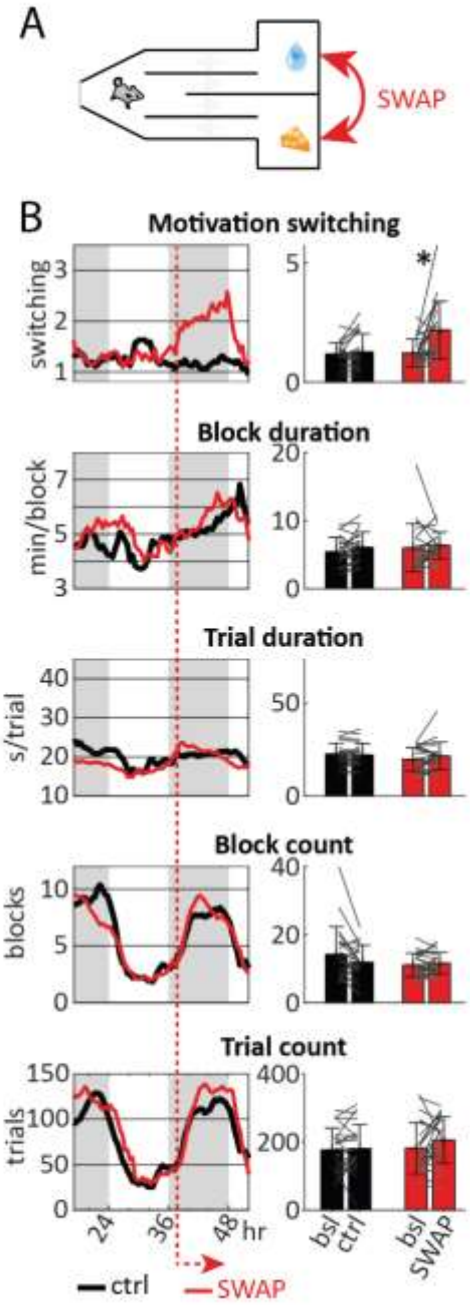


Figure 3, Dynamics of Switchmaze parameters during environmental uncertainty. A, Schematic of the goal module swapping experiment (SWAP); B, Behavioural measures during control and SWAP experiments, from top to bottom: Motivation switching (singles/runs, see Figure 2B,C), Block duration, Trial duration, Block count, Trial count. Time series data (left and middle panels) have a 4 h *bin width and bar graphs (right panels) are measured from the last 6 h before lights-on. N=22. *, p<0.004.*

We next asked how does motivational switching change when the mice encounter environmental challenges. We simulated an environmental uncertainty, similar to depleting resources in patch foraging, by swapping the food and drink areas (Figure 3). This resulted in increased motivation switching without other parameter changes, demonstrating that the motivation switching variable is a useful indicator of behavioural structure changes that would not be evident from traditional metrics.To assess the potential of the Switchmaze for identifying neural underpinnings of motivational switching, we performed two chemogenetic loss-of-function experiments on neural circuits known to be involved in feeding. A broad perifornical region of the hypothalamus, containing the lateral, dorsomedial and tuberal areas, regulates feeding potentially through primary effects on arousal, locomotion and metabolism[28–30], and contains intermingled feeding promoting and inhibiting neural populations[31–33]. The medial prefrontal cortex (PFC) sends axonal projections to the perifornical hypothalamus, affecting feeding in a complex manner depending on behavioural context[34,35]. To test the role of these neural populations in drive switching, we expressed the inhibitory DREADD, hM4Di in either PFC output neurons to the hypothalamus (PFC-hM4Di) or in the perifornical hypothalamus (H-hM4Di). As a control cohort, we used wild-type mice that did not express a transgene, and which were interleaved in groups of hM4Di expressing mice.

Basic behavioural metrics were not changed in PFC-hM4Di or H-hM4Di mice upon activation of the inhibitory DREADDs with C21, as two-way repeated measures ANOVA tests showed no significant cohort-time interaction for food consumed (F(2,27)=0.61, p=0.55), block count (F(2,27)=0.66, p=0.52), trial count (F(2,27)=0.50, p=0.61), block duration (F(2,27)=0.75, p=0.48) or trial duration (F(2,27)=0.69, p=0.51). However, motivation switching was altered significantly (cohort-time interaction F(2,27)=3.99, p=0.03) and paired t-tests revealed the effect was a 46.5 ± 45.5% increase in the PFC-hM4Di cohort (p=0.007, Figure 4). The average switch rate exceeded the 99th percentile of switch rates expected from randomly permuted trial sequences (upper dashed lines in Figure 4). This coincided with 6 out of 11 PFC-hHM4Di animals increasing their switch rate above the 99th percentile of their distribution of random sequences (cyan circles in Figure 4). This result supports previous findings suggesting that a behavioural role of the PFC is to regulate decision sequence stochasticity[27].

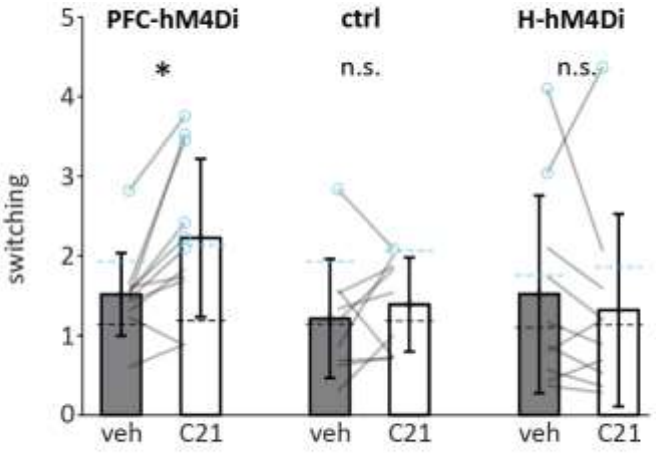


Figure 4, Chemogenetic modulation of motivational switching. A, Motivation switching quantified as singles/runs in PFC-hM4Di (left, n=11), ctrl (middle, n=9) and H-hM4Di cohorts (right, n=10) during the 6 h after vehicle or C21 injection. Two-way repeated measures ANOVA revealed a significant cohort-time interaction F(2, 27)=3.99, p=0.03 and paired t-tests revealed the effect was in the PFC-hM4Di cohort, * p=0.007. Dashed lines denote chance level arising from the mean of medians (black dashed line) and mean of 99th percentiles (cyan dashed line) of 1000 random-permuted behavioural sequences for each animal. Additionally, for each animal, switching rate higher than the 99th percentile of its random permutations is labelled with a cyan circle.

Conclusions

- The open-design (open source hardware, software and publicly available design) based Switchmaze can be used to measure motivational switching in an ethological semi-natural setting.
- Motivational switching is measured as the ratio of single behavioural cycles to runs, which summarizes rodent-typical frequent behavioural switching. This parameter is called motivation switching.
- Motivation switching is increased under environmental uncertainty while standard behavioural parameters, such as trial count and duration, are unchanged.
- Motivation switching is modulated by PFC→Hypothalamus projection neurons.
- Motivation switching is therefore a useful parameter which may be innately modulated by mouse brain in order to generate behavioural sequence stochasticity in order to elude competitors. This could be a previously unnoticed form of behavioural camouflage.

References

1. Pfaff, D.W. (1999) *Drive: Neurobiological and Molecular Mechanisms of Sexual Motivation*, MIT Press.

2. Berridge, K.C. (2004) Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, **81** (2), 179–209.

3. Ghoniem, A., van Dillen, L.F., and Hofmann, W. (2020) Choice architecture meets motivation science: How stimulus availability interacts with internal factors in shaping the desire for food. *Appetite*, **155**, 104815.

4. Caprioli, D., Zeric, T., Thorndike, E.B., and Venniro, M. (2015) Persistent palatable food preference in rats with a history of limited and extended access to methamphetamine self-administration. *Addict Biol*, **20** (5), 913–926.

5. Venniro, M., Zhang, M., Caprioli, D., Hoots, J.K., Golden, S.A., Heins, C., Morales, M., Epstein, D.H., and Shaham, Y. (2018) Volitional social interaction prevents drug addiction in rat models. *Nat Neurosci*, **21** (11), 1520–1529.

6. Reppucci, C.J., and Veenema, A.H. (2020) The social versus food preference test: A behavioral paradigm for studying competing motivated behaviors in rodents. *MethodsX*, **7**, 101119.

7. Eiselt, A.-K., Chen, S., Chen, J., Arnold, J., Kim, T., Pachitariu, M., and Sternson, S.M. (2021) Hunger or thirst state uncertainty is resolved by outcome evaluation in medial prefrontal cortex to guide decision-making. *Nat Neurosci*, **24** (7), 907–912.

8. Rodgers, R.J., Holch, P., and Tallett, A.J. (2010) Behavioural satiety sequence (BSS): separating wheat from chaff in the behavioural pharmacology of appetite. *Pharmacol Biochem Behav*, **97** (1), 3–14.

9. Tierney, A.J., MacKillop, I., Rosenbloom, T., and Werner, A. (2020) Post-feeding behavior in crayfish (Procambarus clarkii): Description of an invertebrate behavioral satiety sequence. *Physiology & Behavior*, **213**, 112720.

10. Wee, R.W.S., Mishchanchuk, K., AlSubaie, R., and MacAskill, A.F. (2022) Internal state dependent control of feeding behaviour via hippocampal ghrelin signalling. 2021.11.05.467326.

11. Burnett, C.J., Funderburk, S.C., Navarrete, J., Sabol, A., Liang-Guallpa, J., Desrochers, T.M., and Krashes, M.J. (2019) Need-based prioritization of behavior. *Elife*, **8**, e44527.

12. Sotelo, M.I., Tyan, J., Markunas, C., Sulaman, B.A., Horwitz, L., Lee, H., Morrow, J.G., Rothschild, G., Duan, B., and Eban-Rothschild, A. (2022) Lateral hypothalamic neuronal ensembles regulate pre-sleep nest-building behavior. *Current Biology*, **32** (4), 806-822.e7.

13. Silasi, G., Boyd, J.D., Bolanos, F., LeDue, J.M., Scott, S.H., and Murphy, T.H. (2018) Individualized tracking of self-directed motor learning in group-housed mice performing a skilled lever positioning task in the home cage. *J Neurophysiol*, **119** (1), 337–346.

14. Woodard, C.L., Bolaños, F., Boyd, J.D., Silasi, G., Murphy, T.H., and Raymond, L.A. (2017) An Automated Home-Cage System to Assess Learning and Performance of a Skilled Motor Task in a Mouse Model of Huntington's Disease. *eNeuro*, **4** (5), ENEURO.0141-17.2017.

15. Erskine, A., Bus, T., Herb, J.T., and Schaefer, A.T. (2019) AutonoMouse: High throughput operant conditioning reveals progressive impairment with graded olfactory bulb lesions. *PLoS ONE*, **14** (3), e0211571.

16. Torquet, N., Marti, F., Campart, C., Tolu, S., Nguyen, C., Oberto, V., Benallaoua, M., Naudé, J., Didienne, S., Debray, N., Jezequel, S., Le Gouestre, L., Hannesse, B., Mariani, J., Mourot, A., and Faure, P. (2018) Social interactions impact on the dopaminergic system and drive individuality. *Nat Commun*, **9** (1), 3081.

17. Puścian, A., Łęski, S., Kasprowicz, G., Winiarski, M., Borowska, J., Nikolaev, T., Boguszewski, P.M., Lipp, H.-P., and Knapska, E. (2016) Eco-HAB as a fully automated and ecologically relevant assessment of social impairments in mouse models of autism. *Elife*, **5**, e19532.

18. Kiryk, A., Janusz, A., Zglinicki, B., Turkes, E., Knapska, E., Konopka, W., Lipp, H.-P., and Kaczmarek, L. (2020) IntelliCage as a tool for measuring mouse behavior - 20 years perspective. *Behav Brain Res*, **388**, 112620.

19. Kempermann, G., Lopes, J.B., Zocher, S., Schilling, S., Ehret, F., Garthe, A., Karasinsky, A., Brandmaier, A.M., Lindenberger, U., Winter, Y., and Overall, R.W. (2022) The individuality paradigm: Automated longitudinal activity tracking of large cohorts of genetically identical mice in an enriched environment. *Neurobiology of Disease*, **175**, 105916.

20. Watts, A.G., Kanoski, S.E., Sanchez-Watts, G., and Langhans, W. (2022) The physiological control of eating: signals, neurons, and networks. *Physiol Rev*, **102** (2), 689–813.

21. Craig, W. (1917) Appetites and Aversions as Constituents of Instincts. *Proc Natl Acad Sci U S A*, **3** (12), 685–688.

22. Hartmann, C., Borges, V., Thönnes, Y., and Karnani, M.M. (2023) Build instructions for a long term behavioural enclosure for measuring motivational switching in mice. *ResearchEquals*. 10.53962/mn5r-7ekh

23. Jendryka, M., Palchaudhuri, M., Ursu, D., van der Veen, B., Liss, B., Kätzel, D., Nissen, W., and Pekcec, A. (2019) Pharmacokinetic and pharmacodynamic actions of clozapine-N-oxide, clozapine, and compound 21 in DREADD-based chemogenetics in mice. *Sci Rep*, **9** (1), 4522.

24. Alexander, G.M., Rogan, S.C., Abbas, A.I., Armbruster, B.N., Pei, Y., Allen, J.A., Nonneman, R.J., Hartmann, J., Moy, S.S., Nicolelis, M.A., McNamara, J.O., and Roth, B.L. (2009) Remote control of neuronal activity in transgenic mice expressing evolved G protein-coupled receptors. *Neuron*, **63** (1), 27–39.

25. Tervo, D.G.R., Kuleshova, E., Manakov, M., Proskurin, M., Karlsson, M., Lustig, A., Behnam, R., and Karpova, A.Y. (2021) The anterior cingulate cortex directs exploration of alternative strategies. *Neuron*, **109** (11), 1876-1887.e6.

26. Biro, L., Sipos, E., Bruzsik, B., Farkas, I., Zelena, D., Balazsfi, D., Toth, M., and Haller, J. (2018) Task Division within the Prefrontal Cortex: Distinct Neuron Populations Selectively Control Different Aspects of Aggressive Behavior via the Hypothalamus. *J. Neurosci.*, **38** (17), 4065–4075.

27. Tervo, D.G.R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., and Karpova, A.Y. (2014) Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, **159** (1), 21–32.

28. Karnani, M.M., Schöne, C., Bracey, E.F., González, J.A., Viskaitis, P., Li, H.-T., Adamantidis, A., and Burdakov, D. (2020) Role of spontaneous and sensory orexin network dynamics in rapid locomotion initiation. *Prog. Neurobiol.*, **187**, 101771.

29. Bonnavion, P., Mickelsen, L.E., Fujita, A., de Lecea, L., and Jackson, A.C. (2016) Hubs and spokes of the lateral hypothalamus: cell types, circuits and behaviour. *The Journal of Physiology*, **594** (22), 6443–6462.

30. Schwartz, M.W., Woods, S.C., Porte, D., Seeley, R.J., and Baskin, D.G. (2000) Central nervous system control of food intake. *Nature*, **404** (6778), 661–671.

31. Jennings, J.H., Ung, R.L., Resendez, S.L., Stamatakis, A.M., Taylor, J.G., Huang, J., Veleta, K., Kantak, P.A., Aita, M., Shilling-Scrivo, K., Ramakrishnan, C., Deisseroth, K., Otte, S., and Stuber, G.D. (2015) Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell*, **160** (3), 516–27.

32. Stamatakis, A.M., Van Swieten, M., Basiri, M.L., Blair, G.A., Kantak, P., and Stuber, G.D. (2016) Lateral Hypothalamic Area Glutamatergic Neurons and Their Projections to the Lateral Habenula Regulate Feeding and Reward. *Journal of Neuroscience*, **36** (2), 302–311.

33. Li, Y., Zeng, J., Zhang, J., Yue, C., Zhong, W., Liu, Z., Feng, Q., and Luo, M. (2018) Hypothalamic Circuits for Predation and Evasion. *Neuron*, **97** (4), 911-924.e5.

34. Padilla-Coreano, N., Batra, K., Patarino, M., Chen, Z., Rock, R.R., Zhang, R., Hausmann, S.B., Weddington, J.C., Patel, R., Zhang, Y.E., Fang, H.-S., Mishra, S., LeDuke, D.O., Revanna, J., Li, H., Borio, M., Pamintuan, R., Bal, A., Keyes, L.R., Libster, A., Wichmann, R., Mills, F., Taschbach, F.H., Matthews, G.A., Curley, J.P., Fiete, I.R., Lu, C., and Tye, K.M. (2022) Cortical ensembles orchestrate social competition through hypothalamic outputs. *Nature*, **603** (7902), 667–671.

35. Clarke, R.E., Voigt, K., Reichenbach, A., Stark, R., Bharania, U., Dempsey, H., Lockie, S.H., Mequinion, M., Lemus, M., Wei, B., Reed, F., Rawlinson, S., Nunez-Iglesias, J., Foldi, C.J., Kravitz, A.V., Verdejo-Garcia, A., and Andrews, Z.B. (2022) Identification of a stress-sensitive anorexigenic neurocircuit from medial prefrontal cortex to lateral hypothalamus. *Biological Psychiatry*.

Translational Validity of Assessing Cognitive Control and Memory Functions in Nonhuman Primates Using Gamified Tasks

Thilo Womelsdorf[1,2,3], Adam Neumann[1], Nathan Traczewski[1], Seema Dunghana[1], Xuan Wen[1,2], Paul Tiesinga[4]

[1]Department of Psychology, Vanderbilt University, Nashville, TN 37240

[2]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37240

[3]Vanderbilt Brain Institute, Nashville, TN 372404

[4]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen 6525 EN, Netherlands.

Translational validity of cognitive assessments requires humans and animals performing similar tasks with similar cognitive strategies. Achieving translational validity is a major challenge for pre-clinical research seeking to understand the circuit mechanisms underlying higher cognitive functions in humans by using nonhuman primate (NHP) or rodent animal models. Higher cognitive functions require cognitive control and the use of relational memories for inferential reasoning that are often difficult to achieve by animals without extensive training and for which computational validity may not be established (Redish et al. 2021 Computational validity: using computation to translate behaviors across species. Phil. Trans. R. Soc. B 377: 20200525).

Here, we address this challenge by introducing a cognitive assessment approach for NHPs and humans aimed at enhanced translational validity for tasks assessing attentional control, response inhibition, working memory, visuo-spatial problem solving, relational object memory, and motivational effort control.

The assessment approach combines five components to enhance translational validity. First, all tasks are programmed in a unified videogame engine platform 'M-USE' (http://m-use.psy.vanderbilt.edu). M-USE presents tasks on touchscreens requiring identical behavioral interactions from humans and NHPs. Humans use iPads, while for NHPs the touchscreens are embedded in Kiosk stations mounted to their home cages and available on a routine basis [2]. Second, cognitive assessment proceeds with multiple tasks in single testing sessions, which allows same-session comparison of performance metrics in NHP similar to what is standard for human cognitive test batteries. We show that NHPs routinely engage with five and more tasks at stable, high performance levels. Third, tasks are gamified by including 3D renderings, animated task elements and feedback. For example, by including an animated token system for secondary rewards, both humans and NHPs receive frequent, intermittent reinforcement through secondary rewards, which is thought to enhance learning and reduce habitual responding. Fourth, task difficulty is titrated to match the cognitive abilities of individual subjects in order to avoid ceiling and flooring effects. Titration is achieved through adaptive psychophysical staircase procedures that increase the likelihood of reaching a user-defined stable plateau performance level. Fifth, task performance is evaluated in a standardized format that surveys the distribution of errors and performance metrics and can be fed into computational models for estimating latent cognitive variables underlying performance. For example, behavioral learning of learn novel reward rules, spatial path or object sequences are tracked by quantifying rule-abiding and rule-violating errors and perseverations, while cognitive augmented reinforcement learning models quantify learning rates, working memory decay and decision confidence that are variables not directly observable, but critically important when accounting for behavior.

We adopted five components and found that each component contributes to improved quality and scope of cognitive data. For example, (*i*) the same tasks are engaging to be performed by humans and NHPs; (*ii*) relational (sequential) object memory of NHPs is evident at higher performance levels than what would be expected from prior studies; (*iii*) spatial learning can be efficiently studied in high and low performing individuals and allows assessing long-term spatial memory in the same behavioral testing session; (*iv*) an effort control task that is prone to low task compliance because is pushes subjects to their motivational break point can be efficiently studied when interleaved with more rewarding tasks; (*v*) adaptive staircase selection of task difficulty on a trial-by-trial basis can shorten the duration of assessing response inhibition in an accuracy based anti-saccade task; and (*vi*) initial computational modeling suggests that the same computational model mechanism accounts for performance across subjects, which promises high computational validity of cross-subject cognitive assessment.

In summary, we delineated an approach for evaluating higher cognitive control and memory functions with enhanced translational validity in NHPs and humans. Our approach is versatile, allowing routine assessment of multiple cognitive functions in NHPs in the same sessions using gamified tasks at adaptive difficulty levels. By standardizing analysis across tasks and species, cognitive assessment is quantified objectively. By improving translational validity of pre-clinical research the proposed approach has the potential to enhance the efficiency of diagnosing circuit dysfunctions and developing treatment strategies.

Ethics Statement

All animal and experimental procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals, the Society for Neuroscience Guidelines and Policies, and approved by the Vanderbilt University Institutional Animal Care and Use Committee (M1700198-01)

References

1. Watson M, Traczewski N, Dunghana S, Banaie Boroujeni K, Neumann A, Wen X, Womelsdorf T (2023) A multi-task platform for profiling cognitive and motivational constructs in humans and nonhuman primates. 1-48; bioRxiv; https://www.biorxiv.org/content/10.1101/2023.11.09.566422v1

2. Womelsdorf T, Thomas C, Neumann A, Watson MR, Banaie Boroujeni K, Hassani SA, Parker J, Hoffman KL. (2021) A Kiosk Station for the Assessment of Multiple Cognitive Domains and Cognitive Enrichment of Monkeys. Frontiers in Behavioral Neuroscience. 15. 1-13 https://doi.org/10.3389/fnbeh.2021.721069.

Sniffing Out Pigs' Immediate Perception of Novel Odours (Pig Odour Hedonics)

R. Grut[1], J. Stenfelt[1] & M.V. Rørvang[1]

[1]Dept. Biosystems and Technology, Swedish University of Agricultural Sciences (SLU), Lomma, Sweden
mariav.rorvang@slu.se

Abstract

Their well-developed olfactory system suggests that odours play an important role in the life of domestic pigs. Yet, we know little about how naïve pigs perceive, interpret and respond to odours. As part of a larger project on odour exploration in pigs, we developed a scale to assess immediate behavioural reactions to novel odours. We divided 184 growing-finishing pigs into 92 opposite-sex pairs of pigs from the same litters. All pig pairs were presented with three out of a total of twelve experimental odours (essential and synthetic odour oils). The pigs were presented with their three odours three times in a row (n=9 trials) alongside an odourless control (distilled water) in a habituation/dishabituation test. We recorded the immediate behavioural reactions to both odours and controls on a scale from 1-7, based primarily on the pig's distance to the odour. Exploratory behaviours were represented on the lower end of the scale (score 1-3) and avoidance behaviours on the upper end of the scale (score 5-7). The median (score 4) was considered a neutral response. In total, 88 pigs did not approach at all (score 0, i.e., excluded). The most observed score for odour was 1, meanwhile, for control the most frequently observed score was 7. Score 6 was the least observed in both treatments. The scale provides a first step to understanding how pigs might perceive and respond to novel odours and could be adapted to assess immediate reactions to other types of appetitive or aversive sensory input.

Introduction

The domestic pig is renowned for its great sense of smell, and its highly developed olfactory system suggests that odours play an important role in how pigs interpret and navigate their environment. Olfactory cues are likely to provide pigs with important information regarding e.g., mate selection, development and direction of parental behaviour, group cohesion, and food-seeking [1]. Yet, we know little about odour exploration in pigs in general, and how naïve pigs perceive, interpret, and respond to olfactory stimuli. From human research, it is well established that certain odours are associated with certain degrees of pleasantness or unpleasantness a phenomenon known as odour hedonics [2]. Odour hedonics in humans are relatively easy to study as the experimenter can simply ask, and the human test subject can easily convey his/her thoughts. In animal research, it is not as straightforward, and to our knowledge, there has been no attempt to map pigs' immediate behavioural reaction as an analogue to human odour hedonics. Understanding how pigs perceive olfactory stimuli when first encountered may enable us to assess how (and if) they attribute valence to different odours. As part of a larger project on odour exploration in pigs, we aimed to develop a scale to assess the immediate behavioural reactions of pigs exposed to novel, natural odours.

Method

Experimental design

The study was conducted at the Swedish Livestock Research Centre in Uppsala, Sweden, and included 184 growing-finishing Hampshire-crossed pigs, divided into 92 opposite-sex pairs of pigs from the same litters. The testing of pairs was done to avoid stress from social isolation, and the pairing of littermates already familiar with each other minimized the risk of mixing aggression. Twelve experimental odours were grouped into four triads (i: vanilla, aniseed, and blood orange; ii: musk, apple, and cinnamon bark; iii: ginger, pine, and jasmine; and iv: cedarwood, thyme, and lavender) and all pig pairs were randomly assigned one of the triads (i.e., three odours) to limit the risk of loss of motivation to explore or sniffing fatigue. The experimental odours consisted of either essential oils or synthetic odour oils, which were chosen based on their origin (herbs, spice, tree, root, fruit, flower, seed or mixture). To standardise the odour samples, 6 drops of odour oil were applied to a piece of filter paper which was taped to the inside of a plastic container specific to each odour. Odourless control samples were prepared in the same manner but with 6 drops of distilled water. The samples were prepared right before each test session started and the odour was contained by an airtight lid to preserve the intensity and limit odour drift and contamination.

Test procedure

The pig pair was tested in experimental pens which had two insertion holes (Ø: 14cm; 55cm apart) for odour and control drilled in the front fixture. The holes were large enough that the pigs could put their snout through them and into the container on the outside of the pen fixture (see Figure 1, [3]), but did not allow the pigs to reach the filter paper. This method of presentation allowed the pigs to investigate the samples without coming in direct contact with the odour oil. The holes used for odour and control remained the same throughout each test session (i.e., for each pig pair) to avoid odour contamination of the control hole, but were balanced across pig pairs exposed to the same odour triad to control for any side-biases. The pigs were presented with their three odours three times in a row (n=9) alongside the odourless control in a habituation/dishabituation test [4]. The first presentation of each odour lasted for 1 minute, starting when the first pig had approached either the odour or control, and was followed by a 2-minute pause before a new 1-minute presentation started. After three consecutive presentations of odour number 1, odour number 2 was presented. The order of the odours in the triad (number 1-3) was balanced across the pig pairs in each group.

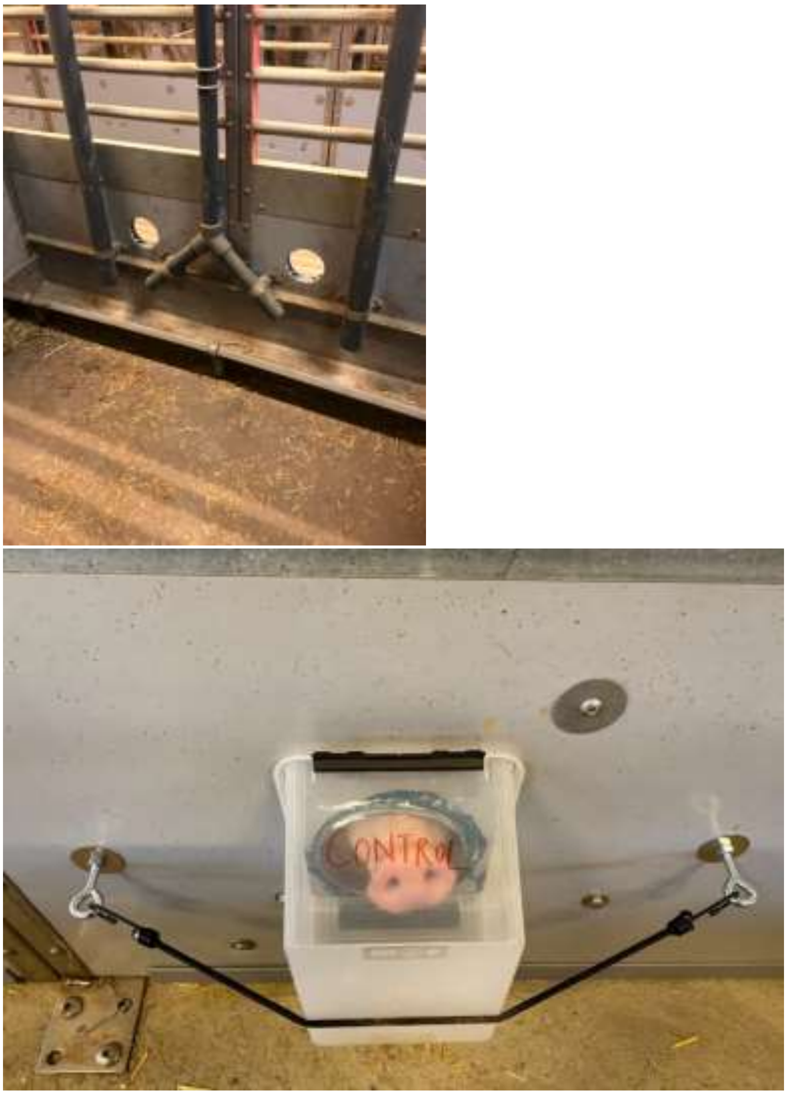


Figure 1. The left picture shows the modified experimental pen with two insertion holes. Reprinted from Rørvang et al. [3]. The right picture shows the odour and control containers, here with the control sample. Reprinted from Rørvang et al. [5]

Reaction scale

The reaction scale was developed by two observers using a subset of the recorded test sessions, and validated by three observers scoring the same test sessions. The third observer who was not involved in the development of the scale and who was further blind to the odours of each test session did the final scoring of all test sessions. A pig's immediate behavioural reaction to each odour was observed 6 seconds after it first approached the sample, both for odours and controls. A pig "approaching" was defined as the pig having its snout within 12 cm of the sample container. The reaction of each individual pig was scored on a scale from 1-7 (see Table 1), primarily based on the pig's distance and orientation to the sample. Exploratory behaviours such as sniffing and oral manipulation were represented on the lower end of the scale (score 1-3). Avoidance behaviours such as headshaking or increasing distance to odour were represented on the upper end of the scale (score 5-6) with no approach at the very end (score 7). The median of the scale (score 4) was considered a neutral response. Pigs that could not be presumed to have detected the odour (e.g. due to sleeping or not being close enough to insertion holes) were scored 0 and excluded from further analysis.

Table 1. Ethogram used to determine pigs' immediate behavioural reaction to a novel odour on a scale of 1-7, adapted from Grut [6].

Reaction score	**Description**
1 – Sniff	The pig has its snout inserted through the insertion hole.
2 – Oral	The pig is within 12 cm of the insertion hole, and makes direct contact with the hole fixture or pen wall (e.g., bites/licks/roots). May include rubbing aimed against the insertion hole.
3 - Approach	The pig is within 12 cm of the insertion hole, but does not make direct contact with the hole fixture or pen wall. May include snout above the insertion hole, but not above pen bars.
4 – Turn	The pig has approached and turned < 90° away from the insertion hole, and is no longer within 12 cm. May include moving ≤ 3 steps.
5 – Leave	The pig has approached and turned ≥ 90° or moved ≥ 4 steps away from the insertion hole and is no longer within 12 cm.
6 - Headshake	The pig shakes its head at least two times from side to side in a sudden movement. May include sniff, oral, approach, turn or leave.
7 – No approach	The pig has had its snout within the front 50% of the solid floor while facing the front pen wall but has not approached within 12 cm of the insertion hole.
0 - No observation	The pig has not had its snout within the front 50% of the solid floor while facing the front pen wall. May include pigs out of camera range (i.e. on the slatted area of the pen)

Data editing and statistical analyses

As an immediate reaction to a novel odour only can be recorded once after which the odour is no longer novel, the data set consisted of six reaction scores per pig, one for each of the three odours and one for each control presented alongside the odour samples. The data was compiled in Microsoft Excel and imported in R (version 4.2.3) [7] using RStudio interface (version 2023.3.0.386) [8] where descriptive analyses were performed by using built-in functions. The frequency of each recorded score was analysed. This resulted in a total of 1104 presentations of which 68 were lost due to camera errors. Thus, the final data set consisted of a total of 1036 reaction scores, 518 for odour and control presentations respectively.

Results

Out of the 1036 presentations where a reaction was scored, a total of 88 were given a score of 0, i.e., no observation, and were therefore excluded. The 88 excluded presentations consisted of 44 odour and 44 control presentations. Overall, for both the odour and control, scores 1 and 7 were most frequently recorded, with a total of 327 recordings of score 1 and 251 recordings of score 7 (see Figure 2).

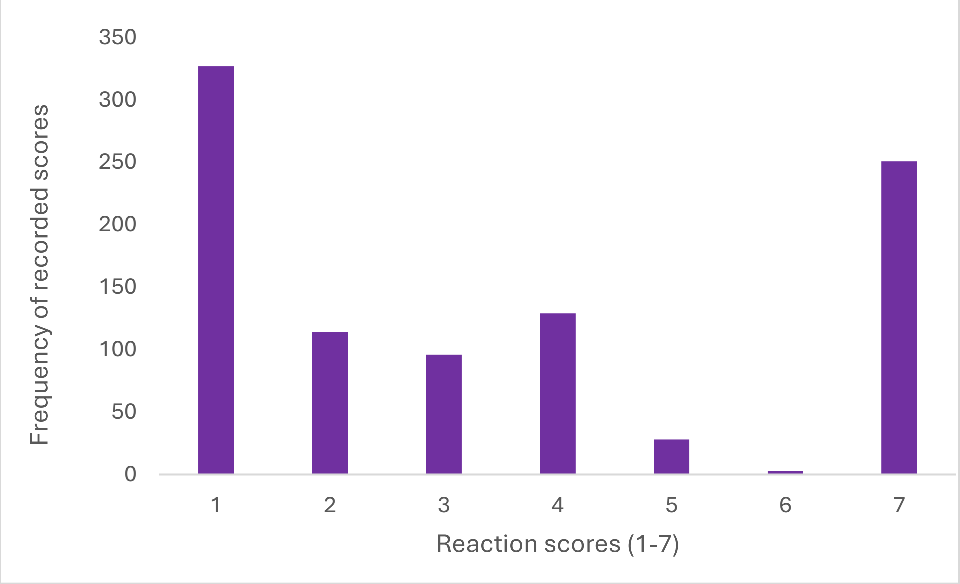


Figure 2. The frequency of observed reactions for each recorded score (1-7) across the entire sample. Observations of score 0 (i.e., no observation) were excluded.

In general, the pigs reacted with more exploratory behaviours (having lower reaction scores: 1-3) towards the odour compared to the control insertion hole (see Figure 3). Overall, scores on the ends of the scale (scores 1 and 7) had the greatest number of recordings. Score 1 was more frequently recorded for odour presentations (215 observations) than for control presentations (112 observations). Conversely, score 7 was more frequently recorded for control presentations (162 observations) than odour presentations (89 observations). For both the control and odour, score 6 was the least frequently recorded reaction score. Score 6 involved head shaking, which was only observed once at a control insertion hole and twice at an odour insertion hole.

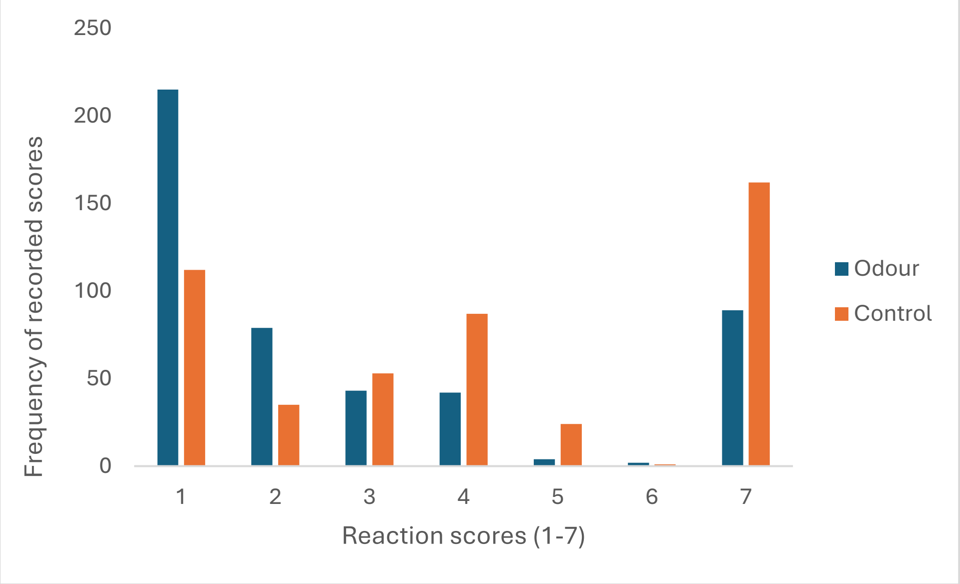


Figure 3. The frequency of observed reactions for each recorded score (1-7) across the entire sample and the distribution between reactions to odour and control presentation. Observations of score 0 (i.e., no observation) were excluded.

Discussion

The aim of this study was to develop a reaction scale that can describe the immediate behavioural reaction of naïve pigs to odours of non-social origin. The observed pigs exhibited a more avoidant behaviour, resulting in higher reaction scores, towards the control compared to the odours. This shows that pigs were more interested in the odours compared to the odourless control, to which the pigs showed more avoidance behaviour (higher reaction scores). This finding is supported by the overall results presented in the article by Rørvang et al. [5], as the odour insertion holes had lower scores as well as longer sniffing durations than the control insertion holes.

On a total of 88 occasions, no reaction from the pigs was observed (score 0). This score included both the lack of reaction from the pigs and instances of the pigs being outside the camera frame. The lack of reactions could be due to the pigs finding neither the odour nor control interesting, or that they did not notice the odour/control being presented, for instance, due to being asleep.

The preliminary result of this work shows that the 7-point scale was successful in describing explorative, neutral and avoidance behaviours that accurately reflect the valence of the pigs' reactions. To our knowledge, this is the first example of a scale developed specifically to assess pigs' immediate behavioural reaction to novel olfactory stimuli, and it provides a first step to understanding how pigs might perceive and respond to novel odours in a commercial production environment. In future research, the scale could be adapted to assess immediate reactions to other types of appetitive or aversive sensory input, and could potentially be further adapted to accommodate similar investigations into olfactory perception and behavioural reactions of other species with well-developed olfactory systems.

Ethical statement

The procedures and details of the experiment were assessed and approved by the Board of Ethical use of Animals in teaching and Research, Swedish University of Agricultural Sciences, Uppsala, Sweden, prior to the experiment. An ethical permit was obtained from the Swedish Board of Agriculture, Uppsala, Sweden ID number: 5.2.18-02900/2020.

References

1. Schild, S.-L.A. & Rørvang, M.V. (2023). Pig olfaction: the potential impact and use of odors in commercial pig husbandry. *Frontiers in Animal Science*, **4**:e1215206.

2. Rouby, C., Pouliot, S. & Bensafi, M. (2009). Odor hedonics and their modulators. *Food Quality and Preference,* **20**(8):545–549.

3. Rørvang, M.V., Schild, S.-L.A., Wallenbeck, A., Stenfelt, J., Grut, R., Valros, A. & Nielsen, B.L. (2023). Rub 'n' roll – pigs, Sus scrofa domesticus, display rubbing and rolling behaviour when exposed to odours. *Applied Animal Behaviour Science,* **266**:e106022.

4. Yang, M. & Crawley, J.N. (2009). Simple behavioural assessment of mouse olfaction. *Current Protocols in Neuroscience,* **48**(1):8.24.1-8.24.12.

5. Rørvang, M.V., Schild, S.A., Stenfelt, J., Grut, R., Gadri, M.A., Valros, A., Nielsen, B.L. & Wallenbeck, A. (2023). Odor exploration behavior of the domestic pig (Sus scrofa) as indicator of enriching properties of odors. *Frontiers in Behavioural Neuroscience,* **17**:e1173298.

6. Grut, R. (2022). Pigs' immediate behavioural reaction when exposed to scent enrichment. [Thesis, Uppsala: SLU, Department of Biosystems and Technology].

7. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. < https://www.r-project.org/ >. Accessed 30 March 2024.

8. Posit Team (2023). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. < http://www.posit.co/ >. Accessed 30 March 2024.

Measuring Episodic Memory in Preschool Children

Tereza Nekovářová[1, 2, 3], Petra Skalníková[1, 2], Petra Eretová[1], Jiří Pešek[1, 3], Martina Píšová[2], Lukáš Hejtmánek[1], Iveta Fajnerová[1]

1: National Institute of Mental Health, Klecany, Czech Republic

2: Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic

3: Department of Psychology, Faculty of Arts, Charles University, Prague, Czech Republic

Background

Memory is a central part of cognition. Episodic memory is the key for remembering personal experiences and for building continuity within ourselves. Episodic memory is usually defined as long-term declarative memory system allowing conscious recollection of an event and its context [1]. Despite its essential role in human cognition, the concept of episodic memory is surprisingly difficult to define.

The term 'episodic memory' was originally proposed to distinguish personally experienced events from general facts (semantic memory) [2, 3]. Traditionally, the terms 'episodic' and 'autobiographical' memory have been used synonymously in the literature, but some authors have questioned the assumption of equivalence between these types of memory and formulated different concepts (e.g. [4, 5, 6]). This implies that autobiographical memory is either synonymous with, or a specific case of, episodic memory.

Episodic memory was originally described as a mental time travel [2, 3], so it involves remembering by re-experiencing episodes from memory and allows us to be aware of the continuity of self in the time. According to this view, episodic memory refers to the conscious recollection of a personal experience. Recollection from episodic memory also implies a first-person subjectivity (i.e. autonoetic consciousness). Autonoetic consciousness (self-awareness), which has been proposed to be driven by the frontal lobes and possibly the parietal cortex [6], is essential for many of the most complex cognitive abilities, including mental time travel.

Conceptualisations of episodic memory that are difficult or impossible to apply to non-human animals have resulted from the focus on unique phenomenal experiences in human memory, such as the subjective re-experiencing of an event (the sense of "having been there") [7]. To bridge this gap, simplified models of episodic/autobiographical memory have been proposed, such as the WWW ("what", "where" and "when") model [8]. Therefore, this approach can also be applied to preverbal infants and preschool children. However, a problem in Tulving's definition [9] of episodic memory is that it combines two distinct components - contextual nature of episodic memory (contains specific information about "what", "where" and "when") and "autonoetic awareness" component (connected with mental time travel).

We aim to study the development of episodic memory in ontogeny - in terms of two main concepts of episodic memory - the first one based on the experience of the self in the remembered event (1st person memory perspective, for which we will use the term "self-perspective" or "autobiographical episodic memory") and the second one based on contextual cues (3rd person perspective - we will use the term "contextual episodic memory").

As noted above, autobiographical memory (i.e. 1st person perspective memory) can be distinguished from simpler 'contextual' episodic memory (3rd person perspective) by the personal or subjective perspective taken on the event and its representation in memory. Traditionally, it has been suggested that during the first 5-7 years of life, the ability to create a fully autobiographical type of memory is limited or even absent [10]. One of the possible reasons for the late emergence of autobiographical memory is that there is not a fully developed self-concept around which memory can be organised. Alternatively, it is suggested that the reason is a limited subjective self. Without these components there can be no full autonoetic consciousness and therefore no autobiographical memory (for a review see e.g. [11]). Nelson and Fivush [12] suggested that it is not until the age of 5 that children fully develop the cognitive abilities needed to encode, store and retrieve autobiographical memories. These abilities include: self-concept, language and narrative, theory of mind, sense of time and place, sense of self, autonoetic awareness, etc. They proposed that up to this age children may have semantic or episodic (contextual) memory, but not true autobiographical memory.

Contrary to this assumption, there are studies suggesting that there may even be early memories of a self-referential nature. Not only can children remember the "what", "where" and "when", but they can also spontaneously verbalise the "who". According to these observations, infants develop a more self-oriented (1st person) perspective during the preschool years (for a review, see [10]).

Here, we would like to present a battery for testing the development of episodic memory in the sense of autobiographical memory, using self-perspective and contextual episodic memory.

Children are tested in two subsequent sessions (after 7-10 days) to test recall of episodic information after prolonged delay, and also to make testing less difficult and less demanding for concentration and mainly.

To address ontogenetic development of possible different memory systems (autobiographical vs. contextual episodic system) and to compare their development with developmental trajectories of another cognitive functions (e.g. verbal memory, self-representation) in children we will use the following standard methods:

Memory for sentences (NEPSY Test Battery) - assessing verbal memory for particular sentences not connected into a meaningful story (episode).

Narrative memory (memory for verbal episodes) (NEPSY Test Battery) - tests memory for a single episode situated in spatial and temporal context, but without the “self-involvement”. We use it to test episodic memory (in means of contextual memory, not autobiographical memory per se).

Several behavioural methods will be applied to address episodic memory processes in ecologically valid situations:

“Hide and seek task”: children participate in the game, after a week delay they are asked for some details from this event (e.g. who participated in a game, where participants were hiding etc.). The game is presented either in 1st person perspective - autobiographical memory task (children play hide and seek game) or in 3rd person perspective - contextual memory task (children watch someone else - toys and puppets -play hide and seek game). This task allows us to compare autobiographical memory (1st person perspective - “my own experience”) with a comparable episode (and to ask for the similar information) but without reference to self (“3rd person perspective”). We will test the developmental trajectory of both types of memory systems.

The virtual episodic memory task (vEMT) - The vEMT was designed to test the episodic-like memory model using a simple paradigm [13]. Participants are asked to collect everyday objects in a virtual household and remember their identity, position, and temporal order. This task was created as a three-dimensional adaptation of the original two-dimensional computer task previously used to demonstrate impaired episodic memory in elderly with the amnestic form of MCI and AD patients [14]. Each trial consists of three: A) Acquisition: collection of objects; B) Item recognition: selection of objects from a set; and C) WHEN-WHERE Recall: returning all items to their original positions in a given sequence. Outcome measures: error rates for object recognition; temporal order and spatial position errors; time and trajectory to finish individual trials. The original vEMT task was slightly modified for the purposes of assessment in children. The set of objects used as task items in the task for adults have been replaced by a set of daily objects and toys. The task procedure is identical, however a number of objects above 7 will not be tested in recruited children.

This test battery allows us to test the development of different aspects of episodic/autobiographical memory, and to compare the development of episodic memory from a 1st and 3rd person perspective. The addition of other tasks will allow us to observe the developmental relationship with other types of memory.

Ethical statement

A signed written informed consent (approved by the Ethical Commission of (blinded for the review) was obtained before the beginning of the study from parents, and each child was verbally asked if he/she was willing to participate.

Acknowledgment

This study was supported by the institutional support from the Ministry of Health Czech Republic – DRO 2021 ("National Institute of Mental Health – NIMH, IN: 00023752"), and by the grant of Czech Science Foundation (no. GACR22-10493S).

References

1. Allen T.A., Fortin N.J. (2013). The evolution of episodic memory. Proc Natl Acad Sci., 110: 10379–86.
2. Tulving, E. (1972). Episodic and semantic memory. Organization of memory, 1, 381-403.
3. Tulving, E. (1982). Synergistic ecphory in recall and recognition. Can. J. of Psych., 36(2), 130.
4. Gilboa, A. (2004). Autobiographical and episodic memory—one and the same?: Evidence from prefrontal activation in neuroimaging studies. Neuropsychologia, 42(10), 1336-1349.
5. Conway, M. A. (2001). Sensory-perceptual episodic memory and its context: Autobiographical memory. Philosophical Transactions of the Royal Society of London Series B: Bio Sci, 356(1413), 1375–1384.
6. Wheeler, M. A., Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. Psychology Bulletin, 121(3), 331–354.
7. Templer, V. L., & Hampton, R. R. (2013). Episodic memory in nonhuman animals. *Current Biology, 23*(17), R801-R806.
8. Clayton, N. S., and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. Nature 395, 272–274.
9. Tulving, E. (2002). Episodic memory: From mind to brain. Annual review of psychology, 53(1), 1-25.
10. Bauer, P. J. (2015). Development of episodic and autobiographical memory: The importance of remembering forgetting. Developmental Review, 38, 146-166.
11. Howe, ML. The co-emergence of the self and autobiographical memory: An adaptive view of early memory. In: Bauer, PJ.; Fivush, R., editors. The Wiley-Blackwell Handbook on the Development of Children's Memory. West Sussex, UK: Wiley-Blackwell; 2014. p. 545-567.
12. Nelson K, Fivush R. The emergence of autobiographical memory: A social cultural developmental theory. Psychological Review. 2004; 111:486–511. [PubMed: 15065919]
13. Fajnerová, I., Oravcová, I., Plechatá, A., Hejtmánek, L., Sahula, V., Vlček, K., & Nekovářová, T. (2017). The virtual Episodic Memory Task: Towards remediation in neuropsychiatric disorders. In 2017 International Conference on Virtual Rehabilitation (ICVR) (S. 1–2). IEEE.
14. Vlcek K., Laczo J., Vajnerova O., Ort M., Vyhnalek M., Hort J. (2009) Impairment of patients with non-amnestic MCI in a novel episodic-like memory test. Psychiatrie. 13(4):211-215.

From Labs To Zoos To The Field: Insights In To Mammalian Whisker And Avian Rictal Bristle Behaviour

R. A. Grant[1] and M. G. Delaunay[1]

1Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK. Robyn.grant@mmu.ac.k

Tactile feelers are present on the faces of many animals, including whiskers in mammals, bristles in birds, barbels in fish and antennae in arthropods. Probably the most similar to one another are the whiskers of mammals and the bristles of birds, since they are ultimately made up of dead cells, and it is the follicle that contains the sensitive mechanosensors and musculature. Whiskers and bristles also have a similar function, they are both tactile sensors, and are more sensitive, moveable and prominent in animals that forage in dark, complex habitats, such as rodents [1], pinnipeds [2] and Caprimulgiformes (nightjars and their relatives) [3]. The most well-documented species is the laboratory rat [1], and many studies have described how rats explore objects with their whiskers [4,1] as well as how these behaviours develop [5,1]. Indeed, in rats, whiskers are an important sense from birth, and even moveable from day 2 [5]. However, there are a lot of things we do not know about tactile facial sensing in other species, even in mammals and birds.

We will present here two studies (Figure 1). Both studies have been ethically approved by the committee at Manchester Metropolitan University, and the local panels of each collaborating zoo institution. Firstly, we show how studying novel object exploration behaviour in rodents has led us to define whisker control behaviours, such as whisker spread reduction and asymmetry (Figure 1.1, top panel, see [6] for more details). We designed a comparative version of this task for zoos to study 16 mammalian species in varying sized species, from harvest mice to harbour seals. We will present this new task and the challenges associated with conducting comparative behavioural work in a zoo setting, including having to consider differences in species colouration, size, trainability, shyness, institution and the enclosure design. We then went on to adapt this study for fieldwork, specifically on grey seals to characterise their natural whisker movements and contact types in the wild. Preliminary observations suggest that object-related whisker behaviour, including protractions, spread alterations and asymmetric positioning, can all be seen in the lab, zoo and field irrespective of whether the object is natural (e.g.. rock) or artificial (e.g. guttering). However, whisker contacts in the field involve a lot more social interactions, than those observed in the zoo or lab.

The second study builds on classic research on rat development, characterizing behavioural responses to whisker stimulation [5]. We applied an amended version of this task to a completely novel setting- that of rictal bristle responses during development in tawny frogmouth chicks in zoos, as well as barn owls, common nighthawks and whip-poorwhills in the field (Figure 1, bottom panel). Overall, we found that, unlike whiskers, rictal bristles are not present at birth, and their emergence coincides with fledging and independent feeding, which supports their role in foraging in older birds. We further discuss flexible working in the zoo and the field, especially when handling and stimulating chick bristles.

The ability to apply lab observations to a zoo and field setting has many benefits. Studying complex behaviours across different species means we can start to answer questions about evolution and function. Furthermore, using lab studies for inspiration helps us to create tasks based on specific observations, which allows us to make more detailed and focused studies in the field. Studying natural whisker-related and bristle-related behaviours in the wild, gives us greater insights into animal behaviour, as well as giving us the ability to study natural whisker contacts, which has implications for designing enclosures and enrichment toys for animals in captivity. However, working in the field and in zoos can be challenging, and is considerably different from the controlled and detailed laboratory studies that many of us are used to. We will explore these challenges in our talk and make recommendations for further studies of this nature.

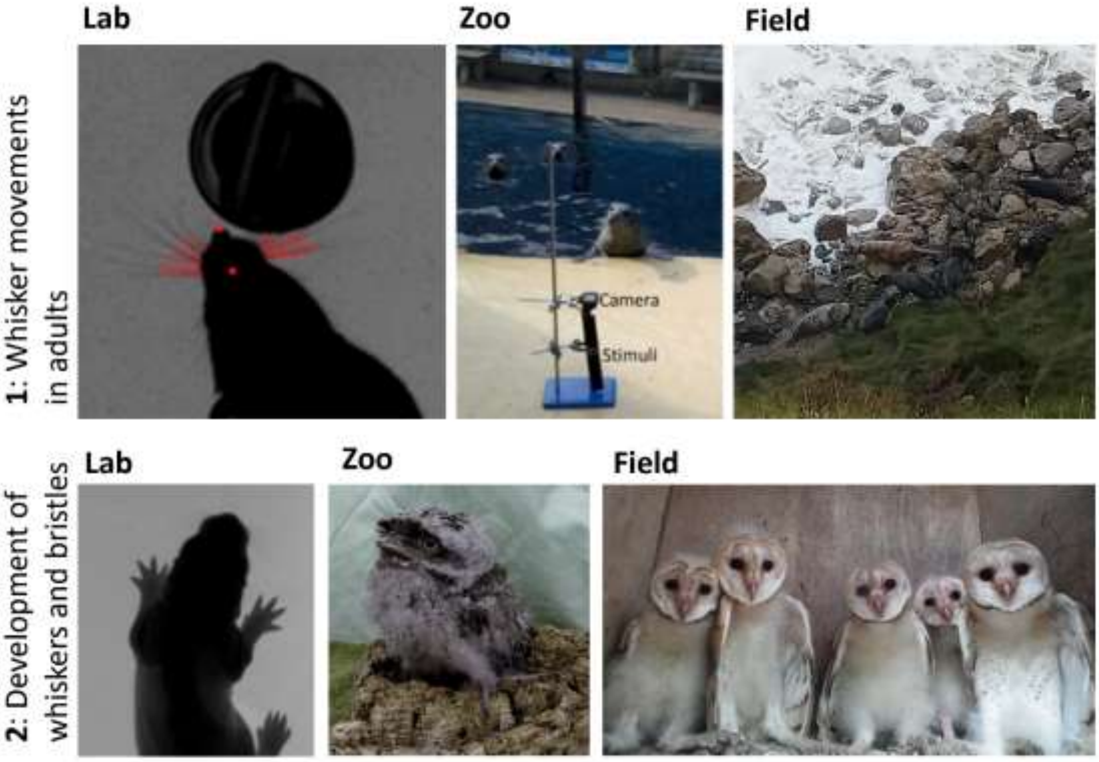

Lab
Zoo
Field
1: Whisker movements in adults
Camera
Stimuli
Lab
Zoo
Field
2: Development of whiskers and bristles

Figure 1 Diagram of presented methods from lab in rats, and then applied to zoo and field settings. Top panel (1.1) shows the first study we will present, investigating whisker movements in adults, especially in response to objects. Shown here in laboratory rat (1.1 left), using a high-speed video camera filming above novel object interactions, with whiskers tracked in red; in harbor seals (1.1 middle) at the zoo, using an action camera filming above a novel object; and in Grey seals in the field, contacting conspecifics and natural objects in the environment with their whiskers. Bottom panel (1.2) shows the second study we will present, including a rat pup in the lab (1.2 left); Tawny frogmouth chick in the zoo (1.2 middle); and a field study with barn owl chicks (1.2 right).

References

1. Grant, R.A. and Goss, V.G.A. (2021). What can whiskers tell us about mammalian evolution, behaviour and ecology? *Mammal Review* **52**, 148-163.

2. Milne, A.O., Muchlinski, M.N., Orton, L.D., Sullivan, M.S., and Grant, R.A. (2021). Comparing vibrissal morphology and infraorbital foramen area in pinnipeds. *The Anatomical Record* **305**, 556-567.

3. Delaunay, M., Larsen, C., Lloyd, C., Sullivan, M., and Grant, R.A. (2020) Anatomy of avian rictal bristles in Caprimulgiformes reveals reduced tactile function in open habitat, partially diurnal foraging species. *Journal of Anatomy* **237**, 355-366.

4. Grant, R., Mitchinson, B., Fox, C., & Prescott, T.J. (2009). Active touch sensing in the rat: Anticipatory and regulatory control of whisker movements during surface exploration. *Journal of Neurophysiology* **101**, 862-874.

5. Grant, R.A., Mitchinson, B., & Prescott, T.J. (2012). The development of whisker control in rats in relation to locomotion. *Developmental Psychobiology* **54**, 151-168.

6. Simanaviciute, U., Brown, R.E., Wong, A., Fertan, E. & Grant, R.A. (2022). Abnormal whisker movements in the 3xTg-AD mouse model of Alzheimer's disease. *Genes, Brain and Behavior*, 21, e12813.

Developing novel whisker movement tests to examine object-related exploration and habituation in Reeler mice

Ugne Simanaviciute[1]*, Emma Hodson-Tole[2], Andrew Spink[3], Robyn Grant[1]

1. Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK

2. Department of Life Sciences, Research Centre for Musculoskeletal Science & Sports Medicine, Manchester Metropolitan University, Manchester, UK

3. Noldus Information Technology BV, Wageningen, The Netherlands

*** Contacting author: ugne.simanaviciute@stu.mmu.ac.uk**

The current battery of tests used for assessing rodent behaviour consists of either expensive and intrusive methods or requires extensive animal training. They also often result in only simple behavioural measures, such as durations or frequencies. Our lab has developed a protocol to measure whisker-related exploratory movements and has shown that they offer an alternative way to measure highly quantitative behavioural changes in mice. We have previously demonstrated this by assessing whisker movement differences in mouse models of a wide range of neurological disorders [1] and shown that whisker movements can detect behavioural deficits in some mouse models, as early or earlier than any other behavioural measure [2, 3], which is especially useful in complex and subtle phenotypes [2]. We chose to investigate homozygous Reeler mice here, since they have a complex behavioural phenotype, and few findings have shown strong behavioural and cognitive deficits, despite their neuroanatomical disruptions [4]. These studies have been ethically approved by the committees at Manchester Metropolitan University and the University of Göttingen. Using our standard whisker task [1] (Figure 1a), we found few differences in Reeler mice whisker movements, compared to the wildtype controls. While we saw some differences in whisker spread and asymmetry, with Reeler mice not decreasing whisker spread and increasing asymmetry as much as we might expect following object contact, these were only slight differences in the during-contact measures. As we would usually only examine changes from pre-contact to during contact [2], this standard analysis did not reveal any significant deficits in Reeler mice. Additionally, these slight changes in whisker positions on the object were only present in Reeler mice that had not previously been exposed to the experimental arena before, suggesting that habituation, or novelty, might also have an effect on their behaviour. Furthermore, we observed that the Reeler mice did not often contact the novel object, resulting in reduced sample sizes for contact-related measures, which limited our findings. Therefore, overall, we saw that i) Reeler mice whisker movements might be affected by exposure to the arena; and ii) Reeler mice were likely avoiding the object.

Therefore, we then went on to develop two novel tasks to explore this in more detail. To address the first observation, we filmed whisker movements with a high-speed 500 fps camera firstly in an open field environment and then during a further habituation period, where we analysed their whisker movements in the open field, and in the first and fifth habituation sessions. We found that habituation affected both the wildtype and Reeler mice with mean angular position, spread and locomotion speed all decreasing over the habituation period. This means that the animals were moving slower, holding their whiskers further back and more spread out, which might suggest that the animals were not as focused on the area ahead of themselves as they got familiar with the environment [5]. Whisker angular position in Reeler mice decreased systematically between the open field, first and fifth habituation period (Figure 1b).

To address the second observation, an infrared 30 fps camera was paired with the high-speed camera to record the whole object exploration experiment, rather than just the small periods of exploration that we usually capture. This new technique showed that while the Reeler mice spent the same length of time around the object, when they got close to the object, they did not touch it and initiate whisker exploratory behaviours (Figure 1c). Therefore, we were unable to use the high-speed video clips in our usual contact-related measure analyses. We posit that this object avoidance might be due to their neuroanatomical changes, especially in the motor cortex [6], which can impact the initiation of sequences of behaviours.

Using these novel observations, we suggest that whisker movements are a powerful behavioural measurement tool. However, making observations in a standardized manner might not suit all mouse models if their impairment does not allow for a close contact with the object. Incorporating whisker measurements within a further battery of behavioural tests allows detection of additional exploratory-related parameters. Therefore, further work should consider incorporating whisker movements into the usual behavioural tests to complement standard tasks that assess object exploration and habituation in particular.

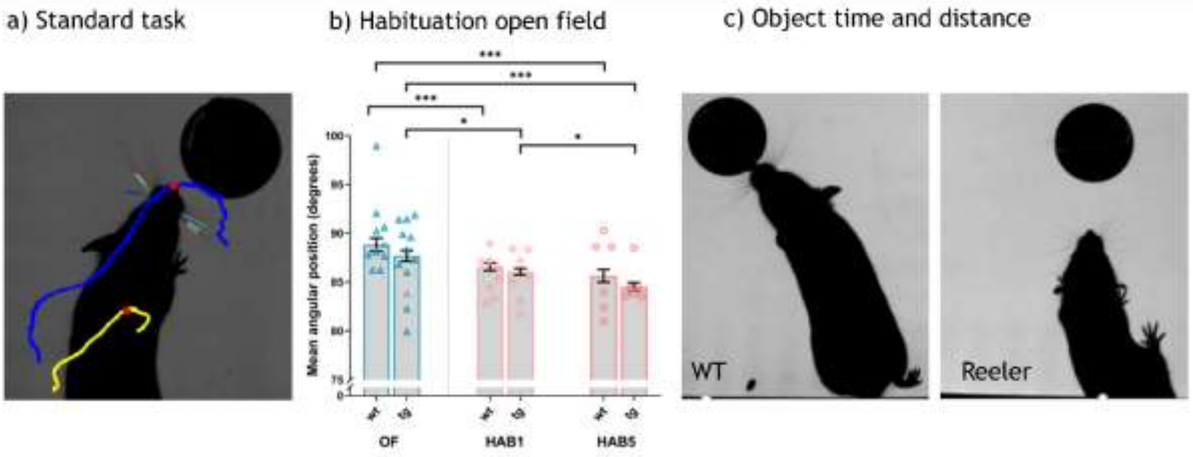


Figure 1. Summary figure. Panel a shows our standard experimental set-up with an object and the tracking, with the nose tracking in blue, body centroid position in yellow, and the whiskers detected in mulitcolour. Panel b shows the results from the habituation study, with the significant findings indicated with asterisks (* p<0.05, *** p<0.001). Panel c shows the difference in distance to the object for the wildtype (WT) and Reeler individuals.

References

1. Simanaviciute, U., Ahmed, J., Brown, R.E., Farr, T.D., Fertan, E., Garland, H., Morton, J.A., Staiger, J.F., Skillings, E.A., Trueman, R.C., Wong, A.A., Wood, N.I. and Grant, R.A. (2019) Measuring whisker movements and locomotion to describe motor and exploratory behaviours in mouse models. *Journal of Neuroscience Methods* **300**, 103-111.

2. Simanaviciute, U., Brown, R., Wong, A., Fertan, E. and Grant, R.A. (2022). Abnormal whisker movements in the 3xTg-AD mouse model of Alzheimer’s Disease. *Genes, Brain and Behavior* **21**, e12813.

3. Garland, H., Wood, N.I., Skillings, E.A., Detloff, P.J., Morton, A.J. and Grant, R.A. (2017). Characterisation of progressive motor deficits in whisker movements in R6/2, Q175 and Hdh knock-in mouse models of Huntington’s disease. *Journal of Neuroscience Methods* **300**, 103-111.

4. Guy, J., Staiger, J.F., 2017. The Functioning of a Cortex without Layers. Front Neuroanat 11, 54. https://doi.org/10.3389/fnana.2017.00054

5. Arkley, K.P., Grant, R.A., Mitchinson, B. & Prescott, T.J. (2014). Strategy Change in Vibrissal Active Sensing during Rat Locomotion. *Current Biology* **24**, 1507-1512.

6. Hafner, G., Guy, J., Witte, M., Truschow, P., Rüppel, A., Sirmpilatze, N., Dadarwal, R., Boretius, S., Staiger, J.F., 2020. Increased Callosal Connectivity in Reeler Mice Revealed by Brain-Wide Input Mapping of VIP Neurons in Barrel Cortex. Cereb Cortex 31, 1427–1443. https://doi.org/10.1093/cercor/bhaa280

Age- and Sex-Related Behavioral Changes During the Lifespan of Wistar Rats

V. Borbélyová[1], A. Feješ[1], P. Sušienková[1], K. Lichá[1], J. Szabó[1], P. Celec[1] and K. Šebeková[1]

1 Institute of Molecular Biomedicine, Comenius University, Bratislava, Slovakia. borbelyova.veronika88@gmail.com

Introduction

Aging is characterized by gradual and progressive changes in brain function, accompanied by changes in behavior over the lifetime [1]. In humans, aging is associated with lower physical activity, higher anxiety, and cognitive decline [2]. In Wistar rats, age-related behavioral changes have been demonstrated by comparisons between young and adult rats of both sexes (2- and 5-month-old) [3], or between young and old rats (4- and 24-month-old) considering only male sex [4]. However, experiments dealing with age- and sex-dependent changes in the behavior of Wistar rats throughout their lifespan are lacking. Therefore, the aim of the present study was to evaluate age and sex-related changes in the behavior of female and male Wistar rats from young to old age.

Materials and Methods

In this study, we used female and male Wistar rats (n=90, Velaz, Prague, Czech Republic). Animals were divided into groups according to sex (females/males) and age: young rats (n = 10/10; age: 1 and 2 months), adult rats (n = 13/13; age: 5 and 10 months), middle-aged rats (n = 9/9; age: 15 months), and old rats (n = 17/9; age: 18 and 25 months). The housing room was temperature controlled (temperature 23±2°C and humidity 55±10%) with a 12:12 light-dark cycle (lights on at 7:00 a.m.). Rats had free access to food and water *ad libitum*. Rats were group-housed (3-5 per cage) in polycarbonate cages (50×36×19cm).

We tested the behavior of both female and male Wistar rats throughout life using a battery of behavioral tests in the following order: open field test (to assess locomotor activity and anxiety-like behavior), novel object recognition test (to assess exploratory behavior and memory) and elevated plus maze test (to assess anxiety-like behavior). We conducted behavioral testing in rats between 9. a.m. and 12 a.m. We tested the animals in only one behavioral test per day. At least 30 minutes before the beginning of each behavioral test, we transferred the animals into the testing room for acclimation.

We performed recordings of all behavioral tests using a camera placed above the apparatus used for the given behavioral test. We analyzed all observed behavioral parameters automatically using a computerized animal observation system and software (EthoVision XT 10.0, Noldus Information Technology, Wageningen, Netherlands). The apparatus used for behavioral testing, as well as the objects, were cleaned after each animal with Incidur spray (Ecolab, Dusseldorf, Germany) to prevent any olfactory disturbance.

Open field test

The open field apparatus consisted of a dark plastic square arena measuring 100cm x 100cm and was virtually divided into a center zone (40cm x 40cm) and a border zone. The arena of the open field apparatus was illuminated with white light (25 lx). We placed rats individually in the center zone of the open field arena and allowed rats to explore the apparatus freely for 5 minutes. We evaluated the following parameters: total distance moved (locomotor activity), number of entries, and time spent in the center zone of the open field apparatus (anti-anxiety behavior).

Novel object recognition test

We performed the novel object recognition test in the arena of open field testing (100cm x 100cm). We tested the animals in the open field test one day before the novel object recognition test. Therefore, rats were tested in an already familiar environment under the same experimental conditions as in the open field test. The novel object recognition test consisted of two trials: trial 1 (the training phase lasting 5 min) and trial 2 (the testing phase lasting 5 min), separated by a retention interval of 1 hour.

During the **training phase**, animals were exposed to two unknown objects (a glass bottle and a plastic bottle) filled with water. In this phase, we placed the rats in the middle of the arena with the two objects situated at the opposite corners of the arena (27cm from the wall and 55cm apart from each other). We evaluated the time spent exploring the glass bottle and plastic bottle for explorative behavior. Following the training phase, we placed rats into the home cage for one hour (retention interval).

Following a 1-hour retention interval, we conducted the **testing phase**, and we replaced one of the objects (the glass bottle) with an unknown object (a metallic can). Consequently, we removed the rat from the home cage and placed it in the middle of the open field arena. We allowed rats to explore the two objects (one familiar, a plastic bottle, and one unfamiliar, a metallic can) again for 5 min. Following the completion of the testing phase, we moved the animals to home cages. To eliminate side preference artifact of the objects, we randomly changed the position of the objects (glass bottle or metallic can) from one side of the arena to the opposite one. We evaluated the time spent exploring the metallic can and plastic bottle for explorative behavior. If the rat remembers the familiar object (the plastic bottle) during the testing phase, it spends more time exploring the new object (the metallic can). In the testing phase of the test, we assessed the short-term memory (recognition index in % for the novel object) of rats using a formula: [(time spent with exploration of novel object / time spent with exploration of familiar object + novel object)] *100.

In both trials, we defined exploration of the objects as time the rat spent in physical contact with the objects with the top of the nose (when the top of the nose was within 0.5 cm around the object) and engaged in active exploration – sniffing.

Elevated plus maze test

We performed the elevated plus maze test in a plus-shape arena elevated to a height of 50cm above the floor. The arena consisted of two open (45cm x 10cm) and two closed arms (45cm × 10cm) extending from a central platform (10cm x 10cm). The closed arms were enclosed by 40-cm-high walls. The illumination of open arms was 100-110 lx and that of closed arms was 3-5 lx. We placed rats in one of the open arms facing the environment outside the arena and allowed the rats to freely explore it for 5 min. The number of entries and time spent in the open arms were automatically analyzed by the EthoVision video tracking software. We evaluated lower anxiety in rats when the rats spent more time in the open arms and made more entries into the open arms.

Statistical analysis

For statistical analysis, we used GraphPad Prism version 9.5.1 for Windows (GraphPad Software, La Jolla, CA, USA). We analyzed the data using a two-way ANOVA (independent factors: sex and age) with a Bonferroni-corrected post-hoc t-test. We considered data statistically significant when P values were lower than 0.05.

Results

Young rats showed higher locomotor activity than the other three groups (1.5-2-fold; p<0.001) in the open field test. Adult rats moved 1.2-fold more compared to old rats (p<0.05). Female rats showed higher locomotor activity than male rats in all age-related groups (by 37%; p<0.01). The old rats entered the open arms of the elevated plus maze test 1.6-2-fold less compared to young and adult rats (p<0.05). Old male rats spent 2.2-fold more time in the open arms than young and adult male rats (by 43-80%; p<0.01). Old rats showed 1.9 times lower exploratory activity compared to young rats (p<0.05), and 2.5-fold lower in comparison with adult and middle-aged rats (p<0.05). Recognition memory did not differ significantly between the groups.

Conclusion

This study describes sex- and age-dependent behavioral differences throughout the lifespan of Wistar rats. Changes in locomotor activity occurred between young and older rats, and we observed further changes in anxiety-like behavior in aged rats. In parallel with previous research, we found lower locomotor and exploratory activity in old rats compared to young and adult rats. In addition, female rats showed higher locomotor activity than male rats at all ages. In contrast with the previous findings, old rats displayed lower anxiety-like behavior in comparison to young, adult, and middle-aged rats. We did not observe differences in short-term memory between age groups of rats. These findings provide basic information about locomotor activity, anxiety-like behavior, and short-term memory of young, adult, middle-aged, and old Wistar rats of both sexes. These findings should be considered when choosing the age of Wistar rats for conducting behavioral experiments.

Ethical Statement

All experimental procedures were approved by the Ethical Committee of the Institute of Molecular Biomedicine, Comenius University, Bratislava, Slovakia, and have been conducted in accordance with EU Directive 2010/63/EU and Slovak legislation.

Acknowledgement

This study was supported by grants from the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA 1/0341/23, and Slovak Research and Development Agency No. APVV-21-0355.

References

1. Shoji, H., Miyakawa, T. (2019). Age-related behavioral changes from young to old age in male mice of a C57BL/6J strain maintained under a genetic stability program. *Neuropsychopharmacology Reports* **39**, 100-118.
2. Gur, R.E., Gur, R.C. (2002). Gender differences in aging: cognition, emotions, and neuroimaging studies. *Dialogues in Clinical Neuroscience* **4**, 197-210.
3. Sudakov, S.K., Alekseeva, E.V., Nazarova, G.A., Bashkatova, V.G. (2021). Age-related individual behavioural characteristics of adult Wistar rats. *Animals* **11**, 2282.
4. Boguszewski, P., Zagrodzka, J. (2002). Emotional changes related to age in rats--a behavioral analysis. *Behavioral Brain Research* **133**, 323-332.

Neural Control of Odour Seeking Behaviour In the Fruit Fly

A. Miriyala[1], S. Waddell[1]

1 Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom.
ashwin.miriyala@dpag.ox.ac.uk

Abstract

Modern psychology tells us that our behaviours arise from inherited and learned traits. Seminal studies in the fruit fly *Drosophila melanogaster* provided a conceptual and experimental framework towards understanding how genetics and experience underlie behaviours. Recent development of sophisticated genetic approaches combined with the ability of flies to learn associations between odours and rewards or punishment, permit a detailed study of how genes provide the nervous system with innate behaviours and the capacity to learn.

Introduction

History of studying genetic contributions to behaviour

In the decades following Mendel's discovery of inheritance of traits, groundbreaking work from Thomas Morgan and Alfred Stutervant in *Drosophila* confirmed the existence of genes, and showed they could be mapped on chromosomes. By the 1950s, the physicist/geneticist Seymore Benzer showed using bacteriophages that genes were divisible, and mutations in their base pairs could alter traits of the phages. Benzer would later come to use *Drosophila* as a model, utilizing its short reproductive span and its extensively mapped genome to study genes and how they contribute to various behaviours.

Benzer induced mutations in *Drosophila* by feeding them the chemical mutagen ethyl methane sulfonate. He then took these flies and put them in a device he called a countercurrent apparatus, which tests their geotactic (tendency to walk upwards against gravity) and phototactic (tendency to walk towards light) behaviour [1]. Over the next decade, Benzer and his students used variations of this approach to discover a variety of mutants that not only would alter specific behaviours of the flies, but importantly these aberrant behaviours would persist across generations. In this way, Benzer pioneered the use of genetic manipulations in *Drosophila* as a model for studying the neurobiology of behaviour.

Measuring behaviour in groups of flies using the T-maze

William G. (aka Chip) Quinn modified the countercurrent apparatus and demonstrated for the first time that flies can be conditioned to associate an odour with an aversive electric shock [2]. Using this learning apparatus Quinn, Yadin Dudai, and colleagues discovered mutants, such as *dunce*, *rutabaga* and *amnesiac* that are defective in learning and memory.

An improved T-maze apparatus was soon designed [3] that permits better control over the pairing of odours and shock or reward, and this device is now widely used throughout the world to test fly learning and memory. To achieve aversive conditioning in the T-maze, flies are exposed to one odour (conditioning stimulus with shock i.e CS+) for 1 minute with twelve 90 V electric shocks delivered at 5 second intervals. After 45 seconds of clean air, the flies are then exposed to a second odour without shock (CS-) for 1 minute. Flies are then tested for their odour preference memory by giving them 2 minutes to choose between the CS+ and CS- odours. A performance index (PI) is calculated as the number of flies that avoid the previously shocked CS+ minus the number of flies going towards the CS-, divided by the total number of flies.

Today, *Drosophila* genetics is so sophisticated that we can target a wide variety of genetic tools to specific subsets of neurons within the fly brain, to test their contribution to learning and behaviour. The Gal4-UAS system is one example of a tool that allows us to manipulate gene expression in specific neurons in the flies' brain [4]. This gives us the means to express a variety of effectors of neural activity (allowing neuronal activation or silencing), for example thermogenetic and optogenetic tools that can be respectively controlled by altering temperature or shining bright lights. Such manipulations permit investigators to control and record from specific neurons while the fly is performing all types of behaviours. In addition, we now have a synapse-level connectome of an entire adult fly brain, including learning-relevant regions such as the mushroom bodies [5].

Measuring behaviour in individual, freely moving flies

A limitation of the T-maze is that it tests the outcome of manipulating a gene or set of neurons by assaying the behaviour of groups of flies. We can however now also monitor learning behaviour of individual flies. For example, several assays have been developed to show that flies can use spatial cues to drive goal-directed behaviour. One of these uses a grid-like maze to show that freely moving flies can keep track of the location of a reward while they continuously explore the maze [6]. In a confined arena, they have even been shown to learn the most efficient path to return to a reward [7]. Olfactory preference and learning can also be assayed in individual flies using a single fly t-maze [8]. This apparatus tracks a fly’s movement in two air streams that converge from opposite ends of a narrow 50 mm chamber.

Measuring behaviour using tethered fly-on-ball assay

Gotz and Buchner [9] developed a method whereby a fly is tethered and placed on a ball that is suspended on a stream of air. The advantages of this assay are that it allows a precise control over the fly’s sensory environment, and allows for monitoring behaviours such as grooming, goal-tracking and sleeping by using machine learning tools (e.g. DeepLabCut [10]) to track the fly’s bodyparts.

This assay was later improved so that the fly can be tethered by fixing its head to a platform while it walks on the ball [11]. This allows for using imaging techniques to monitor activity of neurons in the fly’s brain. One of the most commonly used imaging techniques in *Drosophila* is calcium imaging, whereby specific neurons can be made to express a fluorescent calcium indicator such as GCaMP [12]. In this way, whenever the neurons are active, they release photons that can be detected for example by two-photon microscopy, which is capable of visualizing calcium transients at sub-second resolution. The most recent calcium sensor for the fly has a half-rise time of around 80 ms and a half-decay time of around 200 ms [12]. This sensor is useful to understand how neural activity is phase-locked to motor activity and stimulus presentation, and how activity is correlated across populations of neurons. Recently, voltage imaging has been demonstrated in the fly which provides a much better temporal resolution (up to 1600 Hz) which allows for monitoring action potentials from populations of neurons [9]. This method uses a high-speed camera to capture photons released by neurons that express a voltage indicator such as pAce [13].

There are therefore a variety of assays available to measure behaviour in flies, ranging from observing population level learning down to individual neural activity. In the following sections, I will discuss my own project which involved setting up a fly-ball paradigm that allows for studying how flies use olfactory cues to coordinate goal-directed behaviour. In my earliest studies I used water vapour as a predictor of reward in naive (untrained) thirsty flies, and use a combination of T-maze and flyball experiments to study the neurons involved in seeking a water resource.

Methods

Tethered fly-on-ball

The fly-ball setup was adapted from Seelig *et al.* [11] to allow delivery of olfactory cues to the fly (Figure 1). A camera (Point-Grey, Grasshopper USB3) sends images of the ball to a software (Fictrac; [14]) that tracks dots drawn on the ball to reconstruct the walking trajectory of the fly (Figure 1a). Rotation around the flies x-axis reflects left-right movement, around its y-axis reflects forward-backward movement, and around its z-azis reflects clockwise-counterclockwise rotation (Figure 1c). Another camera (Point-Grey) records the body movements of the fly, and is used with DeepLabCut [7] to track individual points on the fly’s body (Figure 1b).

For odour delivery, a custom python script (executed via a gui built on the tkinter python library) sends commands to mass flow controllers (Sensirion) and air valves (Lee Company) which direct air through vials of water vapour or mineral oil, which is then directed towards the fly (Figure 1d). The rate of airflow to the fly always remains consistent with minimum air fluctuations during the opening and closing of the valves. The odour port that delivers odours to the fly is connected to a servo that is controlled by an Arduino (Figure 1e). The servo is in closed loop with the heading of the fly, so that any turns that the fly makes is mirrored by the angle of the odour port relative to the fly. The servo (and hence the odour port) can only rotate 80 degrees clockwise and anticlockwise, with the 0 degree position being head-on relative to the fly.

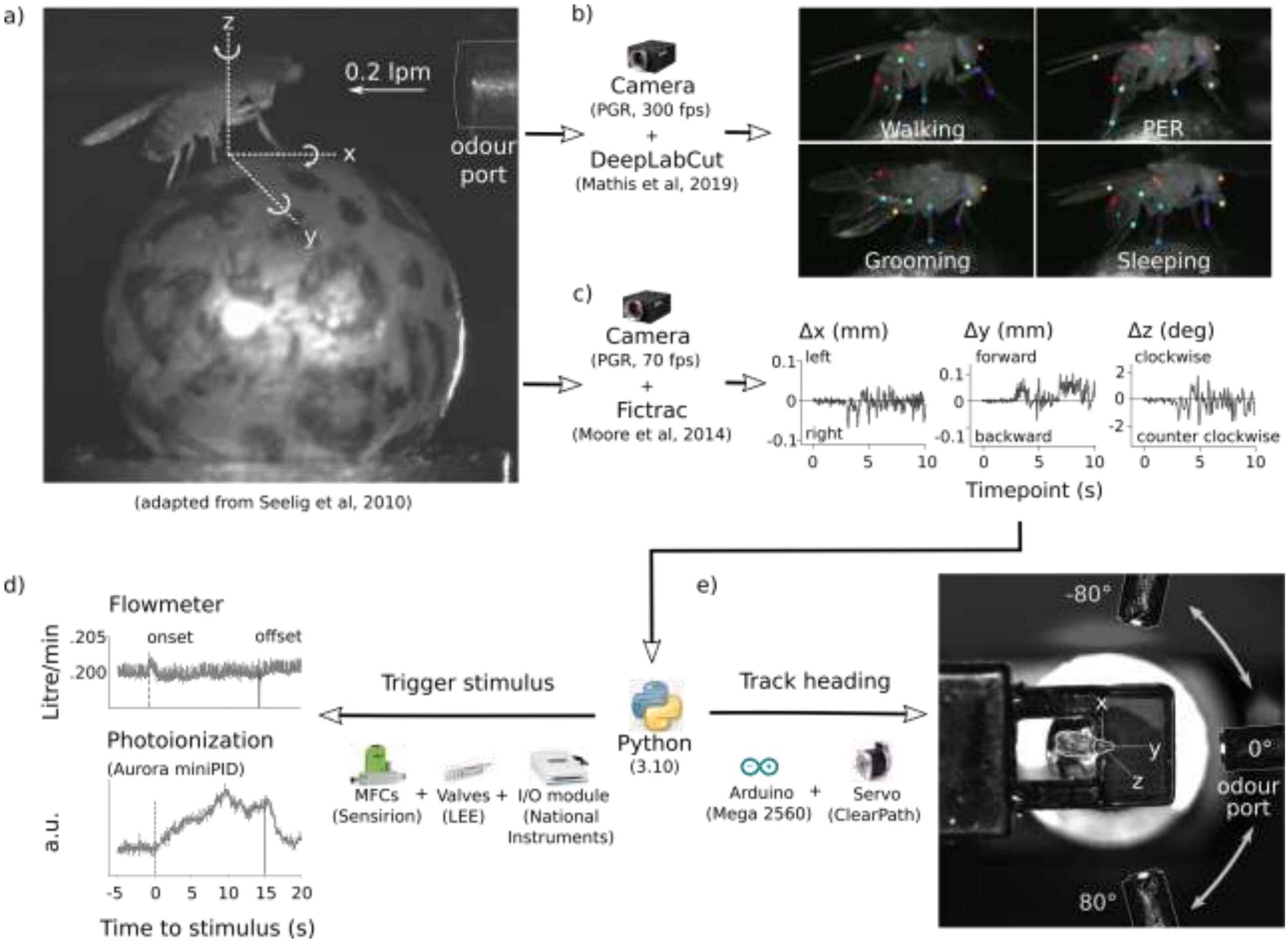


Figure 1
Tethered fly-on-ball setup with closed-loop odour delivery to measure olfactory responses.

Tmaze

The t-maze was used to measure naive (untrained) approach to water vapour (Figure 2). A group of 100 flies are loaded into the elevator of the tmaze and lowered down to a choice point, where they are given 2 minutes to move either towards air that has been bubbled through mineral oil or air that has been bubbled through tap water. A half-preference index is calculated as the number of flies that go towards the water vapour arm minus the number of flies that go towards the mineral oil arm, divided by the total number of flies. To remove side bias, water vapour delivery is switched sides in successive tests, and a full PI is obtained by averaging the PI obtained from left and right delivery (n=1).

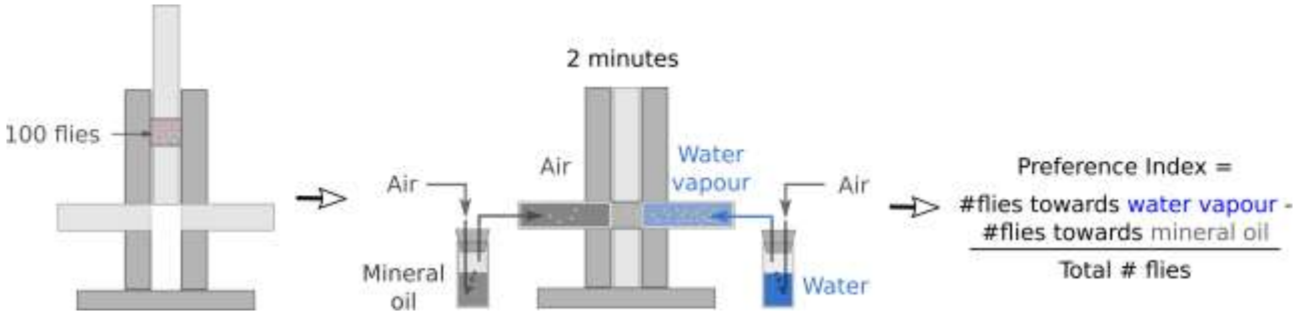


Figure 2
T-maze setup to test naive water vapour approach.

Results

A previous study from our lab [15] implicated specific dopamine neurons that project to the mushroom body in driving naive approach to a water reward in thirsty flies. I first tested whether these dopamine neurons are implicated in using water vapour as a cue in directing approach to water as well. In the T-maze assay, thirsty flies can use water vapour as a cue to instruct approach (Figure 2a). I found that the same dopamine neurons that instruct naive approach to a water reward [15] also instruct naive approach to water vapour (Figure 2b).

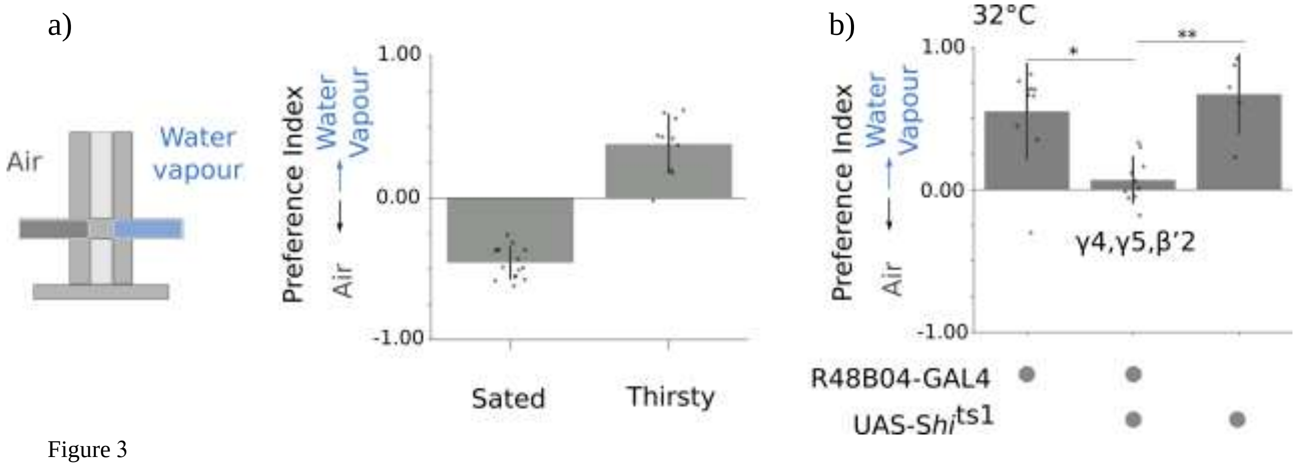


Figure 3
Dopamine neurons direct approach to water vapour
a) [*left*] Schematic of t-maze. [*right*] Performance index of thirsty flies, showing that sated flies avoid water vapour (hence the negative PI values) but thirsty flies approach water vapour. b) By using UAS-*shits1* driven by R48B04-Gal4, we blocked the DANs that project to the γ4,γ5,β'2 compartments of the mushroom body and found a significantly reduced approach to water vapour.

My next goal was to image the dopamine neurons that are implicated in water seeking to understand their role in driving approach behaviour. To do this, I needed to first establish what approach looks like on the flyball setup. I tether thirsty flies and deliver a 15 second duration of water vapour. These flies show a robust response that is reflected by a significant increase in speed and in the absolute rotation (which reflects heading changes i.e rotation around the fly's z-axis; Figure 3).

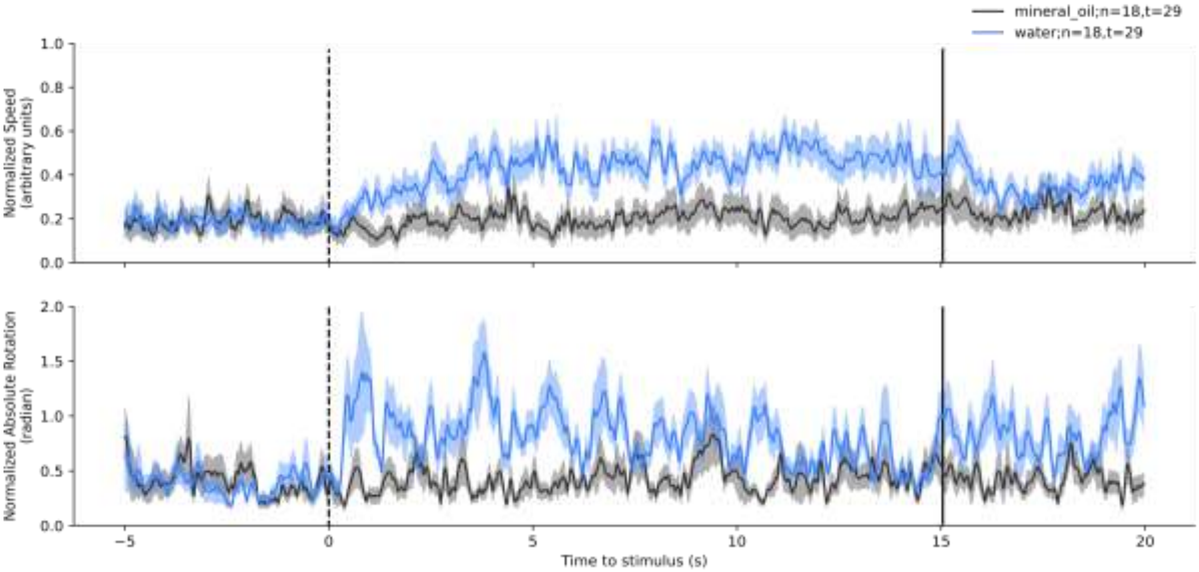


Figure 3
Tethered flies show a robust response to water vapour
top – normalized speed (movement around the fly's x-y axis). In blue is the water vapour trace and in black is the mineral oil trace. The dotted line indicates the onset of water vapour (or mineral oil; air is passed through water or mineral oil and then directed towards the fly, see methods) and the solid line indicates the offset. Outside these regions there is no water vapour or mineral oil delivery, just a clean air flow towards the fly. Solid blue (or black) lines are the mean (n=18 flies, 1 or 2 trials per fly) and the shaded area are the standard error of mean.

In addition to recording walking trajectories, I have trained a machine learning algorithm using DeepLabCut which efficiently tracks the positions of the flies legs, wings, haltere and proboscis (figure 4). This provides a rich dataset that allows us to analyse behaviours with an incredibly fine resolution.

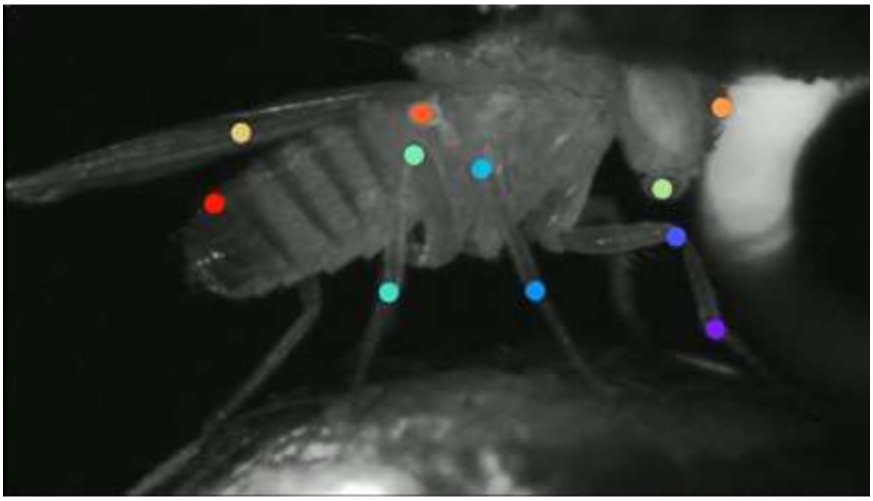


Figure 4
Tracking body parts of the fly
DeepLabCut was used to train a machine learning algorithm that allows to track body parts of a tethered fly on a ball.

Discussion

Using the T-maze, I show that thirsty flies can use water vapour as a cue to search for water reward. On the fly-ball setup, thirsty flies respond to water vapour by increasing their speed and by changing their heading. For future experiments, I will genetically target the dopaminergic neurons that are implicated in driving approach to water and test whether they have a role in controlling speed or heading. Dopaminergic neurons provide teaching signals, and their activity has recently been associated with instructing ongoing fly behaviour [16]. In addition, I will image from these dopaminergic neurons using two-photon microscopy to observe if there are correlations in dopaminergic neuron activity with different features of approach. Since I can track the fly's body parts while they are responding to water vapour, I can also see if there are features in this dataset that reflect goal-directed behaviour.

References

1. Benzer, S.: Behavioral Mutants of Drosophila Isolated By Countercurrent Distribution, Proceedings of the National Academy of Sciences, **58** (1967), no. 3, pp. 1112–1119.

2. Quinn, W.G.; Harris, W.A. and Benzer, S.: Conditioned behavior in Drosophila melanogaster, Proceedings of the National Academy of Sciences of the United States of America, **71** (1974), no. 3, pp. 708–712.

3. Tully, T. and Quinn, W.G.: Classical conditioning and retention in normal and mutant Drosophila melanogaster, Journal of Comparative Physiology A, **167** (1985), no. 2, pp. 263–277.

4. Owald, D.; Lin, S. and Waddell, S.: Light, heat, action: neural control of fruit fly behaviour, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, **370** (2016), no. 1677, p. 20150211.

5. Li, F.; Lindsey, J.W.; Marin, E.C.; Otto, N.; Dreher, M.; Dempsey, G.; Stark, I.; Bates, A.S.; Pleijzier, M.W.; Schlegel, P.; Nern, A.; Takemura, S.-Y.; Eckstein, N.; Yang, T.; Francis, A.; Braun, A.; Parekh, R.; Costa, M.; Scheffer, L.K.; Aso, Y.; Jefferis, G.S.; Abbott, L.F.; Litwin-Kumar, A.; Waddell, S. and Rubin, G.M.: The connectome of the adult Drosophila mushroom body provides insights into function, eLife, **9** (2020), p. e62576.

6. Corfas, R.A.; Sharma, T. and Dickinson, M.H.: Diverse Food-Sensing Neurons Trigger Idiothetic Local Search in Drosophila, Current Biology, **29** (2019), no. 10, pp. 1660-1668.e4.

7. Navawongse, R.; Choudhury, D.; Raczkowska, M.; Stewart, J.C.; Lim, T.; Rahman, M.; Toh, A.G.G.; Wang, Z. and Claridge-Chang, A.: Drosophila learn efficient paths to a food source, Neurobiology of Learning and Memory, **141** (2016), pp. 176–181.

8. Claridge-Chang, A.; Roorda, R.D.; Vrontou, E.; Sjulson, L.; Li, H.; Hirsh, J. and Miesenböck, G.: Writing Memories with Light-Addressable Reinforcement Circuitry, Cell, **149** (2009), no. 2, pp. 405–416.

9. Buchner, E.: Elementary movement detectors in an insect visual system, Biological Cybernetics, **24** (1976), no. 2, pp. 85–101.

10. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W. and Bethge, M.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning, Nature Neuroscience, **21** (2018), no. 9, pp. 1281–1289.

11. Seelig, J.D.; Chiappe, M.E.; Lott, G.K.; Dutta, A.; Osborne, J.E.; Reiser, M.B. and Jayaraman, V.: Two-photon calcium imaging from head-fixed Drosophila during optomotor walking behavior, Nature Methods, **7** (2010), no. 7, pp. 535–540.

12. Zhang, Y.; Rózsa, M.; Liang, Y.; Bushey, D.; Wei, Z.; Zheng, J.; Reep, D.; Broussard, G.J.; Tsang, A.; Tsegaye, G.; Narayan, S.; Obara, C.J.; Lim, J.-X.; Patel, R.; Zhang, R.; Ahrens, M.B.; Turner, G.C.; Wang, S.S.-H.; Korff, W.L.; Schreiter, E.R.; Svoboda, K.; Hasseman, J.P.; Kolb, I. and Looger, L.L.: Fast and sensitive GCaMP calcium indicators for imaging neural populations, Nature, **616** (2023), no. 7954, pp. 884–891.

13. Kannan, M., Vasan, G., Haziza, S., Huang, C., Chrapkiewicz, R., Luo, J., Cardin, J.A., Schnitzer, M.J., Pieribone, V.A., 2022. Dual-polarity voltage imaging of the concurrent dynamics of multiple neuron types. Science 378

14. Moore, R.J.D.; Taylor, G.J.; Paulk, A.C.; Pearson, T.; van Swinderen, B. and Srinivasan, M.V.: FicTrac: A visual method for tracking spherical motion and generating fictive animal paths, Journal of Neuroscience Methods, **225** (2015), pp. 106–119.

15. Lin, S.; Owald, D.; Chandra, V.; Talbot, C.; Huetteroth, W. and Waddell, S.: Neural correlates of water reward in thirsty Drosophila, Nature Neuroscience, **17** (2015), no. 11, pp. 1636–1642.

16. Handler, A.; Graham, T.G.W.; Cohn, R.; Morantte, I.; Siliciano, A.F.; Zeng, J.; Li, Y. and Ruta, V.: Distinct Dopamine Receptor Pathways Underlie the Temporal Sensitivity of Associative Learning, Cell, **178** (2019), no. 1, pp. 60-75.e19.

Human Factors

Beyond Snapshots: Validation of a Continuous Frustration Assessment in a Simulator and Real-World Setting

E. Bosch[1], S. Bohmann[1], U. Slivsek[1,2], K. Ihme[1]

[1] German Aerospace Center, Institute for Transportation Systems, Braunschweig, Germany, firstname.lastname@dlr.de

[2]University of Ljubljana, MEi:CogSci, Ljubljana, Slovenia

Abstract

Scale-based questionnaires are frequently used to assess complex psychological states such as emotions; however, these scales are often utilized for single-instance reporting and as such do not capture the complete dynamics of emotion occurrence and changes. This study aimed to compare a continuous after-study measurement of subjectively experienced frustration, to frustration ratings given on a 5-point Likert scale reported after each condition. Data was collected in a high-fidelity driving simulator and in a real-world study with an automated driving vehicle. We found that the during-study Likert-Scale ratings correlate highly with the mean after-study continuous frustration ratings in both the simulator and real-world setting. The results indicate that the after-study continuous rating is a viable alternative to the during-study Likert-scale when measuring frustration.

Introduction

Traditional emotion questionnaires employ single-instance questionnaires due to the need of multi-item questionnaires to ensure inter-item reliability [22]. However, when investigating time-resolved indicators of emotion, such as physiological data, a time-resolved subjective rating is necessary in order to research how dynamics of emotion occur and how changes in subjective experience and physiological changes are interrelated. Continuous ratings given during an experiment can disrupt the natural progression of emotions and reveal the objective of emotion induction. Therefore, one viable alternative for a continuous scubjective emotion rating is a post-hoc (post study) and single-item assessment of an emotion. One attempt of continuous emotion measurement is the affect rating dial first used by Levenson and Gottman [19]. To receive a continuous emotion rating while couples were interacting, the couples were video-taped during their interaction. Subsequently, they returned to the lab separately and provided a continuous positive-negative emotion rating post-hoc. Other studies have also used continuous post-hoc measurements by recording participants and collecting their rating afterwards [9, 16, 18, 20, 24, 27]. In this paper, we compare such a post-hoc continuous measurement to a during-study 5-point Likert scale frustration rating. Allowing participants to self-rate their emotions after the study circumvents some of the challenges associated with continuous emotion annotation as described in [21].

This study is set in the context of measuring subjectively experienced frustration in fully automated driving. Frustration is especially interesting in the context of automated driving, as the experience of frustration can inhibit the acceptance of new mobility concepts such as automated driving [10, 25]. It is, therefore, highly interesting to understand how frustration can be recognized in this context [5]. For this, traditional subjective ratings ask for participant's emotion ratings once after every condition. However, to acquire highly time-resolved information about emotional responses to different events within an experimental condition and to connect it to possible changes in acquired sensor data, it can be helpful to obtain a time-resolved subjective rating. To see whether relationships between single-instance and post-hoc continuous frustration ratings differ depending on the context, we collected data in a high-fidelity driving simulator and a real automated driving car on a test track. Based on previous research [4], we used an in-vehicle interface to induce frustration. We then explored how well both ratings correlated. For this correlation, different metrics of the continuous frustration rating can be interesting. For example, [26] found that due to a duration neglect, the maximum and end pain ratings during a colonoscopy can be better predictors for an overall experience rating given after the procedure than the mean rating. We therefore compared the during-study Likert-scale frustration not only to mean, but also combined maximum and last-minute values of the after-study continuous rating.

Methods

Summary

Study 1 was conducted at our institute's high-fidelity driving simulator with 50 participants. Frustration was induced by interaction with an in-vehicle user interface. Subjective frustration ratings were collected after each drive on a 5-point-scale and after all drives as a continuous (i.e., highly time-resolved) rating. To test whether the results of the simulator study could be replicated in a real-world setting, we designed Study 2 as close as possible to Study 1 in a real car on a test track with 23 participants. Every participant experienced baseline and frustrating drives which were all driven on the same test track. Subjective frustration ratings were, again, collected after each drive on a 5-point-scale and as a continuous rating after all drives as manipulation check. The participants were brought to a test track before the start of the study, which took about 20 min. The participants were different from the ones in Study 1. Results of this study's camera and EEG data results are published under [6].

Participants

Fifty participants recruited through the institute's participant pool took part in Study 1. Previous studies with similar scope and settings had comparable sample sizes [13, 14, 31]. In total, nine participants were excluded from data analyses, due to motion sickness (2), data saving problems (3), and missing data (4). The n = 41 participants included in the analyses were aged 20 to 59 years (y) (M = 31.54 y, SD = 12.46 y, 12 female, 29 male). Twenty-two participants recruited through the institute's participant pool took part in Study 2. The decision to recruit twenty-two participants was based on the tradeoff of measuring as many participants as possible within a feasible time of availability of the research car and the test track. One participant had to end the experiment early (for urgent private reasons). The n = 21 participants included in the analyses were aged 23 to 58 years (y) (M = 41.71 y, SD = 10.34 y, 5 female, 16 male). As reimbursement for their time, all participants received 5 € per commenced half hour for their participation.

Set-Up

Study 1 was conducted in a driving simulator virtual reality lab with 360° full view. The participants sat in a realistic vehicle mock-up. Study 2 was conducted in our institute's test vehicle on a test track (comparable to SAE Level 4, 28). The participant sat in the driver seat and did not engage in any driving task. A security driver was present at all times on the co-driver seat with access to break and throttle. The car drove with a maximum speed of 30 km/h on a track of roughly 1.6km. In both studies, the UI was displayed on a tablet (Microsoft Surface Pro 7, 12.3') that was attached over the center console of the car.

Stimuli

The participants read a story to immerse into the setting before all drives that told them they were driving to a business meeting. Participants then solved a task on the in-car user interface displayed on the tablet. The participants were told to receive a 2 € reward upon successful completion of their task. In the baseline condition ('*Baseline*'), the participants were asked to visit a website, which could be accomplished easily. They were then asked to press the one button that appeared in different places of the UI. They were told to have no time pressure and to interact with the UI as natural as possible. In the first frustration condition (*'Frust1'*):, the participants received a call from their 'boss', who told them that they were urgently needed for another, more important meeting and needed to turn around immediately to arrive on time. The participants then had to change the destination of the navigation system. Through ambiguous naming of buttons, unclear icons, and unintuitive paths, this was hard to achieve within 7 min. In the second automation condition (*'Frust2')*, a 'boss' called and asked the participant to very urgently join an online conference with clients. Again, the UI was so difficult to understand that it was hard to reach the goal of joining the online conference within the given time.

Measures

To assess the participant's frustration levels, the participants rated their frustration in two different frustration scales. One was an emotion questionnaire that was filled in after every drive ('during-study Likert-scale frustration rating'). It first asked four distraction questions about gaze behavior in line with the cover story (see Supplementary Materials for the exact questions). Afterwards, the participants rated an emotion scale based on the German version of the positive and negative affect scale 'PANAS' [17]. It has a reliability of Raykovs ρ = 0.93 [7] and is a commonly used method to acquire participant's emotions (see, for example, 2, 12, 30). The translated emotions words used were ‘active’, ‘distressed’, ‘interested’, ‘excited’, ‘upset’, ‘scared’, ‘inspired’, ‘proud’, ’enthusiastic’, ‘ashamed’, ‘alert’, ’nervous’, ’determined’, ’attentive’, ’jittery’, ’afraid’ (from the original PANAS) and ‘insecure’, ‘frustrated’, ‘angry’, ‘sad’, ‘surprised’, ’relaxed’ (our own addition) were rated on a 5-point scale from ‘not at all’ to ‘extremely’. We decided to acquire this broad emotion spectrum to hide that we were trying to induce frustration and also added emotion words similar to frustration to test for a latent frustration construct by factor analysis.

The second frustration rating (‘after-study continuous frustration rating’) was obtained after all drives. For this, the participants watched the videos that were recorded during all drives of the whole scene (the participant’s face was not visible) and rated their frustration with a joystick on a level from 0 to 100%. This rating was given continuously, i.e. the participant always held the joystick in the position that corresponded to their frustration level as experienced in the situation shown in the video. The joystick was moveable only in one direction and automatically returned to zero-position when not touched. The participants saw a visual feedback of their current rating, which was presented next to the video. They were asked to move the joystick according to the frustration level that they felt in the situation shown in the presented video. By this, a continuous frustration rating for each drive and each participant was collected. We decided for this continuous measure in addition to the common method of questionnaires to receive a subjective rating not only once per drive, but for every timepoint during the drive. In the Simulator study, the time between the last drive and giving the after-study continuous frustration rating was about 10 min, in the Real-world study it was about 45 min.

Procedure

All participants arrived and filled in an informed consent and a data privacy statement. Before the start of Study 1, participants were informed of potential risks of driving in simulators (e.g., the experience of simulator sickness) according to the simulator safety concept. In Study 2, Participants were brought to a test track, which took about 20min. Before the start of the study, the participants were informed about potential risks of driving in an automated vehicle on a test track with safety driver (e.g., the experience of motion sickness) according to the vehicle safety concept. The participants were informed that they could take a break or abort their participation at any time. All participants provided written informed consent to take part in the study and the video recording. The participants were told the cover story that the study investigated differences in gaze behavior between manual and automated driving modes. This was done to conceal the true aim of frustration induction and enable natural emergence of emotions. To reduce effects that came from unfamiliarity, all participants experienced automated driving scenarios before the start of the experiment until they said to be adapted to the simulator or the automated riving car, respectively. This took five minutes on average. After the all drives, the participants were informed about the true goal of the experiment (evoking frustration) and the necessity to conceal this goal with a cover story. They then gave the continuous frustration rating for all drives. The whole procedure took 2 hours on average. The collected data was handled and saved in line with the European General Data Protection Regulation. A project-internal ethics committee reviewed and approved the study.

Experimental Design

This data collection was part of a larger study as described in [5, 6]. Therefore, Participants also drove manual driving modes in Study 1. In a 2 (driving mode: automated vs. manual) x 2 (frustration induction: frustration vs. baseline) within-subject design, each participant experienced six drives in total. Three of these were driven by the participants themselves (manual driving mode) and in three the car drove automatically (automated driving mode). Both driving modes consisted of one baseline drive and two frustration-inducing experimental drives each. The order of the drives was balanced by a balanced Latin square design for all participants, which means that every condition was driven in every position, and also the order of the drives was balanced (see for example 15). The same was true for Study 2, where only automated driving conditions existed (see [5]).

Data Analysis

As our factor-analytical approaches did neither reveal a fitting measurement model for negative affect nor for a latent frustration construct, we correlated the ‘frustrated’ item ratings after each condition (‘during-study Likert-scale frustration rating’) with the after-study continuous rating’s mean of each respective condition. This was done by spearman rank correlation due to the ordinal nature of the during-study Likert-Scale. Considering a heuristic perspective to the experience of affective episodes [11, 26] we first fitted an ordinal logistic regression model (Model 1) with the predictor variable ‘mean after-study continuous frustration’. We then extended this model to include a linear combination of the peak and the mean of the last minute of the continuous frustration rating as predictors in Model 2. We report both models’ pseudo-r-squared values and used a likelihood ratio test to compare Model 1 and Model 2. In Model 3, we fitted an ordinal logistic regression model with the predictor variable ‘Peak-end value of after-study continuous frustration’ only and then compared it to Model 2 by likelihood ratio.

Results

Figure shows a comparison of during- and after-study frustration ratings. Over both the simulator and real world drives, the Spearman’s rank correlation coefficient of emotion scale rating per drive and mean continuous frustration rating per drive was 0.57, which is a high correlation according to Cohen [8]. The correlation was as high when only considering the simulator setting and 0.69 when only considering the real-world study (see Table 5).

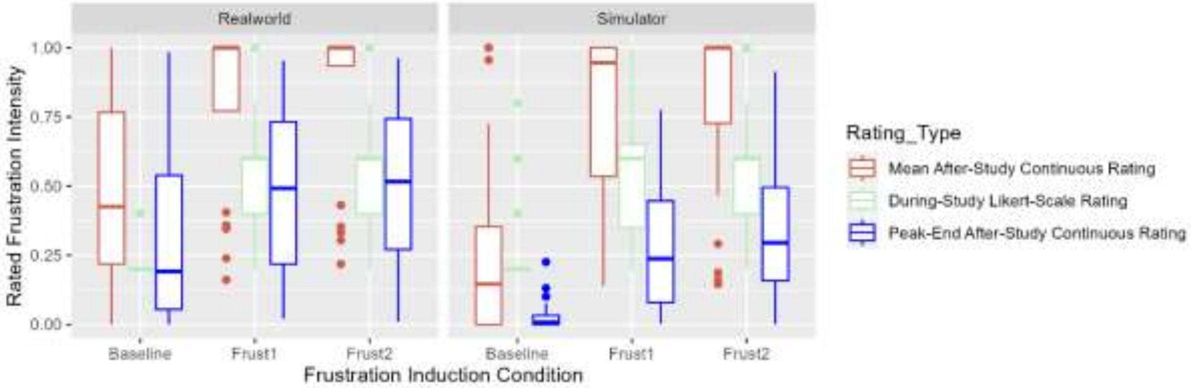


Figure 1: Comparison of during-study Likert-Scale rating and after-study continuous rating by using the mean and peak-end rating of the continuous rating. Likert Scale of Emotion Scale divided by 5 to have a comparable axis.

Table 5: Spearman's rank correlations of after-study continuous frustration rating with during-study Likert-Scale frustration ratings.

Setting	Mean of After-study continuous rating with During-study Likert-Scale rating	Peak-end value of After-study continuous rating with During-study Likert-Scale rating
Both settings	0.57	0.53
Simulator	0.57	0.54
Real-world	0.69	0.63

The ordinal logistic regression model with only the predictor variable ‘mean continuous after-study frustration’ (Model 1) yielded a 33.4% explanation of the variance in the data based on the Cragg and Uhler's pseudo R-squared [23]. Model 2 that added the peak and last-minute-mean values of the continuous after-study rating (‘Peak-end value’), yielded a pseudo R-squared value of 34.6%. A likelihood ratio test was performed to compare Model 1 and Model 2. The test statistic was *LF = 3.20*, and the p-value was *p = .07*. Since the p-value is higher than 0.05 and the increase in explained variance is minimal, we do not reject the more parsimonious Model 1. Model 3 with only the predictor variable ‘Peak-end value’ yielded a pseudo R-squared value of 21.0% and a likelihood ratio test to compare Model 2 and 3 resulted in a test statistic of *LF = 32.29* with a *p < .001*. We therefore accept the hypothesis that a model including ‘continuous after-study frustration’ is significantly better than a model that only contains ‘peak-end value’. Overall, we therefore prefer the most parsimonious model, using only the mean of ‘continuous after-study frustration’ as predictor, as it achieves no worse goodness of fit than model with additional predictors.

Discussion

In this study, we aimed to compare subjective frustration ratings given on a 5-point Likert Scale after every drive (‘during-study Likert-scale frustration rating’) to a continuous frustration rating given after all drives (‘after-study continuous frustration rating’). We did this comparison in the setup-up of a high-fidelity driving simulator and a real-world study with an automated driving car on a test track. As a result, we found that the ratings given after every drive correlate highly with the continuous frustration rating given after all drives in both set-ups. Previous research has compared a continuous emotion rating to a partner’s emotion rating [20], to emotions expected by induction methods [18] or not compared it to another rating [16, 24, 27]. These findings suggest that a post-hoc continuous rating can be used in studies where a higher time-resolution of a subjective rating is necessary. The higher mean after frustration rating compared to the during study frustration rating might occur because the emotion of frustration became more salient for the subjects as a result of the instruction.

One disadvantage of single-item measurements compared to multi-item measurements of a latent construct is that they are more prone to measurement error and therefore have a lower reliability in many cases [1]. Considering that low reliability reduces the correlation with other variables, the rather strong correlation we found between the continuous rating and the frustration item indicate that the post-hoc rating is a viable alternative to the after-drive frustration scale. Applying heuristics that have previously been found to yield a better fit than the mean of a continuously given rating [11, 26] did not result in a meaningfully improved model fit in comparison to only taking the mean post-study continuous frustration rating as predictor for the during-study Likert-scale rating. Using only the peak-end heuristic as proposed by [26] resulted in a significantly worse model fit than using mean and the peak-end values as predictors. This indicates that the peak-end value does not improve the model fit that can be achieved by only using the mean after-study continuous rating as predictors. [26]’s heuristics do not seem to apply to the relationship between after-study continuous rating and during-study Likert-scale rating in our study.

One limitation of our study is that the Likert scale frustration after each drive was measured by a single frustration item, so that measurement error cannot be considered. Probably due to small sample size and skewed indicators, longitudinal confirmatory factor analysis models resulted either in bad model fits or estimation problems like negative error variances. On the other hand, our struggles to fit a model dovetail with reports of structural ambiguity of the German version of PANAS e.g. by [29]. Future research could induce frustration in a larger study sample, for example in an online study, and do a similar comparison of after-drive and after whole study ratings. We encourage researchers to factor analyze indicators of emotional constructs, particularly in German language. A limitation of the comparison to [26]’s heuristics of memory bias is that in our study, the continuous rating was given later than the single-item scale rating. This is opposed to [26]’s study design and might explain why adding the peak and end ratings did not improve the model fit.

Conclusion

This study set out to compare whether a continuous frustration rating given after a study yields results comparable to a 5-point Likert scale rating given after every experimental condition. Our experiments confirmed that the correlation between the two ratings is high. This suggests that memory effects that might bias the rating after all drives can be neglected in future studies and, when in need of a continuous rating, this is a viable alternative to the during-study Likert-Rating in future studies. Further research using a multidimensional during-study frustration rating and more participants is needed.

Acknowledgements

This work was partly supported by the project F-RELACS which receives funding from the German Federal Ministry of Education and Science in the program KMU-innovativ under the grant 16SV7930.

Data availability

This manuscript’s data will be made publicly available after acceptance under [3].

References

1. Allen, M. S., Iliescu, D., and Greiff, S. 2022. Single Item Measures in Psychological Science. *European Journal of Psychological Assessment* 38, 1, 1–5.
2. Barańczuk, U. 2018. Emotion regulation mediates the effects of temperament traits and posttraumatic stress disorder symptoms on affect in motor vehicle accident survivors. *Transportation research part F: traffic psychology and behaviour* 58, 528–535.
3. Bosch, E. 2023. *Continuous vs. Single-Instance Frustration Rating Dataset. DOI*=10.17605/OSF.IO/MVRJK.
4. Bosch, E., Ihme, K., Drewitz, U., Jipp, M., and Oehl, M. 2020. Why drivers are frustrated: results from a diary study and focus groups. *European Transport Research Review* 12, 1, 52.
5. Bosch, E., Käthner, D., Jipp, M., Drewitz, U., and Ihme, K. 2023. Fifty shades of frustration: Intra- and interindividual variances in expressing frustration. *Transportation research part F: traffic psychology and behaviour* 94, 436–452.
6. Bosch, E., Klosterkamp, M., Guevara, A., Kaethner, D., Bendixen, A., and Ihme, K. 2022. Multimodal Estimation of Frustrative Driving Situations Using a Latent Variable Model. In *2022 13th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 11–16. DOI=10.1109/CogInfoCom55841.2022.10081636.
7. Breyer, B. and Bluemke, M. 2016. Deutsche version der positive and negative affect schedule PANAS (GESIS panel).
8. Cohen, J. 1988. *Statistical power analysis for the behavioral sciences, 2nd edn. Á/L*. Erbaum Press, Hillsdale, NJ, USA.
9. Cowie, R., McKeown, G., and Douglas-Cowie, E. 2012. Tracing Emotion. *International Journal of Synthetic Emotions* 3, 1, 1–17.
10. Ferreri, N. R. and Mayhorn, C. B. 2022. Identifying and understanding individual differences in frustration with technology. *Theoretical Issues in Ergonomics Science*, 1–19.
11. Fredrickson, B. L. and Kahneman, D. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology* 65, 1, 45–55.
12. Frison, A.-K., Wintersberger, P., and Riener, A. 2019. Resurrecting the ghost in the shell: A need-centered development approach for optimizing user experience in highly automated vehicles. *Transportation research part F: traffic psychology and behaviour* 65, 439–456.
13. Hoque, M. and Picard, R. W. 2011. Acted vs. natural frustration and delight: Many people smile in natural frustration.
14. Ihme, K., Unni, A., Zhang, M., Rieger, J. W., and Jipp, M. 2018. Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. *Frontiers in human neuroscience* 12, 327.
15. Kim, B. G. and Stein, H. H. 2009. A spreadsheet program for making a balanced Latin square design. *Revista Colombiana de Ciencias Pecuarias* 22, 4, 591–596.
16. Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. 2019. Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *Int J Comput Vis* 127, 6-7, 907–929.
17. Krohne, H. W., Egloff, B., Kohlmann, C.-W., and Tausch, A. 1996. Untersuchungen mit einer deutschen Version der" Positive and negative Affect Schedule"(PANAS). *Diagnostica-Gottingen-* 42, 139–156.
18. Laurans, G., Desmet, P. M. A., and Hekkert, P. 2009. The emotion slider: A self-report device for the continuous measurement of emotion. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–6. DOI=10.1109/ACII.2009.5349539.
19. Levenson, R. W. and Gottman, J. M. 1983. Marital interaction: physiological linkage and affective exchange. *Journal of personality and social psychology* 45, 3, 587–597.

20. Levenson, R. W. and Ruef, A. M. 1992. Empathy: A physiological substrate. *Journal of personality and social psychology* 63, 2, 234–246.
21. Metallinou, A. and Narayanan, S. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8. DOI=10.1109/FG.2013.6553804.
22. Moosbrugger, H. and Kelava, A. 2020. *Test theory and questionnaire construction*. Berlin: Springer.
23. Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *biometrika* 78, 3, 691–692.
24. Ong, D. C., Wu, Z., Tan, Z.-X., Reddan, M., Kahhale, I., Mattek, A., and Zaki, J. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE transactions on affective computing* 12, 3, 579–594.
25. Partala, T. and Saari, T. 2015. Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Computers in Human Behavior* 53, 381–395.
26. Redelmeier, D. A. and Kahneman, D. 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66, 1, 3–8.
27. Ruef, A. M. and Levenson, R. W. 2007. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment*, 286–297.
28. SAE International. 2014. *Automated Driving Levels of Driving Automation are Defined in New SAE International Standard J3016*. SAE International Troy, MI.
29. Seib-Pfeifer, L.-E., Pugnaghi, G., Beauducel, A., and Leue, A. 2017. On the replication of factor structures of the Positive and Negative Affect Schedule (PANAS). *Personality and Individual Differences* 107, 201–207.
30. Zhang, M., Ihme, K., and Drewitz, U. 2019. Discriminating drivers' emotions through the dimension of power: evidence from facial infrared thermography and peripheral physiological measurements. *Transportation research part F: traffic psychology and behaviour* 63, 135–143.
31. Zhang, M., Ihme, K., Drewitz, U., and Jipp, M. 2021. Understanding the Multidimensional and Dynamic Nature of Facial Expressions Based on Indicators for Appraisal Components as Basis for Measuring Drivers' Fear. *Frontiers in Psychology* 12, 622433.

Identifying Team Process Behaviours of Clinical Nursing Teams Working on Lower-acuity Hospital Units

L. Whitehair, S. Provost and J. Hurley

Faculty of Health, Southern Cross University, Coffs Harbour, Australia. steve.provost@scu.edu.au

Introduction

Team-based nursing has been promoted as a means to improve the quality of patient care and safety and has been widely implemented in the Australian health care system. A systematic review of evidence for the effect of team processes used by clinical healthcare teams on team performance, patient safety, and the relationships between these factors was conducted by Whitehair [1]. Team processes make up the ways in which members interact with other members to organise task work to meet collective goals [1]. Whitehair found that, in general, the evidence suggests that team processes contribute significantly to clinical team performance and/or patient safety, as reported elsewhere [2, 3]. However, despite the identification of a large number of articles investigating team processes, few of these utilised a quantitative approach. In addition, a consensus on the relationships between team processes and team performance and/or patient safety was made difficult due to the variety of teams, clinical settings, team processes, and outcomes measured. Two thirds of the identified studies focused on interdisciplinary clinical teams working in high-acuity clinical settings (largely surgical operating teams) and/or medical emergency teams. Few studies investigated interdisciplinary clinical teams working in lower-acuity clinical settings. The two studies involving lower-acuity settings focused on the impact of training and emergency situations, and utilised simulated environments to conduct observations. The question of whether findings from studies situated in high-acuity settings can translate to lower-acuity settings remains unanswered.

Despite assertions that nurses are pivotal to patient safety, it is not known how clinical nursing teams working in general hospital units go about their everyday work to minimise adverse events and keep patients safe. Dekker et al. [4] argue that safety can be envisaged as something that organisations do in order to be constantly on the alert to the possibility of failure. Reliable performance can only be determined through a review of the collective behavioural processes enacted by clinical workers within hospital organisations. In order to conduct such a review an objective and efficient tool for the measurement of relevant behaviours as they occur in the nursing unit is required. However, no such tool exists for team performance of single nursing teams working in lower-acuity settings. Here we describe the development of such an instrument, the sequential behavioural observation tool for nurses (SBOT-N).

The Sequential Behavioural Observation Tool for Nurses (SBOT-N)

A qualitative analysis of team behaviours observed in a high-performing, lower-acuity, paediatric unit employing a collaborative nursing model is reported by Whitehair, Hurley and Provost [5]. Team-relevant behaviours were categorized using a mixture of a-priori theoretical expectations [6, 7] and researcher-based interpretation where these expectations were not met. This analysis revealed 20 team processes able to categorise the various team-related behavioural events taking place on the unit [1]. However, in the exploratory study described below two of these team processes were not observed, and the SBOT-N thus consists of 18 coded events constituting team processes evidenced in specific behaviours. Table 1 contains a list of these events and their definitions. Specific instances of behaviours corresponding to these events are provided in detail by Whitehair [1]. For example, observing a team leader developing strategies and participating in goal setting with team members or seeking input/viewpoint/comment/advice regarding a situation/event from team members would be coded as "1: *Actively collaborating to meet team goals*", while willingness to help team members when asked and willingness to admit mistakes regarding individual actions would be coded as "2: *Showing respect and looking after each other*".

Table 1. Coded team process events of the SBOT-N derived from the qualitative study [5] and their definitions.

Code	**Team Process**	**Definition**
1	Actively collaborates to meet team goals	TM demonstrates the belief in the importance of team goal's over individual TMs' goals. TM takes into account other viewpoints and behaviour during group interaction.
2	Showing respect and looking after each other	The shared belief that TMs' interests will be protected and everyone will be treated with respect.
3	Creating a happy work atmosphere	TL uses humour or shares a joke when working with other TM to create or establish positive working relationships.
4	Structuring work processes	TL determines the sequencing for actions to be carried out, coordinates how these activities will be paced and plans future steps for the team.
5	Problem solving	TL considers a problem presented by a TM, provides an interpretation of the problem, and looks for a solution and goals to set.
6	Requesting task/procedure relevant information	TL proactively acquires task/procedural relevant information.
7	Anticipating TMs' needs	TM anticipates other TM's needs based on an accurate understanding about their responsibilities. TM shifts workload among members so that a balance can be achieved when workload or pressures are high.
8	Reviewing individual patient/processes.	TL considers all patient or process related elements and where needed interprets and verbalises understanding of a patient matter to other TM. TM identifies any future problems and shares this information with regard to any future plans for the patient and/or any constraints.
9	Developing TM knowledge	TL/CNE/NUM provides information to other TM about human disease/anatomy/physiology, or a task/procedure related to work matters
10	Task distribution	TM assigns tasks or roles to other team member
11	Cross-check communication	The exchange of information between TMs where the message is acknowledged and/or clarified.
12	Ordering a TM	TM gives a direct order to other team member.
13	Developing TM skills	TL/CNE/NUM provides information about a skill or demonstrates how to perform a skill to other TM.

14	Establishing performance expectations and/or acceptable interaction patterns.	TL/CNE/NUM verbalises expectations about required behaviours for working in the unit to other TM.
15	General situation assessment	TL/NUM considers all team related elements and where needed interprets and verbalises understanding of a matter to other TM. TM identifies any future problems and shares this information with regard to any future plans and/or any constraints.
16	Adjusting strategies	TM adjusts strategies through the use of backup behaviour and/or reallocation of team resources. TM alters a course of action or team repertoire in response to changing conditions (internal or external).
17	Building team confidence and motivation	TM promotes confidence or motivates other TM in relation to an aspect of teamwork.
18	Regulating team emotions	TM demonstrates regulation of TM's emotions with regard to accomplishing team goals.

Note. Some team processes may be evidenced in the behaviour of any team member, while other relate to specific roles, such as team leaders or unit managers. The definition includes reference to the relevant roles using the following abbreviations: NUM = Nurse Unit Manager; CNE = Clinical Nurse Educator; TL = team leader; TM = team member

Whitehair [1] conducted a pilot/proof of concept study utilising the SBOT-N. She followed nursing teams in surgical and medical units over 12 shifts (generally 8-hours in length), recording the occurrence of team process events on a hand-held electronic device. The study had approval from the relevant health authority (North Coast NSW Human Research Ethics Committee, approval number LNR 127).

The total number of coded events is shown in Table 2. The most frequently observed team process was *active collaboration to meet team goals*, followed by *showing respect and looking after each other* and *creating a happy work atmosphere*. Some behaviours, such as those relating to *regulating team emotions and building confidence and motivation* were very infrequently observed. It should be noted that team leaders were not necessarily in any kind of supervisory relationship with team members, perhaps explaining why events contributing to social cohesion were more frequently observed than those which would traditionally be associated with meeting organisational training goals.

Table 2. Overall frequency of coded events across 12 shifts

Code	Team process events	Total
1	Actively collaborates to meet team goals	315
2	Showing respect and looking after each other	75
3	Create a happy work atmosphere	67
4	Structuring work processes	57
5	Problem solving	56
6	Requesting task/procedure relevant information	51

7	Anticipating team members' needs	49
8	Reviewing individual patient/processes	43
9	Developing team member knowledge	39
10	Cross-check communication	32
11	Task distribution	25
12	Ordering a team member	25
13	General situation assessment	20
14	Establishing performance expectations and /or acceptable interaction patterns	18
15	Adjusting strategies	13
16	Developing team member skills	9
17	Building team confidence and motivation	8
18	Regulating team emotions	1

Evidence for the ability of the tool to discriminate levels of team behaviours between different teams is provided by examining the occurrence of coded events across different shifts, as well as the number of team process events occurring in total for a shift. Figure 1 shows the occurrence of each team process across a surgical unit afternoon shift (approximately 2-10 pm). Each team process is indicated by differently coloured dots. Actively collaborating to meet team goals (Code 1, blue dots) occurs across the entire shift, with the exception of the last hour (during which the nursing staff would have been writing up their notes and conducting hand-over). A number of other team processes occur relatively frequently also. This can be contrasted with the pattern for another Surgical Unit afternoon shift shown in Figure 2. There are far fewer team processes occurring, and it is noticeable that interactions decline as the shift progresses. Finally, Figure 3 shows the occurrences of team processes for a Medical Unit afternoon shift in which team behaviours appear to be almost entirely absent. This interpretation is correct. On this shift one team member was allocated to "specialling" a patient, which involved being isolated in the patient's room for the entirety of the shift. The other two, very experienced, team members employed patient allocation to the remaining load, and worked largely independently with the exception of covering for each other for meal-breaks. The total number of coded team events for each shift is shown in Figure 4. It is clear from this figure that the SBOT-N is differentiating between shifts, and one would assume the degree of teamwork being employed in those shifts. These raw scores cannot be directly interpreted, however, as they will have been influenced by the length of the shifts, by the degree of business, by the number of team members, and by the relative levels of seniority and capabilities of the staff. How one might account for these factors in order to obtain a more exact measure of overall team behaviour awaits further research.

Figure 1. Occurrences of coded team process events across a single surgical unit afternoon afternoon shift (Shift 9).

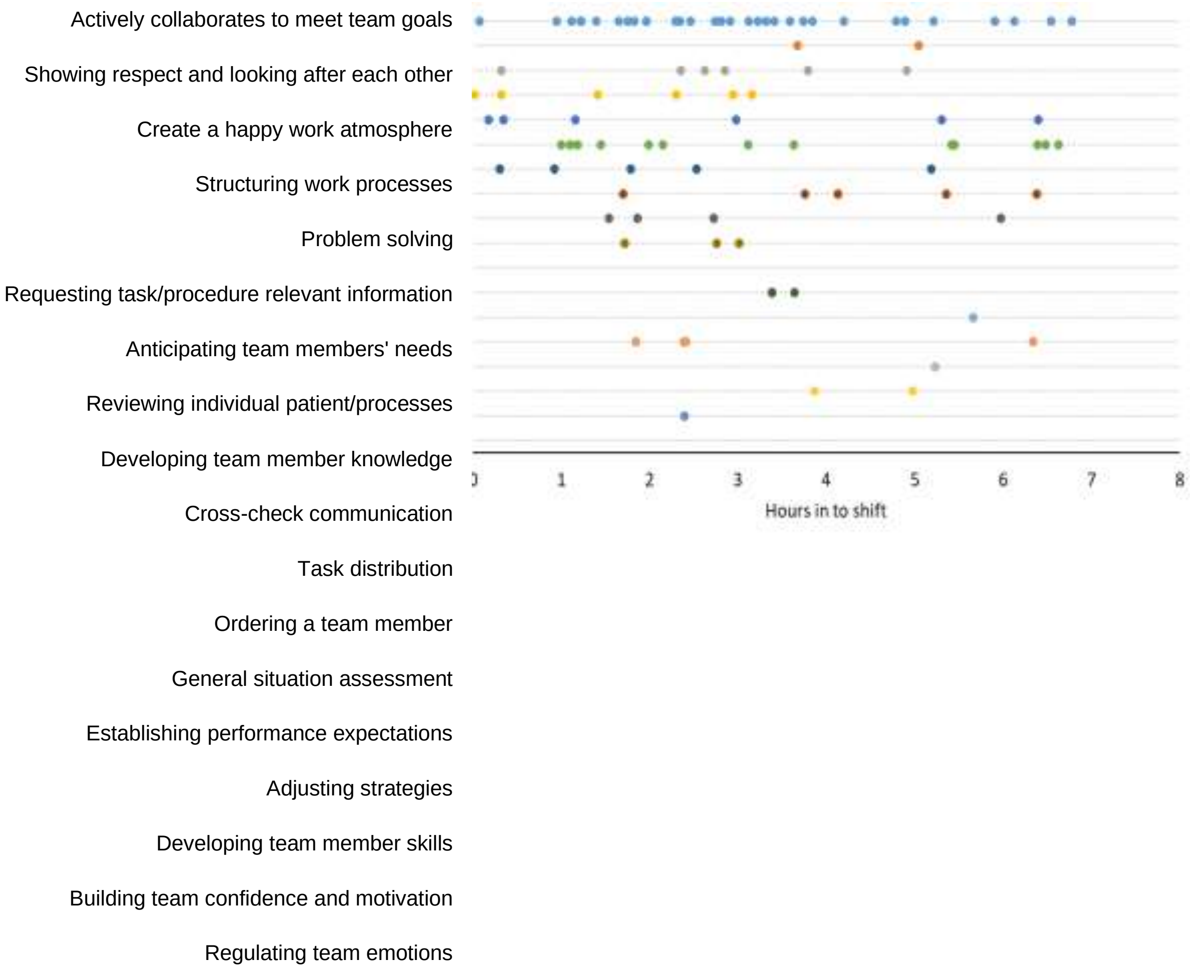


Figure 2. Occurrences of coded team process events across a single surgical unit afternoon afternoon shift (Shift 2).

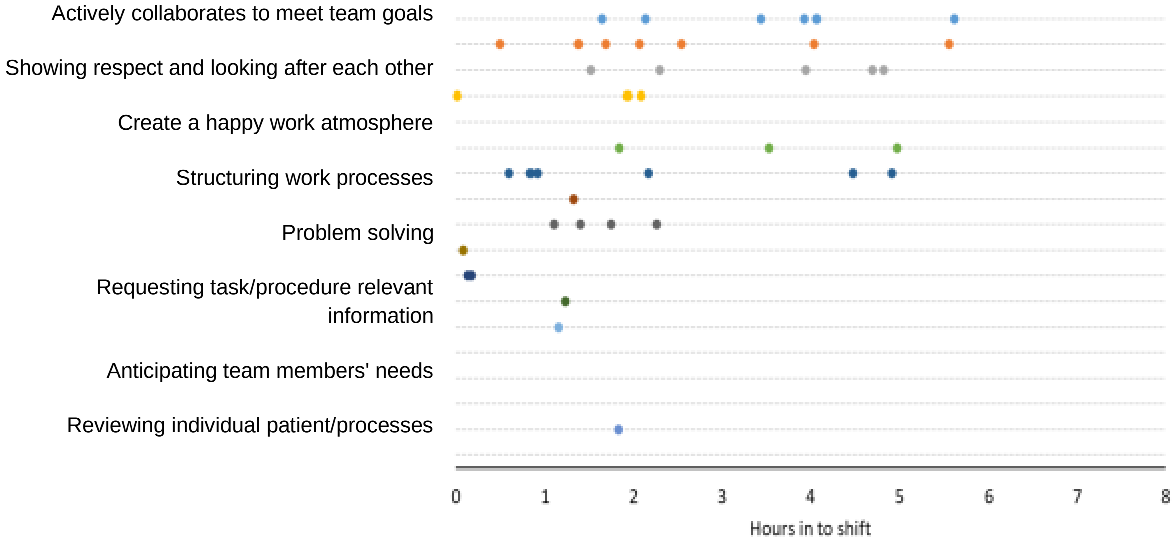


Figure 3. Occurrences of coded team process events across a single medical unit afternoon afternoon shift (Shift 11).

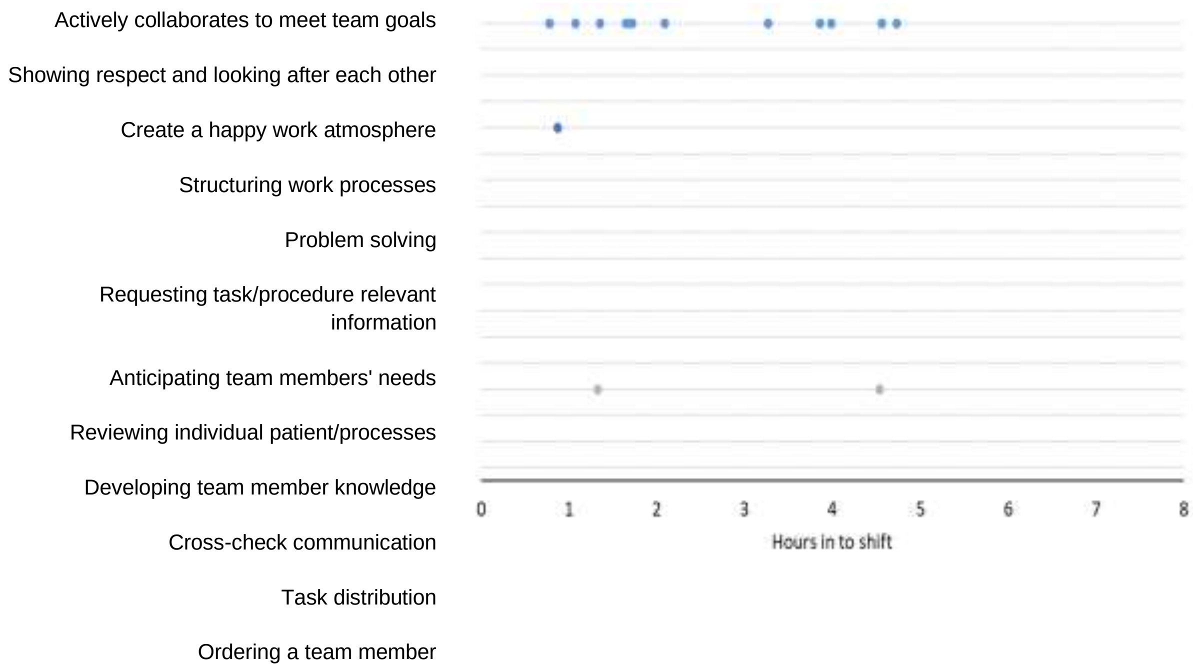
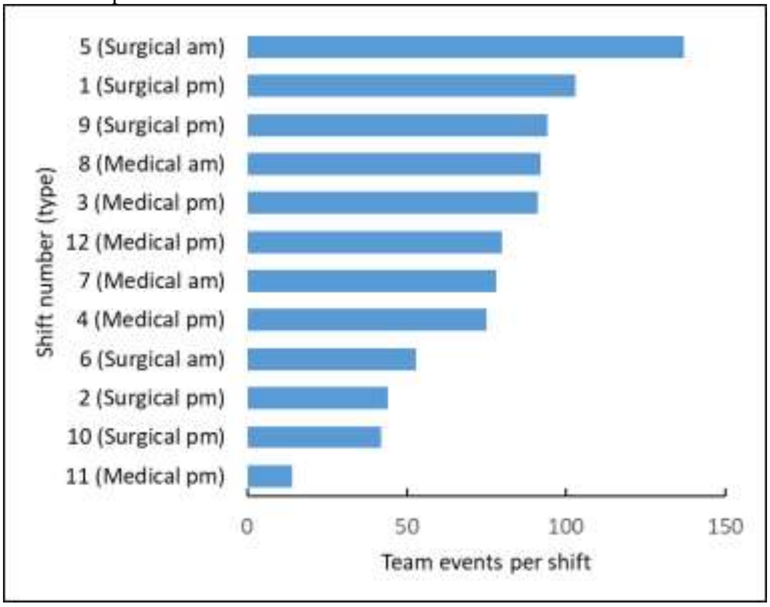
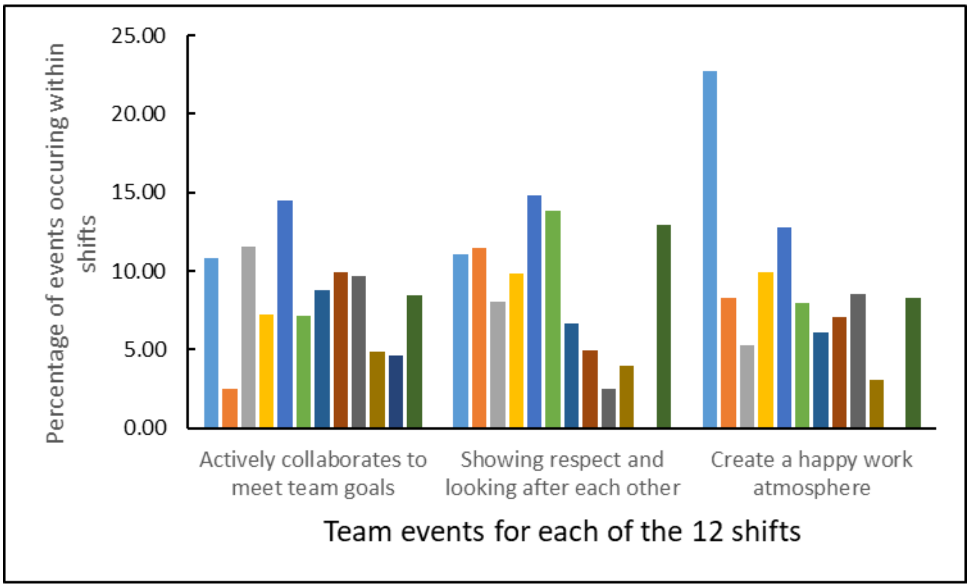


Figure 4. Overall number of team process events observed in each of the 12 shifts examined.



In addition to considering the overall prevalence of team behaviours, the SBOT-N also allows for differences in the utilisation of separate team processes to be examined. Figure 5 shows the percentage of team processes observed in each shift for three of the coded team process events for each of the 12 shifts. It can be seen, for example, that *Creating a happy work environment* constituted a substantial percentage of the coded events for Shift 1 (light blue bars), but not so for Shift 3 (gray bars). However this relationship was almost reversed for *Actively collaborates to meet team goals*, where this team process was more prevalent in Shift 3 than in Shift 1.

Figure 5. The percentage of total team process events contributed to by 3 coded events (*Actively collaborates to meet team goals*, *Showing respect and looking after each other*, and *Create a happy work environment*) for each shift. Different shifts are indicated by different coloured bars.



Examination of the individual team processes leads to consideration of how they might be related. There are a number of such relationships, and a complete description is beyond the scope of this paper. The following examples are provided to illustrate the potential of the SBOT-N for more detailed analysis of theoretically relevant differences. A more complete description is provided by Whitehair [1]. In a number of cases the group processes were correlated with each other quite highly. For example, there was a strong relationship between *Actively collaborates to meet team goals* and *Reviewing individual patient* processes (r = .79), indicating that these two team process events shared more than 60% of the variance. However, the relationships between some team processes appear somewhat paradoxical at first. For example, the correlation between *Structuring work processes* and *General situation assessment* was negative (r = -.59). Specific behaviours relating to *Structuring work processes* include “Provide information about what actions a TM should do next” and “Plan goals and direct TM activities”, while those relating to *General situation assessment* include “Ask for updates on unit happenings” and “Check on the progress of TM with regard to their work activities”. It seems plausible that a team leader who provides a more structured work environment may not require to be updated as frequently as one who does not do so. The discovery that some team processes may be inversely related to each other is obviously of some theoretical interest, and would not have been revealed without the quantitative approach taken here. The ability of the SBOT-N to reveal such subtle interrelationships would appear to be a strength. Our hope is that further data collection from a more varied set of sources than described here may inform the understanding of different leadership styles within a team environment.

Conclusions

The SBOT-N allows for dynamic and objective measurement of team process behaviours in real time over extended periods. Further development and validation of the tool is clearly warranted. Inter-rater reliability of the coding system is the most important first step in such a process. This could be achieved through the video recording of team interaction, allowing observation by multiple coders. If successful the SBOT-N has potential to support research regarding the nature of teamwork and theories pertaining to it. By linking evidence for teamwork processes with outcomes it also has the potential to support evidence-based efforts to increase patient safety and staff satisfaction. Information gained through its use may be utilised for the provision of feedback in education and training scenarios. Finally, it is possible that with slight adjustments the tool may be appropriate for examining team process behaviours in “everyday” workplaces beyond that of the lower-acuity hospital unit.

References

1. Whitehair, L. (2021). *Identifying Team Process Behaviours of Clinical Nursing Teams Working on Lower-acuity Hospital Units: Development of an Observation Tool*. Doctor of Philosophy (PhD), Southern Cross University. http://doi.org/10.25918/thesis.196

2. Künzle, B., Kolbe, M., & Grote, G. (2010). Ensuring patient safety through effective leadership behaviour: A literature review. *Safety Science*, **48**, 1-17. http://dx.doi.org/10.1016/j.ssci.2009.06.004

3. Schmutz, J., & Manser, T. (2013). Do team processes really have an effect on clinical performance? A systematic literature review. *British Journal of Anaesthesia*, **110**, 529-544. http://dx.doi.org/10.1093/bja/aes513

4. Dekker, S., Hollnagel, E., Woods, D., & Cook, R. (2008). *Resilience engineering: New directions for measuring and maintaining safety in complex systems*. Final report. Lund, Sweden: Lund University School of Aviation.

5. Whitehair, L., Hurley, J., & Provost, S. (2018). Envisioning successful teamwork: An exploratory qualitative study of team processes used by nursing teams in a paediatric hospital unit. *Journal of Clinical Nursing*, **27**, 4257-4269. http://dx.doi.org/10.1111/jocn.14558

6. Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, **26**, 356-376. http://dx.doi.org/10.5465/AMR.2001.4845785

7. Salas, E., Shuffler, M. L., Thayer, A. L., Bedwell, W. L., & Lazzara, E. H. (2015). Understanding and improving teamwork in organizations: A scientifically based practical guide. *Human Resource Management*, **54**, 599-622. http://dx.doi.org/10.1002/hrm.21628

Animal Welfare

Home cage monitoring of group-housed animals in highly enriched enclosures

G. Ciminelli[2], J. Andres[2], K. Stupples[1], C. L. Witham[1]*

1 Centre for Macaques, MRC Harwell Institute, Salisbury, UK. c.witham@har.mrc.ac.uk

2 Biosciences Institute, Newcastle University, Newcastle-upon-Tyne, UK

Home cage monitoring is a key goal for animal welfare as it allows behavior assessment in a familiar setting without disturbance. There are challenges in adapting existing methods to species such as primates that are housed in groups in large, highly enriched enclosures. We propose an object segmentation method that detects the animals and many different objects in their enclosure. We demonstrate how this can be used with macaques to detect simple behaviors including foraging.

Introduction

Home cage monitoring methods are useful for assessing animal behavior in a familiar setting without human presence or other disturbance to the animals. Home cage monitoring methods are both a key method for improving animal welfare as they cause less disturbance to the animals [1] and also a method for assessing animal welfare. The majority of the current monitoring methods are aimed at rodents especially mice [2]. These methods may require custom cages and work best for singly housed animals although some can work with group-housed animals [3].

When using home cage monitoring for welfare monitoring it is important that the methods do not impact the welfare of the animals that are being monitored. For example the level of enrichment in the home cage or enclosure should not be reduced to facilitate the monitoring. For species such as primates the home enclosures may include outdoor space and are usually heavily enriched with a variety of different items [4]. These include large pieces of fixed enrichment such as platforms and poles, moveable enrichment attached at one or two points such as hoses and swings and freely moving objects such as balls and toys. For primates with access to outdoor areas there may also be a variety of plants including trees and bushes. This makes developing automated methods for primates and many other species difficult.

Existing methods for monkeys tend to be based on singly housed animals [5] or only cover a small part of an enclosure [6]. Primates can be housed under many different conditions from single housing in some laboratories to large outdoor enclosures in primate breeding centers, zoos and sanctuaries. Developing models that can be used under many different conditions is challenging but by focusing on some simple group-level measures such as foraging and enrichment use we hoped to develop a flexible home enclosure monitoring approach that can be applied under many housing conditions.

The aim of this project was to develop a new approach to monitoring group-housed rhesus macaques. The new method must be flexible, useable on a single camera and capable of extracting the basic group-level measures. We aimed for a single model that could be used on multiple groups of animals that varied in age from newborn to fully grown adults rather than tailoring individual models to each enclosure and group of animals. We have developed an approach that uses the Segment-Anything-Model [7] to quickly create a ground-truth dataset for both animals and objects in the enclosures and then the YOLO v8 model from Ultralytics [8] to train a custom segmentation model.

Methods

Animals

The video used to develop and validate the methods in this project was recorded at a rhesus macaque breeding center (MRC Centre for Macaques, UK). The center houses 170-200 rhesus macaques in groups of 3-20 (aged between 0 and 16 years). All the macaques are F2 generation or greater and were bred in the UK. The groups include breeding groups (consisting of a single adult male, 2-7 adult females and their offspring) and single sex stock groups (2-12 juvenile and adolescent animals). The animals are housed in enclosures measuring 9m (length) x 3m (width) x 2.4m (height) with an adjoining caging area. All enclosures are similar in shape and size but the enrichment varies between enclosures and is changed in each enclosure every 4 months. The artificial lights come on at 7am and go off at 7pm and the enclosures also have natural light through a large bay window. The macaques are fed once a day with a mixture of complete primate diet, forage mix and two types of fruit and vegetables per day. Videos are recorded from a single CCTV camera on each enclosure for husbandry purposes. The colony is licensed by the UK Home Office and the project was reviewed by the colony's Animal Welfare and Ethics Review Body.

Equipment

There was one CCTV camera per enclosure (Axis P1455-LE; 22 enclosures in total). The video was sent via power other ethernet cables to a central hub (Axis Camera Station S1148). Up to 12 months of video are stored and are accessed via the Axis Camera Station Client. Videos were recorded at 1280 x 720 pixels and 15 frames per second and the videos for analysis were saved in mp4 format on network-attached storage. An example video frame is shown in Figure 1A. Two custom Linux-based deep-learning machines were used for training the models and analyzing the videos (www.scan.co.uk; 3XS Deep Learning DBP G2-18C). Each machine had two NVIDIA GeForce RTX 3080 Turbo V2 10GB graphics cards. All analysis was carried out in Python using the following packages:

- YOLO v8 (version 8.0.162) [8]
- Opencv (version 4.8.0)
- Segment-Anything-Annotator [9]

Ground truth labelling

We used the Segment-Anything-Model [7,10] in an implementation of the label-me image annotator software [9,11] to segment the images and create a ground-truth image set. A random selection of 280 frames from 20 enclosures were extracted from the videos, resized to 640 x 640 pixels, and saved as PNG image files. We labelled 27 different objects/areas including the monkeys, the floor, various enrichment objects (both fixed and freely moving), the floor and the door. An example of the different labels is shown in Figure 1B.

To check the effectiveness of using SAM three people used the image labeler to label the same set of 20 images. Half the images were labelled using manual segmentation (drawing an outline around each object) and half were labeled using the SAM labeler. Using the SAM labeler reduced the amount of time required to label the images (time required to label images reduced by between 20 and 45% for three labelers) and also improved consistency between the three labelers. A custom written python script was used to convert the output of the image labeler (JSON files) to the correct format for the Yolo segmentation model (text files).

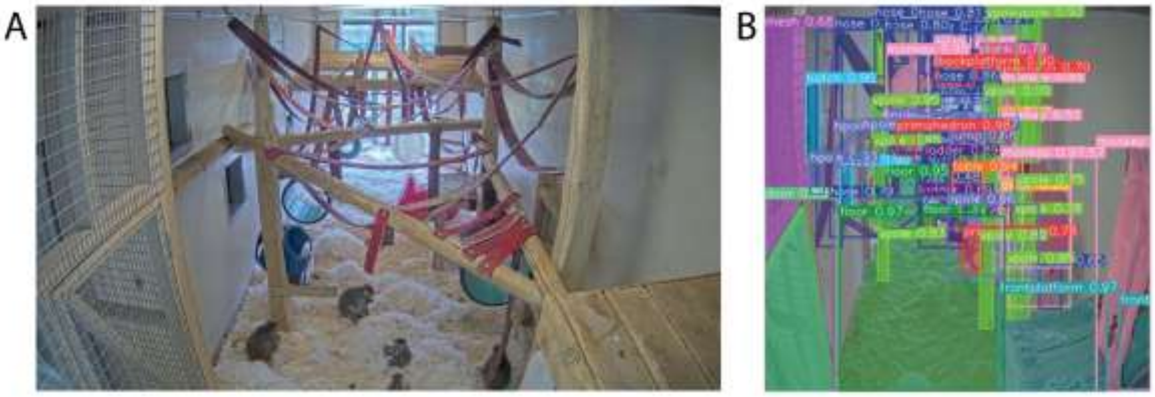


Figure 1. Examples of CCTV images and labelling. A) Example of a single frame from video. B) Example of a resized and Yolo-labelled frame.

Segmentation model

The segmentation model was trained using the YOLO v8 framework. The ground-truth dataset was divided into 210 frames for training, 35 frames for validation and 35 frames for testing (75%:12.5%:12.5% split). The default YOLOv8n-seg model was used as the base and the training time was approximately 1 hour. The precision-recall curve and confusion matrix were used to assess the model performance across the different objects. The mean-average precision for all objects was 0.721 at a threshold of 0.5. In general the objects closest to the camera and the objects that kept a similar shape (such as the fixed enrichment) performed best.

Interpretation of segmentation output

The YOLO segmentation model provides four measures for each object detected, these are the segmentation mask, the bounding box, the object class and the confidence score. To save on storage space we converted the segmentation mask to a contour using the Opencv. There is a challenge in interpreting this output in a meaningful, behavior related way. So far we have used three different methods to interpret the output of the segmentation model (summarized in the diagram in Figure 3):

1. Mask overlay. For the fixed objects such as the floor and the wooden platforms and poles an average mask was created by converting the contours back to a binary mask and averaging across frames. We calculated the amount of overlap between the mask for a single monkey and the average mask for the fixed object and if it was above threshold counted this as the monkey being on or in front of that object. This is the method we used for counting the number of monkeys on the floor foraging (see Application).
2. Simple changes in one dimension. The door of the enclosure is located in the bottom of the video frame. When the door opens the x-coordinate of the center of the mask changes. We used the simple change in center x-coordinate to detect when the door opened. This gave the time when the animals were fed (see Application).
3. The third method is to calculate the difference in x-y position of the mask centroid across frames. The sum of the absolute differences gives the total amount of movement of the object. This method works best when there is only one of the object in the room (otherwise tracking is needed to monitor the movement of a single object or animal). Calculating the movement of an enrichment object can be used as a proxy measure of how much the enrichment is being used by the monkeys.

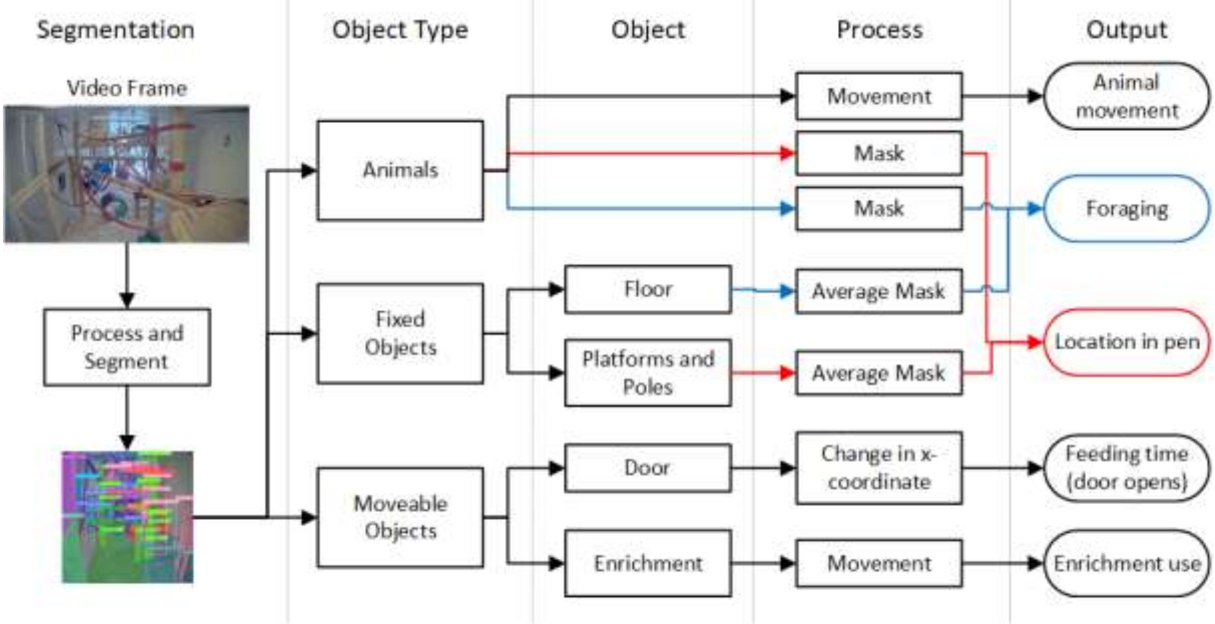


Figure 2: Diagram showing processing of segmentation output.

Application

We used this approach to count the number of animals on the floor using the mask overlap between the individual monkey masks and the average floor mask. This serves as a proxy measure of the amount of time the animals spend foraging as there is a strong relationship between the monkeys being on the floor and them foraging (macaques tend to prefer to be at height when resting/socializing). We took the video footage for a day and ran the object segmentation model on one frame per second (the video was recorded at 15 frames per second). The time the care staff entered the enclosure to feed the animals was determined by measuring deviations in the x-coordinate position of the door. We then counted the number of animals foraging over the two hours after feeding. Figure 3 shows the number of animals foraging over a 12 hour day (the output has been smoothed with a Gaussian kernel for clarity). We compared the number of animals foraging on different days of the week for 12 groups (different fruit and vegetables are fed each day). We found that foraging was highest on Saturdays when the animals received spinach and watercress and lowest on Sundays when the animals were fed whole oranges. As a result of this analysis the fruit and vegetable schedule was changed to encourage more foraging on low-forage days. We are currently analyzing whether the change has been successful using the approach detailed in this paper.

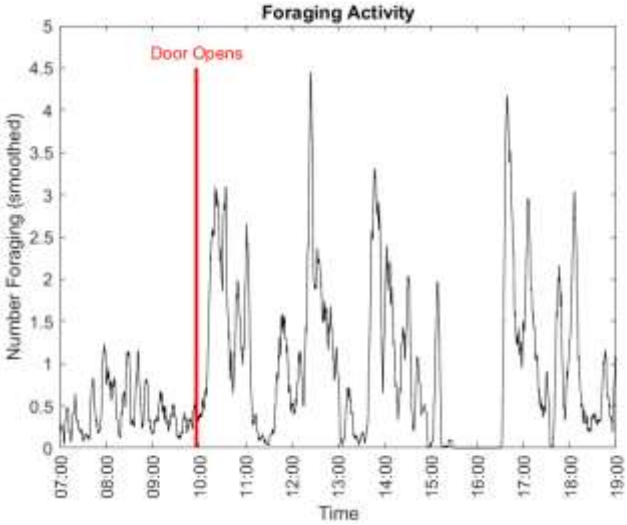


Figure 3: Number of animals foraging over a single day

Discussion

The combination of the segment-anything-model for ground-truth labeling and the YOLO v8 segmentation model allows for the rapid labeling and training of an object segmentation model for a complex enriched environment. It also provides a good method for segmenting animals that can adopt many different postures. The model is robust to changes in the enrichment in the enclosure and movement of different objects and could be adapted to many different species and enclosure types. It is affordable requiring just a single camera per enclosure and can be trained and run on a machine with a single GPU card. Using this model we can estimate the number of animals foraging and the amount of times a piece of enrichment is used. This approach does have several drawbacks:

1. We cannot identify individual animals with this approach so all measures are at a group level.
2. The measures of behavior are not direct measures of behavior.
3. We make a number of assumptions such as that animals on the floor are foraging and that the movement of an object means it is being used by an animal.
4. With a single camera the animals are often occluded especially when they are distant to the camera.

Some of these disadvantages could be solved by either adding more cameras or combining this approach with other methods such as face recognition.

References

1. Gaburro S, Winter Y, Loos M, Kim JJ and Stiedl O (2022) Editorial: Home Cage-Based Phenotyping in Rodents: Innovation, Standardization, Reproducibility and Translational Improvement. Frontiers in Neuroscience. **16**:894193.
2. Kahnau, P., Mieske, P., Wilzopolski, J. et al. (2023). A systematic review of the development and application of home cage monitoring in laboratory mice and rats. BMC Biology **21**, 256

3. Bains RS, Forrest H, Sillito RR, Armstrong JD, Stewart M, Nolan PM and Wells SE (2023) Longitudinal home-cage automated assessment of climbing behavior shows sexual dimorphism and aging-related decrease in C57BL/6J healthy mice and allows early detection of motor impairment in the N171-82Q mouse model of Huntington's disease. Frontiers in Behavioral Neuroscience **17**:1148172
4. Coleman K, Novak MA. Environmental Enrichment in the 21st Century (2017). Institute for Animal Research Journal. **58**:295-307.
5. Bala, P.C., Eisenreich, B.R., Yoo, S.B.M. et al (2020). Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. Nature Communications **11**, 4560
6. Witham, C.L. (2018). Automated face recognition of rhesus macaques. Journal of Neuroscience Methods **300**, 157-165
7. Kirillow, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiai, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R. (2023). Segment Anything. arXiv 2304.02643.
8. https://docs.ultralytics.com/. Accessed 1 January 2024.
9. https://github.com/haochenheheda/segment-anything-annotator. Accessed 1 January 2024.
10. https://segment-anything.com/. Accessed 1 January 2024.
11. http://labelme.csail.mit.edu/Release3.0/. Accessed 1 January 2024.

Quantifying facial expressions of the horse with optical motion capture and surface electromyography; a proof of concept

I.H. Smit[1], Y. Mellbin[2], K. Ask[2], N.C.R. te Moller[1], J. Lundblad[2]

1: Clinical Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands. i.h.smit@uu.nl

2: Dept. of Anatomy, Physiology and Biochemistry, Swedish University of Agricultural Sciences, Uppsala, Sweden

Introduction

Facial expressions are often assessed in the study of affect due to the distinction between voluntary control and emotional expression [1] and the effect of emotional experience on contractions of certain facial muscles [2]. The core purpose of facial expressions is to convey information to conspecifics through social functions, but contractions in facial muscles may stem from physiological or emotional origins as well [3]. The use of facial expressions remains one of the most extensive methods of communication in humans [4]. In the study of animals, which lack self-report, mainly the emotional part of the neuroanatomy of facial expressions has been proposed as a valuable tool for both clinical[5] and welfare applications [6]. Furthermore, animals present an extensive facial repertoire [7]. The use of face-based methods to evaluate affect in animals has received much attention, and the recognition and prevention of pain has had a large impact on research in animals [8]. Especially horses have attracted attention in this field, since pain evaluation of prey animals is more difficult due to them hiding potential pain-related behaviours [9]. Some of these methods have been developed into clinical tools involving both grimace scales [10–12] and dynamic-based research tools employing Facial Action Coding systems [5], which relies less on subjective interpretation of grimaces.

With the Equine Facial Action Coding System (EquiFACS) [12], codes are assigned to morphological changes due to muscle contractions in the equine face, without attributing any interpretation of the source or underlying reason for the contraction, such as emotion or voluntary expression. However, this system requires training and is time-consuming to work with, requiring frame-by-frame analysis of video material. Furthermore, grimace scales and FACS both require some sort of subjective judgement to assign categorical values of the facial expressions produced. Therefore, multimodal approaches, using kinematic analysis of motion together with physiological and electromyographic data, have been proposed to better understand the facial expressions of emotions in humans [4]. Furthermore, no method for quantifying intensity of the muscle contraction exists in most of the FACS-systems for animals.

This study proposed the concept of quantifying facial expressions of the horse using surface electromyography (EMG) as well as describing a method for collecting such data in order to provide an objective tool to quantify and validate visual tools. Surface EMG measures muscle activation non-invasively using surface electrodes placed on the skin over muscles of interest [14]. While surface EMG measures activation of the muscle, not all muscle activations lead to contraction nor to visible movement and different levels of activation might lead to different movement amplitudes [15]. To compare the presence of movement and movement amplitude to the degree of muscle activation, optical motion capture (OMC) was used to determine the movement of the skin overlying the muscles of interest in the three-dimensional space as a proxy for facial expressions. A proof of concept regarding the facial expressions around the eye (i.e. blinking or raised eyebrow) is described, since that is a movement that occurs often and entails a clear noticeable change. We hypothesized that 1) eyelid movement can be captured using both surface EMG and OMC and 2) the amplitude and/or speed of the movement correlate with surface EMG amplitude.

Material and methods

Study design

Eight adult standardbred trotters of similar conformation were included in the study. The horses underwent a general clinical exam as well as a specialized neurological exam of the cranial nerves in order to ensure that the horse had proper function of the facialis nerve, which controls the facial muscles and to exclude detrimental influences from impaired function of other cranial nerves. The horses were introduced to the experimental area in advance in order to acclimatize them to the environment before the study started. The horses were placed in a small confinement which they were already accustomed to, with the head placed outside of the confinement and they were always accompanied by another horse that was placed in the same room. For this proof of concept, data from neutral states were evaluated, even though the experiment involved the horses experiencing certain stimuli as well.

Data collection

Prior to applying the electrodes, the skin surface overlaying the muscles of interest on the left side of the head was clipped and thoroughly cleaned (ethanol 80%) to improve signal quality and electrode adhesion. Six bipolar leads were placed over five muscles that are associated with facial expressions: m. frontalis, m. masseter, m. zygomaticus, m. caninus, m. nasolabialis – ventral root and m. nasolabialis – dorsal root (figure 1). The electrode locations were standardized between horses. For the m. frontalis, electrodes (pre-gelledAg/AgCl; 3mm electrode; 22mm inter-electrode distance; Ambu BlueSensor N) were placed at 30% distance between the corner of the eye and the point of insertion of the forelock. Surface EMG data were collected in bipolar derivation (referred to the average reference), pre-amplified, sampled at 2000Hz and A/D converted with a 24-bit resolution (SAGA, TMSi, Oldenzaal, the Netherlands). A ground electrode was placed over the spina scapula on the left side of the horse. Surface EMG and OMC data were frame-synchronized using a synchronization signal sent from the surface EMG equipment at 200Hz and data collection lasted for approximately 30 minutes.

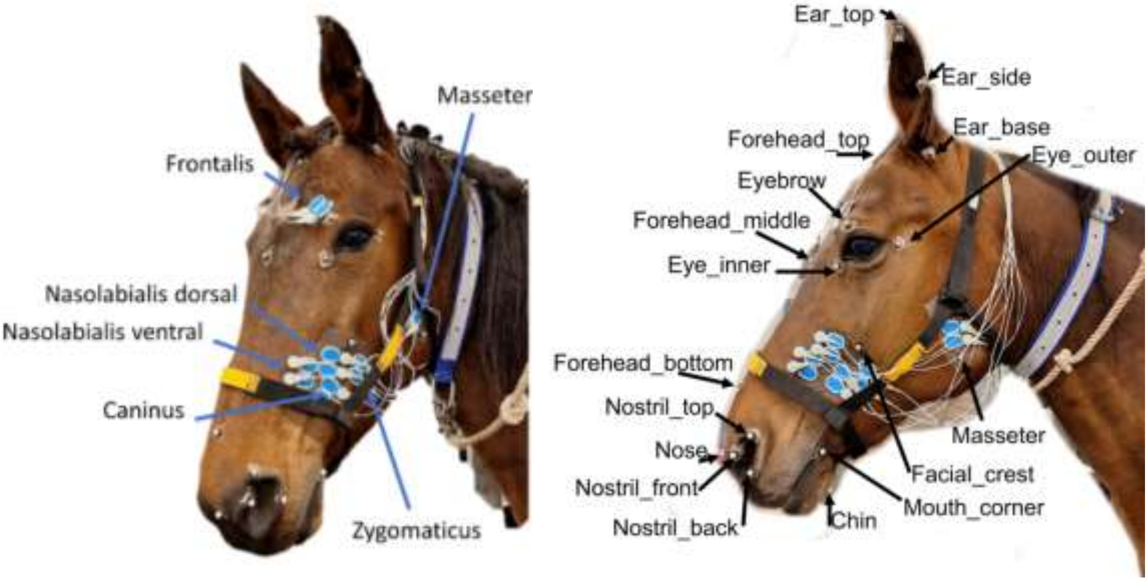


Figure 1. Visual representation of the placement of surface EMG electrodes (left) and OMC markers (right).

Seventeen reflective, spherical 8mm OMC markers were attached to both 'rigid’ and 'moving’ landmarks on the head of the horses’ head using double-sided adhesive tape (figure 1). Rigid landmarks used were: over the midline of the face (forehead_top, forehead_middle and forehead_bottom), at facial crest, the base of the ear and at both corners of the left eye (eye_inner and eye_outer). Moving landmarks used were: on the ear (ear_top and ear_side), on the eyebrow, the lower jaw (masseter), chin, corner of the mouth and around the nose (nose, nostril_front, nostril_back and nostril_top). An Optical Motion Capture (OMC) system with six 2-megapixel IR cameras (figure 2; Miqus M3, Qualisys AB, Sweden) running at 200 Hz was used for capturing the movement of the markers. The head was also recorded in full HD at 25 Hz with one video camera placed on the right side and one straight in front of the horse. The video cameras (Miqus Video Color, Qualisys AB, Sweden) were synchronized and calibrated with the IR cameras in the motion capture system.

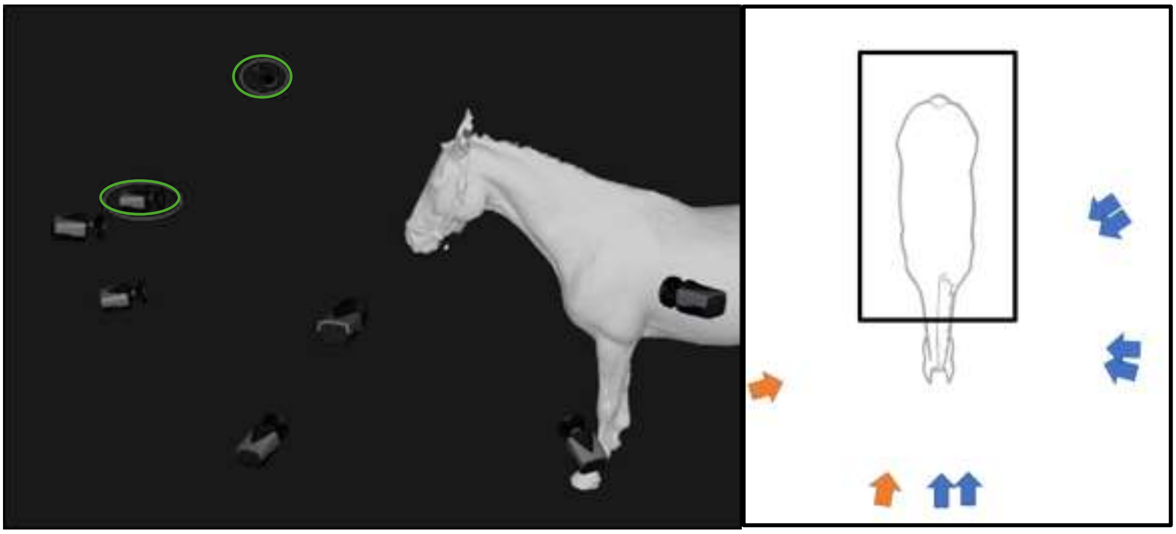


Figure 2. Illustration of placement of optical motion capture cameras. Circled cameras (left) and orange arrows (right) represent video cameras, blue arrows (right) represent infrared cameras.

Data processing

All data analyses were performed in MATLAB (2022b, MathWorks, Natick, USA). For each horse, a segment of 60 seconds was selected for further analyses.

Surface EMG signals were bandpass filtered using a 4th order Butterworth filter (zero-lag; 40-450 Hz cut-off frequencies), rectified and subsequently enveloped (25 Hz; 4th order Butterworth; zero-lag). For the OMC data, the reconstruction of the 3D coordinates of each marker was automatically calculated by using motion capture software (Qualisys Track Manager, version 2023.2). Each marker was identified and labelled using an automated model and manually checked. Raw data consisting of the 3D data of the designated markers were exported to Matlab for further analysis using custom written scripts.

Results

As proof of concept to evaluate the agreement between the surface EMG and OMC data, a random sample of one horse is presented. The activation of the frontalis muscle above the eye is presented in relation to the distance between the medial corner of the eye and the eyelid, since blinking and eye movement are fairly common. The data were plotted on the same timescale to give an indication of the agreement between the two methods (figure 3). In general, the displacement of the upper eyelid marker followed the general activation and relaxation of the frontalis muscle over the course of the sample. However, similar amplitudes of the surface EMG signal did not always result in equal changes in the distance between the corner of the eye and the upper eyelid.

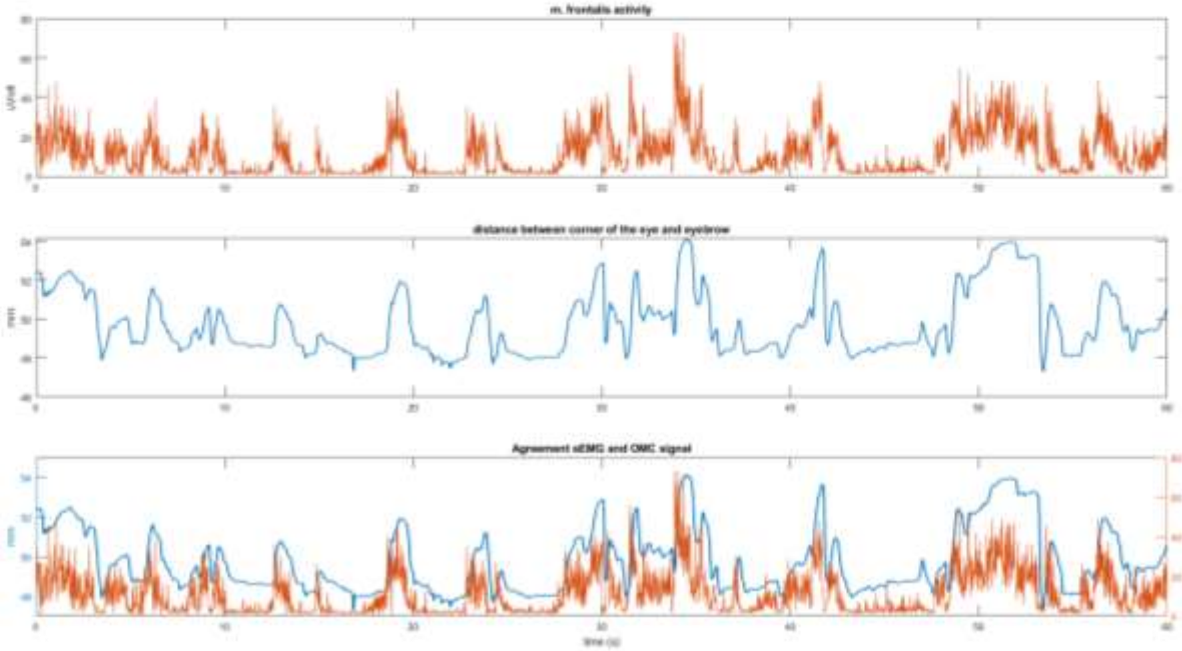


Figure 3. Muscle activity of the m. frontalis (eyebrow raiser; top panel) in µVolt and distance between the corner of the eye and the upper eyelid (middle panel) in millimetres (mm) plotted over time. The two signals are overlayed (bottom panel) to illustrate the relation between the two measurements. The data is a representative sample of one horse.

Discussion

Using surface EMG, we successfully quantified activation of facial muscles in the horse which could be correlated to macroscopic movement, as captured by an optical motion capture system. Both systems could record the facial activity independently from each other, which suggests that both methods could be viable to quantify intensity and timing of facial expressions in horses and other species. The setup in this experiment allowed for the study of other facial muscles as well and this concept enabled both systems to record muscle activation of both the lower and upper part of the face. However, further scrutiny of these activation patterns and their relation to emotion-related facial expressions is required, especially for the lower face where movement amplitudes might be smaller.

This setup could be of value when studying the effect of different emotional states and experiences on the facial expressions of the horse, ensuring quality measurements by quantifying the measured expressions. Further considerations for such a study would involve controlling the context of the experience, thereby ensuring correct biological interpretation of the signals, as well as including a large number of individuals in order to account for inter-individual variation. Furthermore, a multimodal approach, incorporating physiological markers, could provide additional pieces of information to the puzzle. It is also important to evaluate the effect of placing this type of equipment on the horse's head, as the electrodes and markers in themselves could have an effect on horses' facial expressions. In the current setup, this could be accounted for by studying the other side of the horse's face by EquiFACS using video validation. Such a validation, however, needs to take lateralization into account, since facial expressions and thus muscle activations might be different between the left and right side, depending on the active hemisphere and the type of experience studied [16].

The use of surface EMG on facial muscles has been used to monitor facial expressions related to emotion, or emotional valence in humans. Several studies have focused on the m. corrugator supercilii, the human equivalent of the m. frontalis, which is used to frown. Different groups have shown increased activity after negative stimuli and decreased activity after positive stimuli [15,17,18]. Importantly, it has been reported that the muscle activity patterns are "not always readily detected on the overt face" when video recordings of subjects are evaluated [17]. It is known that in humans, the coupling between facial muscle activation and visible movement is not perfect [19]. This may explain the observed occasions where muscle activity and measured eyelid movement were not perfectly matching and the differences in activation amplitude and movement amplitude. This shows the promise of this technique to also track subtle, non-overt changes in emotional state.

Combining OMC and surface EMG might prove beneficial, since the acquisition of surface EMG signals depends on electrode placement and anatomy of the face. Equine facial muscles are often thin in diameter and many muscles are overlapped by others [20], causing a risk of picking up signals from other muscles than the muscle of interest [21]. Different setups with smaller electrodes and/or inter-electrode distances to decrease pick-up volume, might help to further develop the application of surface EMG on equine facial muscles. Alternatively, needle EMG could be used to simultaneously capture signals from several facial muscles. However, needle EMG is an invasive technique causing pain when applied on humans [22] and thus may have a high risk of affecting the facial expressions an animal may show. Complementing the electromyographic evaluation with OMC could therefore provide a more extensive quantification of the facial activity of the whole face without needing to resort to invasive methods. Still, further review of the placement of markers and electrodes, in combination with anatomical studies and needle EMG would be beneficial to validate the OMC-system for recording facial expressions and how they relate to muscle activation patterns.

In conclusion, the setup of a simultaneous OMC and surface EMG recording could record facial expressions of the horse in a non-invasive way and both methods correlated well in regards of facial activity, both in terms of muscle activity and movement, above the eye. The concept could be of value for refining visually based tools, as well as providing intensity scores for Facial Action Coding Systems in animals. Even though both systems provided good data, the setup is complicated and sensitive for movement. As of now the setup is favourable for experimental use, further adaptations should be made for it to be suitable for use in the field and for recording facial expressions in a free environment.

Ethical statement

This study was approved by the Uppsala regional ethical committee assigned by The Swedish Board of Agriculture according to EU directives for animal experiments (decision number 5.8.18-14430/2023).

References

10. Ross ED, Gupta SS, Adnan AM, Holden TL, Havlicek J, Radhakrishnan S. Neurophysiology of spontaneous facial expressions: I. Motor control of the upper and lower face is behaviorally independent in adults. Cortex. 2016;76: 28–42.
11. Karmann AJ, Maihöfner C, Lautenbacher S, Sperling W, Kornhuber J, Kunz M. The Role of Prefrontal Inhibition in Regulating Facial Expressions of Pain: A Repetitive Transcranial Magnetic Stimulation Study. J Pain Off J Am Pain Soc. 2016;17: 383–391.
12. Kavanagh E, Kimock C, Whitehouse J, Micheletta J, Waller BM. Revisiting Darwin’s comparisons between human and non-human primate facial signals. Evol Hum Sci. 2022;4: e27.
13. Straulino E, Scarpazza C, Sartori L. What is missing in the study of emotion expression? Front Psychol. 2023;14.
14. Rashid M, Silventoinen A, Gleerup KB, Andersen PH. Equine Facial Action Coding System for determination of pain-related facial responses in videos of horses. PLOS ONE. 2020;15: e0231608.
15. Descovich KA, Wathan J, Leach MC, Buchanan-Smith HM, Flecknell P, Farningham D, et al. Facial expression: An under-utilized tool for the assessment of welfare in mammals. Altex. 2017;34: 409–429.
16. Waller BM, Julle-Daniere E, Micheletta J. Measuring the evolution of facial ‘expression’ using multi-species FACS. Neurosci Biobehav Rev. 2020;113: 1–11.
17. McLennan KM, Miller AL, Dalla Costa E, Stucke D, Corke MJ, Broom DM, et al. Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. Appl Anim Behav Sci. 2019;217: 1–15.
18. Torcivia C, Mcdonnell S. In-Person Caretaker Visits Disrupt Ongoing Discomfort Behavior in Hospitalized Equine Orthopedic Surgical Patients. Animals. 2020;10.

19. van Loon JPAM, Van Dierendonck MC. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A scale-construction study. Vet J. 2015.
20. Dalla Costa E, Minero M, Lebelt D, Stucke D, Canali E, Leach MC. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS ONE. 2014;9: 1–10.
21. Gleerup KB, Forkman B, Lindegaard C, Andersen PH. An equine pain face. Vet Anaesth Analg. 2015;42: 103–114.
22. Wathan J, Burrows AM, Waller BM, McComb K. EquiFACS: The equine facial action coding system. PLoS ONE. 2015;10: 1–35.
23. Basmajian JV. Muscle Alive: Their Functions Revealed by Electromyography. Baltimore: The Williams & Wilkins Company; 1962.
24. Tassinary LG, Cacioppo JT. Unobservable Facial Actions and Emotion. Psychol Sci. 1992;3: 28–33.
25. Mandal MK, Ambady N. Laterality of facial expressions of emotion: Universal and culture-specific influences. Behav Neurol. 2004;15: 23–34.
26. Schwartz GE, Fair PL, Salt P, Mandel MR, Klerman GL. Facial Muscle Patterning to Affective Imagery in Depressed and Nondepressed Subjects. Science. 1976;192: 489–491.
27. Cacioppo JT, Petty RE. Attitudes and cognitive response: An electrophysiological approach. J Pers Soc Psychol. 1979;37: 2181–2199.
28. Naik GR. Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges. 2012.
29. Kainer RA. Clinical Anatomy of the Equine Head. Vet Clin North Am Equine Pract. 1993;9: 1–23.
30. Mesin L. Crosstalk in surface electromyogram: literature review and some insights. Phys Eng Sci Med. 2020;43: 481–492.
31. Gans BM, Kraft GH. Pain perception in clinical electromyography. Arch Phys Med Rehabil. 1977;58: 13–16.

Setting up an observation strategy to record feeding synchronization in dairy cows

M. Battini, S. Celozzi and S. Mattiello

Department of Agricultural and Environmental Sciences, University of Milan, Milan, Italy.

monica.battini@unimi.it

Introduction

In recent years, research on animal welfare has been moving from the use of negative indicators, such as lameness, abnormal behaviour, or fear, to the use of indicators of positive welfare, which refer to the presence of positive affects, which may improve the animals' quality of life [1]. Unfortunately, most of the current protocols for on-farm welfare assessment still focus mainly on negative aspects, and therefore interpret the lack of suffering as a condition of welfare [2]. Thus, there is a need to identify animal-based indicators of positive welfare, and to set up a feasible strategy for data collection.

Social facilitation and imitation are probably the main triggers that increase the probability that a behaviour is performed by other nearby animals of the same group (i.e., allelomimetic behaviour). Therefore, synchronization is indicative of social cohesion [3] and it is common in social species, like cattle [4]. It may yield several benefits, mainly consisting in better defence against predators and an increased opportunity to spend time grazing, thanks to the possibility of sharing the time for vigilance among all the individuals [5]. There is no clear agreement about the optimal level of synchronization that should be achieved in a group, and suggested values range from 70% to 90% [6–8]. However, it has to be reminded that behavioural synchrony may vary throughout the day, and this is particularly important when planning the observations. For example, [8] suggest recording synchronicity at feeding within one or few hours after the morning feed delivery, as the motivation to eat is higher.

Some authors suggest that feeding synchronization can be assessed by instantaneous scan sampling; however, to our knowledge, no well–established data collection rule is available for its evaluation for on-farm welfare assessment purposes [8,9]. Scan sampling represents a compromise between obtaining the most accurate estimate possible (i.e., continuous sampling) and improving the efficiency of data collection, by reducing time and effort [10,11].

In this study, we tested the application of the instantaneous scan sampling method for collecting information on feeding synchronization, simulating different duration of observations and different scan intervals in order to identify a feasible observation strategy that allows the inclusion of feeding synchronization in on-farm welfare assessment protocols for dairy cows.

Material and methods

The research took place on four dairy cattle farms (A, B, C and D) situated in Northern Italy. The study involved observations across seven distinct contexts, each serving as a case study. These contexts were characterized by variations in housing systems (loose housing, LH; tie stalls, TS), types of feed (total mixed ration, TMR; hay, HAY; fresh grass, GRA), and feeding place to cow ratio (sufficient, ≥1; insufficient, <1). Additionally, an infirmary pen (INF) was included in the sample to provide an example of a situation where feeding synchronization might be negatively impacted by health conditions. (Table 1). In Farm D, the same animals were observed twice, because they received two different types of feed: ventilated hay (at 8:30 a.m.) and fresh grass (at 11:00 a.m.). Table 1 shows further details of the wide range of the farm contexts' characteristics in our sample. This high variability among contexts, which makes our sample highly heterogeneous, interestingly represents the possible range of common situations which may be experienced across dairy cattle farms.

One trained assessor collected data through direct observations at feeding. Given that feed delivery can be considered as the main motivating factor to elicit feeding behaviour [12], observations started immediately after the morning feed delivery, which occurred around 8:30 a.m. (except for fresh grass distribution in Farm D, as specified above). One observation session was carried out for each context, using an instantaneous scan sampling method [13]. Each session lasted one hour, and scans were performed at 5-min intervals, for a total of 13 scans for each session. At each scan, the assessor counted the number of animals that were at the feeding rack. For each context, three variables were calculated: 1) maximum percentage of animals feeding at any one-time point during the entire observation session (MAX); 2) percentage of scans with ≥80% of animals feeding at the same time (≥80%); 3) mean percentage of animals feeding at the same time (MEAN).

In order to identify a feasible and reliable assessment strategy in terms of sampling interval and duration of the observation session, we subsequently generated a set of simulated databases generated from the initial one (Control database), and for each database we calculated MAX, ≥80% and MEAN. First, we calculated these variables when the scan interval was either 10- or 15- rather than 5-min. Then we calculated the same three variables simulating a shorter observation period: from the Control database, we selected scans from 1 to 7 to have a 30-min session, from which we generated three databases with different scan intervals (5-, 10-, and 15-min). We compared MAX, ≥80% and MEAN obtained from the simulated databases with those from the Control database using Wilcoxon matched-pairs test (IBM® SPSS® Statistics, vers. 26).

Results

The level of synchronization widely differed among contexts (Figure 1), with infirmary pen showing the lowest (animals in poor health condition) and tie stall showing the highest degree of synchronization (animals with individual access granted to the feeding rack). The peak of synchronization is always achieved withing the first 30 minutes of the observation session.

Using data from the 60-min observation session, the results showed no significant differences for MAX and ≥80% depending on the scan interval, whereas significant differences were found for the mean percentage of animals feeding at the same time (MEAN) between 5-min and 15-min interval ($p<0.05$) (Table 2).

No significant differences were observed for the considered variables depending on the duration of the observation session (60 minutes *vs* 30 minutes), neither using a 5-min scan interval, nor a 10-min or 15-min scan interval and using 10-min scan intervals, the maximum percentage of animals feeding at the same time (MAX) is absolutely identical to that obtained observing the animals for 30 or for 60 minutes (Table 3).

Discussion

The three selected variables were highly consistent within contexts, regardless of the scan interval or duration of the observations, but varied widely across contexts. The use of instantaneous scan sampling method was feasible, but a 60-min observation session with scan sampling every 5 minutes is time consuming and may be exhausting for the assessor, especially in large farms where high numbers of animals have to be counted in a short period of time. Our simulations using different scan intervals and duration of the observation session suggest that this approach can be refined to improve the feasibility. Using a 10-min scan sampling strategy yields information comparable to that obtained with a 5-min scan sampling strategy. This is in agreement with [10] who found that a 10-min scan interval accurately captures lamb grazing behaviour. In our study, employing a 10-min scan sampling approach offers assessors the flexibility of longer intervals between scans, enabling the collection of additional data between scans, or the possibility to restore before the next scan. This has the potential to save time when implementing a comprehensive welfare assessment protocol, where time is one of the major constraints, and to improve the precision of the assessment, avoiding excessive tiredness of the assessor.

Reducing the observation session to 30 minutes could enhance feasibility, even when using a 10-min scan interval, thus making the assessment of this potential indicator less time-consuming. Therefore, it is not surprising that the maximum percentage of animals feeding simultaneously remains nearly identical whether observed for 30 or for 60 minutes. However, if the goal is to capture the evolution of behaviour over time, it is suggested to extend the observation to 60 minutes.

Conclusions

Despite relying on a limited sample size, our prelilminary findings confirm that observations on feeding synchronization can effectively highlight exptected differences among contexts. Based on our results, the maximum percentage of animals feeding at the same time emerges as the most informative and reliable variable for assessing feeding synchrony. To improve the feasibility of data collection, we suggest conducting the observations on feeding synchronization following the morning feed distribution, using an instantaneous scan sampling method with 10-min scan intervals and an overall duration of the observation session of 30 minutes.

Ethical statement

The study was approved by the Animal Welfare Organisation (OPBA) of the University of Milan (permit number: OPBA_58_2022), in compliance with the Directive 2010/63/EU.

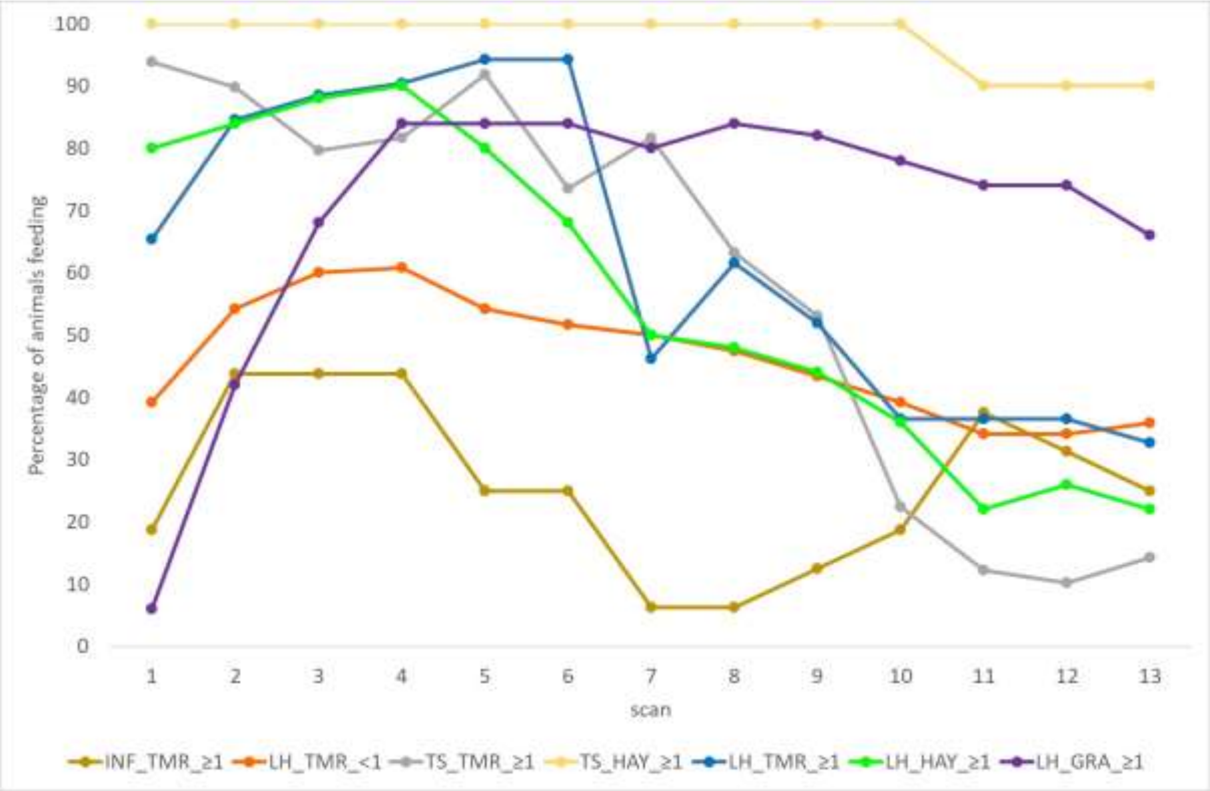


Figure 1. Trend of the percentage of animals feeding at the same time during 60 min after feed delivery (5-min interval scans) in the seven contexts. INF=infirmary pen; LH=loose housing; TS=tie stall; TMR=total mixed ration; HAY=hay; GRA=fresh grass; ≥1=feeding place:cow ratio ≥1; <1=feeding place:cow ratio <1.

Table 1. Characteristics of the seven contexts in which observations on feeding synchronization were carried out.

Farm	Context[1]	N. of cows	Feeding place: cow ratio	Type of feed	Husbandry system	Notes
A	INF_TMR_≥1	16	1.5	TMR	Loose house, cubicles	Infirmary pen
A	LH_TMR_<1	120	0.7	TMR	Loose house, cubicles	Lactating cows
B	TS_TMR_≥1	49	1.0	TMR	Tie stall	Lactating cows
B	TS_HAY_≥1	10	1.0	Hay	Tie stall	Dry cows
C	LH_TMR_≥1	52	1.0	TMR	Loose house, cubicles	Lactating cows
D	LH_HAY_≥1	50	1.7	Ventilated hay	Loose house, cubicles	Lactating cows
D	LH_GRA_≥1	50	1.7	Fresh grass	Loose house, cubicles	Lactating cows

[1] INF=infirmary pen; LH=loose housing; TS=tie stall; TMR=total mixed ration; HAY=hay; GRA=fresh grass; ≥1=feeding place:cow ratio ≥1; <1=feeding place:cow ratio <1

Table 2. Comparison of maximum percentage of animals feeding at the same time (MAX), percentage of scans with ≥80% of animals feeding at the same time (≥80%) and mean percentage of animals feeding at the same time (MEAN) during a 60-min observation period using three different scan intervals (5-min, 10-min and 15-min).

	MAX (% of cows)			≥ 80% (% of scans)			MEAN (% of cows)		
Context[1]	5-min	10-min	15-min	5-min	10-min	15-min	5-min	10-min	15-min
INF_TMR_≥1	44	44	44	0	0	0	26	24	23
LH_TMR_<1	61	60	61	0	0	0	46	45	45
TS_TMR_≥1	94	94	94	38	57	60	59	61	59
TS_HAY_≥1	100	100	100	100	100	100	98	97	98
LH_TMR_≥1	94	94	90	38	29	20	63	59	54
LH_HAY_≥1	90	88	90	23	43	40	57	55	56
LH_GRA_≥1	84	84	84	38	43	40	70	66	63

[1] INF=infirmary pen; LH=loose housing; TS=tie stall; TMR=total mixed ration; HAY=hay; GRA=fresh grass; ≥1=feeding place:cow ratio ≥1; <1=feeding place:cow ratio <1

Table 3. Comparison of maximum percentage of animals feeding at the same time (MAX), percentage of scans with ≥80% of animals feeding at the same time (≥80%) and mean percentage of animals feeding at the same time (MEAN) using two different durations of the observation sessions (60 minutes vs 30 minutes) at 5-min and 10-min intervals.

	5-min scan interval						10-min scan interval					
	MAX		≥ 80%		MEAN		MAX		≥ 80%		MEAN	
	(% of cows)		(% of scans)		(% of cows)		(% of cows)		(% of scans)		(% of cows)	
Context[1]	60 min	30 min	60 min	30 min	60 min	30 min	60 min	30 min	60 min	30 min	60 min	30 min
INF_TMR_≥1	44	44	0	0	26	29	44	44	0	0	24	23
LH_TMR_<1	61	61	0	0	46	53	60	60	0	0	45	51
TS_TMR_≥1	94	94	38	71	59	85	94	94	57	100	61	87
TS_HAY_≥1	100	100	100	100	98	100	100	100	100	100	97	100
LH_TMR_≥1	94	94	38	71	63	80	94	94	29	50	59	74
LH_HAY_≥1	90	90	23	71	57	77	88	88	43	75	55	75
LH_GRA_≥1	84	84	38	57	70	64	84	84	43	50	66	60

[1] INF=infirmary pen; LH=loose housing; TS=tie stall; TMR=total mixed ration; HAY=hay; GRA=fresh grass; ≥1=feeding place:cow ratio ≥1; <1=feeding place:cow ratio <1

Reference

1. Farm Animal Welfare Council. Farm Animal Welfare in Great Britain: Past, Present and Future.; 2009.
2. Welfare Quality Consortium. Welfare Quality ® Assessment Protocol for Cattle.; 2009.
3. Gautrais J, Michelena P, Sibbald A, Bon R, Deneubourg JL. Allelomimetic synchronization in Merino sheep. Anim Behav. 2007;74(5):1443-1454. doi:10.1016/j.anbehav.2007.02.020
4. Bouissou MF, Boissy A, le Neindre P, Veissier I. The social behaviour of cattle. In: Keeling L, Gonyou H, eds. Social Behaviour in Farm Animals. CAB International; 2001:113-145.
5. Dávid-Barrett T, Dunbar RIM. Cooperation, behavioural synchrony and status in social networks. J Theor Biol. 2012;308:88-95. doi:10.1016/j.jtbi.2012.05.007
6. Stoye S, Porter MA, Stamp Dawkins M. Synchronized lying in cattle in relation to time of day. Livest Sci. 2012;149(1-2):70-73. doi:10.1016/j.livsci.2012.06.028
7. Arnold GW, Dudzinski ML. Social organization and animal dispersion. In: Arnold GW, Dudzinski ML, eds. Ethology of Free-Ranging Domestic Animals. Elsevier Scientific Publishing Company; 1978:51-96.
8. Napolitano F, Knierim U, Grass F, De Rosa G. Positive indicators of cattle welfare and their applicability to on-farm protocols. Ital J Anim Sci. 2009;8(sup1):355-365. doi:10.4081/ijas.2009.s1.355
9. Mattiello S, Battini M, De Rosa G, Napolitano F, Dwyer C. How can we assess positive welfare in ruminants? Animals. 2019;9(10):1-27. doi:10.3390/ani9100758
10. Pullin AN, Pairis-Garcia MD, Campler MR, Proudfoot KL. Validation of scan sampling techniques for behavioural observations of pastured lambs. Animal Welfare. 2017;26(2):185-190. doi:10.7120/09627286.26.2.185
11. Pullin AN, Pairis-Garcia MD, Campbell BJ, Campler MR, Proudfoot KL. Technical note: Instantaneous sampling intervals validated from continuous video observation for behavioral recording of feedlot lambs. J Anim Sci. 2017;95(11):4703-4707. doi:10.2527/jas2017.1835
12. DeVries TJ, von Keyserlingk MAG. Time of feed delivery affects the feeding and lying patterns of dairy cows. J Dairy Sci. 2005;88(2):625-631. doi:10.3168/jds.S0022-0302(05)72726-0
13. Bateson M, Martin P. Measuring Behaviour - An Introductory Guide. 4th Edition. Cambridge University Press; 202

Rhythmicity as a welfare indicator – looking into the effect of extrinsic and intrinsic motivation in group housed horses

S. Rey[1], H.R Nasser[2] and M. Cockburn[1]

[1] Animals, Products of Animal Origin and Swiss National Stud, Agroscope, Haras national suisse HNS, Avenches, Switzerland. sonia.rey@agroscope.admin.ch [2] Digital production group, Agroscope, Posieux, Switzerland.

Introduction

The synchronization of an organism to its environment gives indications on its wellbeing, where a high level of synchronicity reflects a good state of physiological and psychological welfare, whereas a desynchronization with its environment reflects the disturbance of the animal's state. It has therefore been proposed that the rhythmicity of animals can serve as a welfare indicator in livestock, as they have shown to follow animal specific rhythmic patterns [4] and [6]. This rhythmicity of animals can be measured by motion sensors and calculated by a Fourier Transformation, resulting in the degree of functional coupling (DFC) [6]. The Fourier transformation calculates the rhythmicity of the animal's locomotion over the past 7 days, which results in the DFC as an output. The DFC is presented on a level from 1 to 0, where 0 presents no rhythmicity and a poor state of welfare and 1 presents a good state of welfare with high rhythmicity. The DFC has shown promising results in extensively managed sheep [5] and in housed dairy cows [2], and has been used to asses activity and feeding behaviour in wild horses [1]. In theory, the DFC can be computed from any sensor data that reflects and organisms' locomotion patterns, such as accelerometers, feeding stations or automatic milking systems, thus it would be a feasible opportunity to gain data on the animal's welfare. However, to date it is unclear how well the DFC reflects the animal's wellbeing in strictly managed husbandry systems, where rhythmicity could be affected by extrinsic factors such as feeding regimes. It therefore cannot be distinguished if the rhythmic pattern detected by the sensors is intrinsically or extrinsically motivated. In the current project we aim to evaluate this circumstance in a two- phased trial in domesticated horses.

Methods

Ethical statement: The experiment was approved by the Freiburg Cantonal authority under the license 2023-40-FR. All relevant guidelines for animal handling were respected.

Two controlled trials were carried out in four paddock trail group housing systems (see Figure 2). Each group housing system contained five female warmblood *Equus Caballus*. Each paddock trail offered a shelter with wood shaving littering material, free access to water and a feeding station at the other end of the trail (see Figure 2). Data was collected at 1 Hz by wireless MSR loggers (X, Y and Z axis) attached to the horses' leg for a period of 14 days per treatment (see Figure 1).

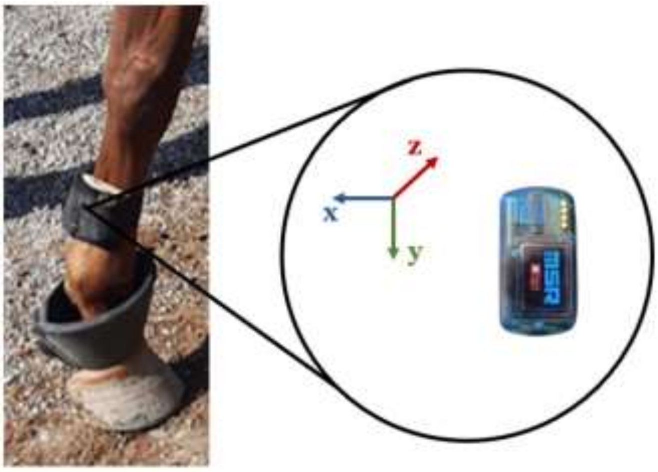


Figure 1. Accelerometer placed on the horse's leg, capturing the orientation of each acceleration axis.

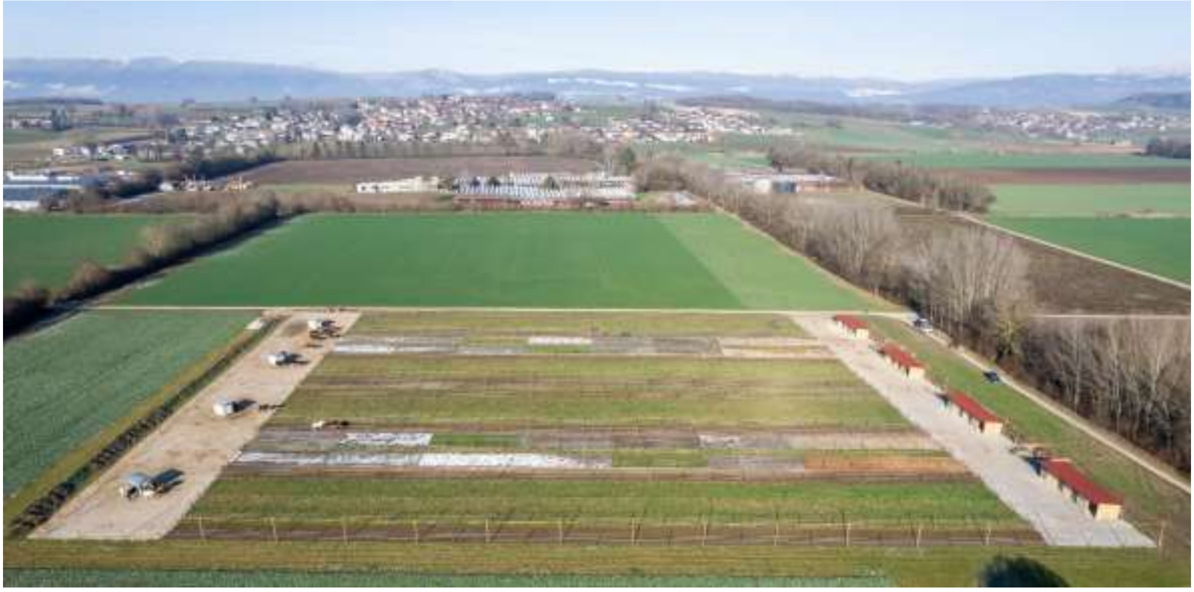


Figure 2. Overview of paddock trail setup during trial 1. The shelters are located on the right, the feeding station are on the left. The four groups are separated by an electronic fence. The grass is fenced off and trails connect functional areas for resting and feeding.

Trial 1: Rhythmicity as a consequence of extrinsic factors:

In the paddock trail system, the feeding regime of the automatic feeding stations was programmed to three different feeding regimes (see Table 1). No additional roughage was offered during the course of this trial. Each treatment was carried out for a period of four weeks (two weeks habituation followed by two weeks of data collection):

A Loose hay three times per day for two hours
B Loose hay six times per day for one hour
C Ad libitum Hay with hay net 24 hours per day

Table 1. Experimental design of scheduled feeding.

	Treatment 1	**Treatment 2**	**Treatment 3**
Group 1	B	A	C
Group 2	B	C	A
Group 3	A	B	C
Group 4	C	A	B

We hypothesized that rhythmicity in horses with set feeding regimes is synchronized to the feeding regimes.

Trial 2: Rhythmicity as a consequence of intrinsic factors:

The same equines were kept on the same paddock trail system displayed above, however, the intrinsic motivation for lying behaviour was adjusted by an alteration of the shelter design, as well as a change of their orientation. Automatic feeding was set to provide hay three times per day for a duration of two hours each, additionally hay was offered in hay nets and hay bells at the shelters for an extended feeding period. Additionally, straw was offered ad libitum in the shelters. We tested the effect of shelter layout on intrinsically motivated rhythmicity and lying behaviour by implementing four treatments for a duration of four weeks each (see Figure 3 and Table 2).

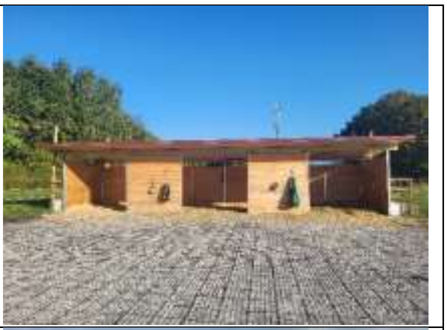
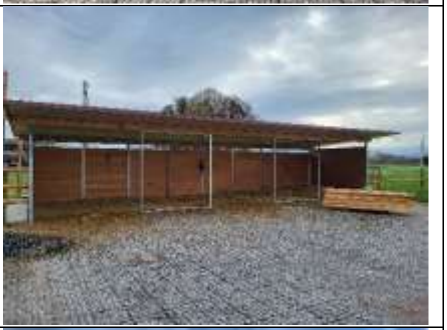
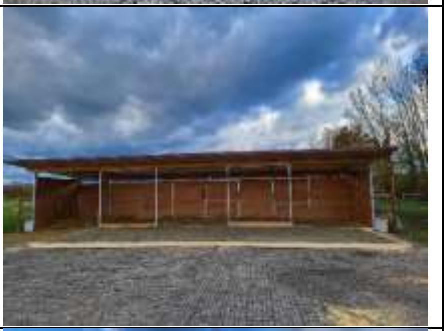
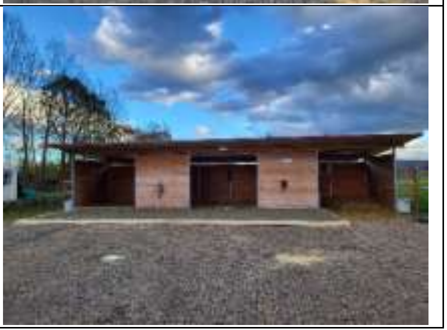
Initial layout: Shelters were closed with two exits.	
No wood: The wood of one long side of the shelters was removed.	
No wood + Sand: The wood of one long side of the shelters was removed and sand was supplied to extend the lying area.	
Sand: Shelters were closed with two exits, but sand was supplied to extend the lying area.	

Figure 3. Stable design layout.

Table 2 Experimental design of altered stabled design.

Weeks Groups	3-4	7-8	11-12	15-16
1	No wood	Initial layout	Sand	No wood + Sand
2	No wood + sand	Sand	Initial layout	No wood
3	Sand	No wood + sand	No wood	Initial layout
4	Initial layout	No wood	No wood + Sand	Sand

We hypothesized that the rhythmicity of horses will change according to their intrinsic motivation to rest in different shelter layouts.

Data analysis

For both trials, we analyzed the Degree of Functional Coupling (DFC) from the sum of the square root of the x, y and z axis (√(x^2 + y^2 +z^2)) of the accelerometer data using the R package "digiRhythm" (README (r-project.org)) [3].Seven DFC values were computed for each horse and treatment. Additionally, we computed the lying behaviour from the y-axis during this time period using the "triact" [7]. All special occurrences were documented in the animal health protocol. Cameras were installed inside the shelters to observe horse behaviour; however, these videos were not systematically analyzed.

Statistics

The data will be analyzed with a general linear mixed effects model.

Results

At the time of writing this paper, no results are present as data analysis is still ongoing. However, we will present the results of both trials during the conference.

Discussion

This research will provide further evidence on the validity of using rhythmicity as a welfare indicator in managed livestock. However, we would like to also point out shortcomings in the study design. Despite their separation and different treatments, all groups of equines had visual, oral and olfactory contact over the fence. Thus, causing a synchronization between all groups. Further, the horses were group housed in the same groups for over six months prior to onset of the trial. They were well adjusted to their herds. However, after moving them to the paddock trail system, the lying behaviour of all horses reduced notably. Thus, we assume that the welfare of the horses in this trial was comprised. Yet this could in fact give us further evidence, if rhythmicity can serve as a welfare indicator in managed husbandry systems. If the DFC is high in systems with synchronized feeding regimes, despite a poor state of welfare, it cannot be a good welfare indicator in managed systems. However, it low despite the synchronized feeding schedules it could well be a good indicator. The link between lying behaviour and DFC will provide further evidence on the value of rhythmicity as a welfare indicator in managed livestock.

References

1. Berger, A., K.-M. Scheibe, K. Eichhorn, A. Scheibe, and J. Streich. (1999). Diurnal and ultradian rhythms of behaviour in a mare group of Przewalski horse (Equus ferus przewalskii), measured through one year under semi-reserve conditions. *Applied Animal Behaviour Science* **64**(1):1-17.

2. Fuchs, P., F. Adrion, A. Z. Shafiullah, R. M. Bruckmaier, and C. Umstätter. (2022). Detecting ultra-and circadian activity rhythms of dairy cows in automatic milking systems using the degree of functional coupling—A pilot study. *Front. Anim. Sci* **3**.

3. Nasser, H.-R., Schneider, M., and Cockburn, M. (2023). DigiRythm: an R package for rhythmicity assessment using the degree of functional coupling. Manuscript submitted for publication.

4. Rufener, C., J. Berezowski, F. Maximiano Sousa, Y. Abreu, L. Asher, and M. J. Toscano. (2018). Finding hens in a haystack: Consistency of movement patterns within and across individual laying hens maintained in large groups. *Scientific Reports* **8**(1):12303.

5. Sarout, B. N. M., A. Waterhouse, C.-A. Duthie, C. H. E. C. Poli, M. J. Haskell, A. Berger, and C. Umstatter. (2018). Assessment of circadian rhythm of activity combined with random regression model as a novel approach to monitoring sheep in an extensive system. *Applied Animal Behaviour Science* **207**:26-38.

6. Scheibe, K. M., A. Berger, J. Langbein, W. J. Streich, and K. Eichhorn. (1999). Comparative Analysis of Ultradian and Circadian Behavioural Rhythms for Diagnosis of Biorhythmic State of Animals. *Biological Rhythm Research* **30**(2):216-233.

7. Simmler, M. and Brouwers, S. (2023). Triact Package for R: Analyzing the Lying Behavior of Cows from Accelerometer Data. Available at SSRN: https://ssrn.com/abstract=4560833 or http://dx.doi.org/10.2139/ssrn.4560833.

Bridging the Gap - Integrating Wild Animal Welfare into Behavior Studies

Janire Castellano Bueno [1,2], Vittoria Elliott [1,3]

1 Wild Animal Initiative, 2 Newcastle University, 3 Smithsonian Institution

Background

Recently, spurred on by an enhanced comprehension of mental health in humans and a growing body of research unveiling the association between displayed behaviour and the affective state of animals, the study of behaviour has reclaimed a central role in animal welfare assessment. Yet, this renaissance in understanding behaviour has not been met with an equivalent integration of welfare science into animal behaviour research. Contrary to evidence from the literature and recommendations from frameworks like STRANGE [1], the incorporation of the welfare perspective into animal behaviour studies is still not common practice, despite the value and potential that it offers for enhancing both disciplines..

This is especially true when it comes to wild animal species. Neglecting the welfare state of study animals hampers our understanding of individual differences and exacerbates the replicability crisis currently confronting the animal behaviour field, as well as most other scientific fields [2].

Measuring behaviour with welfare science

This paper dives into the intersection of the welfare of wild animals and their behaviour. Methods for measuring behaviour are discussed, offering insights into how they can be instrumental for better understanding welfare. Technological advancements have facilitated the development of a broader suite of non-invasive tools for behaviour encoding, offering a streamlined approach to behavioural observation. These methods include: (1) non-invasive behavioural and body condition tracking, (2) deep learning algorithms and (3) advanced population and behavioural simulations. Here we will focus on the integration of these tools with conventional observational methods to achieve a comprehensive understanding of animal behaviour and welfare. We will also provide an introduction to some of the complementary tools available from the physiological sciences that can support interpretation.

Emphasizing the importance of considering both behaviour and welfare within the scientific method, this poster provides examples from diverse species in controlled and wild settings. The importance of integrating animal welfare into animal behaviour studies in the wild is further underscored, with case study examples elucidating opportunities and addressing common misconceptions surrounding methods for measuring behaviour and welfare. Challenges in applying findings and methods from captive settings to wild animals are explored, and through a series of case studies, methodological adaptations necessary for understanding welfare and behaviour in the wild are discussed.

The scientific rationale behind advocating for this integrative approach will be underscored, elucidating how the incorporation of welfare considerations can enhance the overall quality of research. This encompasses improvements in replicability, a reduction in bias, and, notably, potential cost savings.

Using a series of case studies we will describe various behavioural methods and how to use them to assess and better understand welfare. The studies will also emphasize the value of using multiple independent indicators and pairing behavioural measures with physiological ones to obtain a more holistic understanding of the individual welfare condition. Through the selection of examples we will emphasize the opportunity and importance of using non-invasive minimally disturbing methods for assessing animals in the wild.

The behavioural and physiological measures we will explore include drone-based image collection for assessment of body condition and health; remote camera recording to monitor and assess behaviours; cognitive bias testing using field-based experimental design; thermal imaging of behavioural and physiological stress responses; faeces, hair, feather, and saliva collection for assessment of stress and other hormones, indicators of immune system response, and telomere attrition .

We will discuss these methods using a set of on-going studies being conducted to develop, assess, and implement techniques for measurement of animal welfare in the wild. The following will be included among the studies we explore:

- Quantifying the impact of sea ice coverage on the welfare of gray seal pups. (Daire Carroll)
- Does population density influence the welfare of wild newts? (Luiza Figueiredo-Passos)
- Social connections and their welfare implications in wild jackdaws (Alex Thornotn)
- Impacts of land-use on social networks in mixed-species bird flocks, with implications for the short-term and long-term welfare of Himalayan birds. (Josh Firth)

Conclusion

This holistic approach not only contributes to resolving crucial questions in animal behaviour research but also advocates for a paradigm shift in acknowledging the symbiotic relationship between welfare and behaviour in both captive and wild settings. By embracing this integrated perspective in both captive and wild settings, researchers can not only address existing challenges but also pave the way for a more scientifically robust future in the study of animal behaviour and welfare.

References

1. Webster, M., Rutz, C., (2020). How STRANGE are your study animals?, *Nature* **582**: 337-340

2. Baker, M., (2016). 1,500 scientists lift the lid on reproducibility. *Nature* **May**: 26;533(7604):452-4.

Technology for Measuring Behaviour

Monitoring Work Stress: The Role of Physiological Measures in Enhancing Workplace Well-being

P. Shams[1] and H. Blidh[2]

1 Karlstad Business School, Karlstad University, Karlstad, Sweden. poja.shams@kau.se

2 Research and Development, Swedwise, Karlstad, Sweden. henrik.blidh@swedwise.se

Introduction

In the fast-paced environment of modern workplaces, understanding and managing work stress is crucial for both employees and organizations [1]. Just as marketing researchers use tools like Electrodermal Activity (EDA) to explore consumer emotions [2], professionals can employ similar techniques to monitor and understand stress responses in the workplace. Recognizing physiological indicators of stress, such as changes in heart rate or sweat secretion, can provide real-time insights into an employee's emotional state during the workday. This understanding is essential, as work stress affects not only cognitive functions [3] but also emotional and physical well-being [4]. By acknowledging the role of physiological responses in stress and employing methods to measure these responses, organizations can develop more effective strategies to enhance employee well-being and productivity. This approach aligns with the broader need to integrate diverse data sources, including physiological measures, into our understanding of work stress, thereby fostering a healthier and more productive work environment.

Electrodermal activity (EDA)

Electrodermal Activity (EDA) refers to the skin's electrodermal phenomena with psychological implications. Emotional arousal stimulates eccrine sweat glands, leading to increased sweat production, which effectively conducts electricity. This process alters the skin's electrical properties; the greater the arousal, the more pronounced the changes in these properties. EDA measurement involves assessing the skin's electrical conductance, resistance, impedance, or admittance, based on the chosen recording technique, and is quantified in microsiemens (µS) [5].

EDA is categorized into two types of activity: tonic and phasic [6]. Tonic activity, or the Skin Conductance Level (SCL), represents the baseline electrodermal activity and changes gradually. In contrast, phasic activity is a response to specific, discrete stimuli, particularly those that are emotionally arousing. When encountering a stimulus perceived as personally significant, the brain, through the sympathetic nervous system, activates the eccrine sweat glands. This activation results in increased sweat secretion, leading to a rapid spike in skin conductance. This transient increase in skin conductance is known as the Skin Conductance Response (SCR), also referred to as the electrodermal activity response (EDR) or peak.

EDA in measuring experience

Traditional methods of assessing workplace stress have primarily relied on self-reported questionnaires and interviews. These tools gather subjective data on employees' perceptions of stress, their coping mechanisms, and the impact on their mental and physical health. While valuable for providing insights into the personal experiences of workers, these methods are susceptible to biases like social desirability and recall inaccuracies. Responses can be influenced by an individual's desire to present themselves in a certain way or by their inability to accurately remember past emotional states. In contrast, Electrodermal Activity (EDA) emerges as a superior alternative for measuring stress in the workplace. Unlike traditional methods, EDA offers objective, real-time data on physiological responses to stress. This technique bypasses the subjectivity and potential biases of self-reports. By measuring the skin's electrical conductance, which varies with emotional arousal, EDA provides a direct, physiological indicator of stress levels. This allows for a more accurate and reliable assessment of the stress environment in the workplace, aiding to the understanding of the stressor in the workplace. As such, EDA represents a significant advancement in the field of work life science, offering a more nuanced and precise understanding of employee stress.

EDA as work life experience measurement tool

The diary method has long been a staple in field studies for collecting data on workplace factors [7,8]. In this approach, participants are asked to regularly record their activities throughout the day. This self-reporting provides valuable insights into the contexts and activities that trigger stress, allowing for a subjective but detailed account of an individual's daily experiences. However, the diary method relies heavily on the participant's ability to accurately recall and report their feelings, which can be influenced by various biases.

Integrating EDA with the diary method significantly enhances the reliability and depth of data collection. EDA offers real-time, objective measurement of physiological responses indicative of stress. A tool that processes EDA signals can continuously monitor stress levels, providing immediate feedback to the user when elevated stress is detected. This immediate indication of stress complements the diary entries, as it prompts the user to record specific activities or events coinciding with the physiological signs of stress.

By combining EDA with the diary method, researchers and participants gain a more comprehensive understanding of stress triggers. The EDA tool's real-time feedback ensures that high-stress moments are not overlooked or forgotten, addressing a key limitation of the traditional diary method. Additionally, the concurrent use of EDA data and diary entries allows for a deeper analysis of the relationship between specific activities and stress responses. This dual approach not only enhances the accuracy of stress measurement but also provides richer data for developing targeted interventions and stress management strategies in the workplace.

The SWEA (Stress-work-emotion-algoritm) Toolbox

The SWEA Toolbox is an innovative solution designed for the comprehensive monitoring and analysis of workplace stress. It represents a significant advancement in stress management technologies, combining real-time physiological data collection with sophisticated analytical capabilities.

Front-End User Interface:
At the heart of the SWEA Toolbox is its user-friendly front-end interface. This component is tailored for individual users to interact with the system in real-time. When the EDA sensor detects elevated stress levels, the user receives a prompt to annotate the event. This annotation process involves commenting on the current activity or stressor, providing contextual details that are invaluable for later analysis. This feature not only aids in identifying immediate stress triggers but also encourages self-awareness among users regarding their stress responses.

Back-End Researcher Interface:
The back-end of the SWEA Toolbox is where the complex data handling and analysis occur. This interface is designed for researchers and organizational leaders who are analyzing stress patterns within the workplace. The system collects and stores all EDA recordings, along with the corresponding user comments.

Integrated into this interface are advanced machine learning models, including Support Vector Machines (SVMs), which analyze the EDA signals. The SVM model is trained to recognize stress patterns. Using the method described in pyEDA [9], a convolutional autoencoder was first trained to extract features from the EDA signal. The autoencoder takes snippets of the EDA signal as input and compresses them into a lower-dimensional representation of the signal in that time window, which is then used as input to the SVM model to train detection of stress patterns. The accuracy of the SVM model is then tested using a cross-validation method to ensure reliability and validity in real-world scenarios. This involves dividing the data into training and test sets, where the model is trained on one set and validated on the other to test its ability to generalize to new, unseen data. The SVM model uses these features to detect subtle fluctuations in stress levels, allowing for a nuanced analysis of stress patterns over time. Markers shown in Figure 1 correspond to identified stress events, where the model detects significant deviations from baseline stress levels.

In addition to signal analysis, the SWEA Toolbox employs Natural Language Processing (NLP) models to analyze user comments. These models, working in tandem with our SVM-based analysis, parse the text for keywords and sentiments that are indicative of stress, linking these linguistic cues to physiological data. This holistic approach allows us to correlate self-reported experiences with empirical data, providing a comprehensive view of an individual's stress landscape.

Comprehensive Stress Analysis:
The SWEA Toolbox's integration of real-time monitoring, user interaction, and advanced data analytics offers a comprehensive approach to understanding workplace stress. It enables the identification of common stressors, patterns in stress responses, and potential areas for intervention. This tool is invaluable for organizations aiming to foster a healthier work environment and for researchers seeking deeper insights into the dynamics of work-related stress.

Pilot Study using SWEA Toolbox

A pilot study was conducted using the SWEA Toolbox to gain insights into the stress events experienced by healthcare workers in their daily professional lives. This study involved six participants, all healthcare professionals, who used the tool for a duration of one week. During this period, they generated over 100 stress events, providing a rich dataset for analysis.

The hardware chosen for measuring the EDA signals in this study was the Empatica E4 wristband. This device is renowned for its ease of use in real-time physiological monitoring. It's particularly suitable for field studies like this one, as it allows participants to go about their normal workday without intrusive equipment. The Empatica E4 continuously recorded the EDA signals, capturing the physiological markers of stress as they occurred.

Throughout the week, participants interacted with the SWEA Toolbox's user interface to log contextual information about each stress event as prompted by the elevated EDA signals detected by the Empatica E4. This process involved the participants briefly noting the activity or situation they were involved in at the moment of increased stress, providing valuable qualitative data to accompany the physiological measurements.

The combination of the Empatica E4's EDA data and the self-reported contextual information from healthcare workers offered a comprehensive view of the stressors prevalent in their work environment. The data collected from this pilot study laid the groundwork for more detailed analysis, using SWEA's back-end machine learning and NLP models. The goal was to identify patterns and common triggers of stress, which could inform strategies for stress management and intervention in healthcare settings.

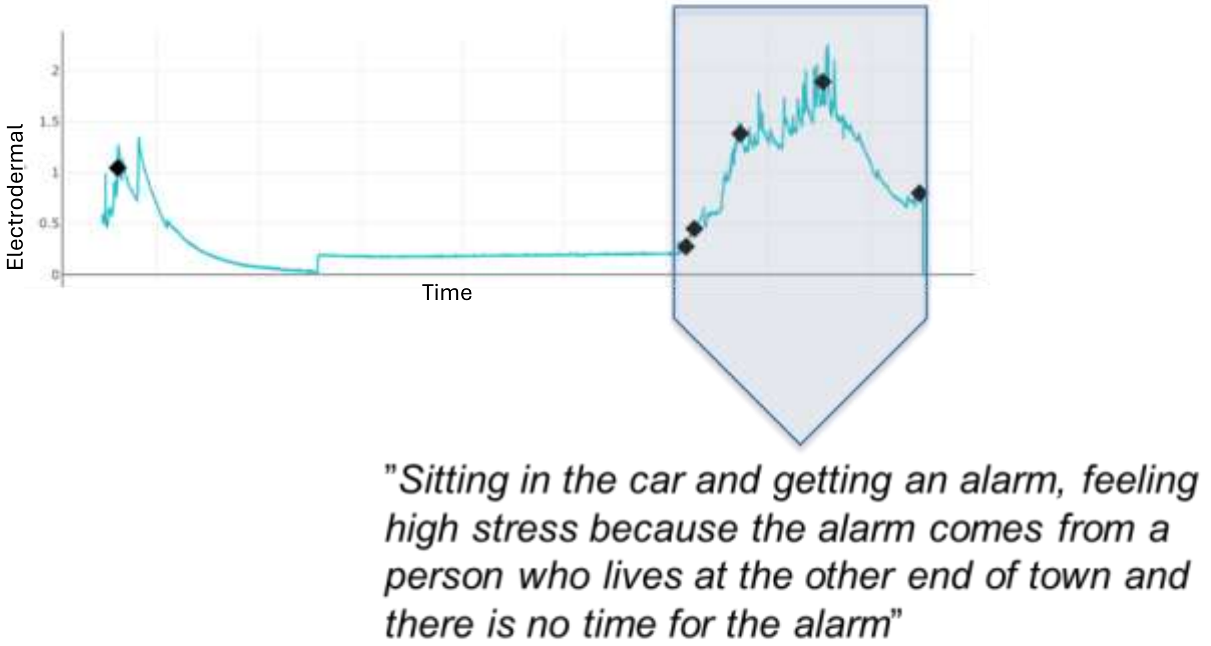
This pilot study represented a significant step forward in understanding the complex nature of workplace stress in high-pressure environments like healthcare, demonstrating the potential of the SWEA Toolbox as a powerful tool for stress analysis and management.

Results

The results of the pilot study using the SWEA Toolbox among healthcare workers revealed significant insights into the sources of stress in their work environment. Notably, a major contributor to stress was the interaction with digital tools intended to aid their work. These tools, equipped with alarms to notify healthcare workers about patients in distress needing assistance, were found to be a significant stressor.

A key issue with these digital tools, as identified in the study, was the lack of contextual awareness regarding the healthcare workers' activities at the time the alarms were triggered. The alarms did not account for what the worker was doing when the notification was received. This lack of synchronization often led to situations where healthcare professionals were interrupted during critical tasks or were already engaged in other demanding activities.

The study found that these interruptions created a sense of disempowerment and loss of control among the healthcare workers. The constant and unpredictable nature of the alarms, combined with their inability to respond immediately in every situation, contributed to heightened stress levels. This was particularly evident in the EDA data collected by the Empatica E4 devices, which showed spikes in stress levels corresponding with the receipt of alarm notifications (Figure 1).



”*Sitting in the car and getting an alarm, feeling high stress because the alarm comes from a person who lives at the other end of town and there is no time for the alarm*”

Figure 1:Iillustration of Electrodermal Activity (EDA) signal that has been processed, alongside commentary from a healthcare worker.

These findings highlight a crucial area for improvement in the design and implementation of digital tools in healthcare settings. There is a need for these systems to be more attuned to the workflow and current activities of healthcare professionals, perhaps by incorporating smarter, context-aware algorithms that can prioritize and time notifications more effectively. Addressing this issue could significantly reduce stress levels and improve the overall efficiency and job satisfaction of healthcare workers.

Conclusion

In conclusion, the exploration of workplace stress through the SWEA Toolbox pilot study, particularly among healthcare workers, has yielded critical insights into stress management and measurement. The integration of EDA monitoring with the Empatica E4 wristband and a diary method of data collection provided a comprehensive approach to understanding stress triggers and responses in a high-pressure environment.

The study’s results particularly highlighted the significant stress caused by interactions with digital tools in the healthcare sector. The lack of contextual awareness in these tools, especially regarding alarm systems, led to increased stress levels due to interruptions and a sense of loss of control among healthcare workers. These findings underscore the necessity for more intelligent and context-sensitive digital tools in healthcare environments.

The SWEA Toolbox, with its combination of real-time physiological data, user input, and advanced analytical capabilities, has demonstrated its potential as an effective tool for both understanding and managing workplace stress. By providing objective, real-time data alongside subjective contextual information, the SWEA Toolbox offers a nuanced view of workplace stress, paving the way for more targeted and effective stress management strategies.

This study not only advances our understanding of stress in high-stress professions but also sets a standard for future research and development in the field of work life science. It shows the importance of considering both physiological and contextual factors in stress measurement and the potential of technology-driven solutions in enhancing workplace well-being.

Ethical Statement

The project "Monitoring Work Stress: The Role of Physiological Measures in Enhancing Workplace Well-being" utilizes Electrodermal Activity (EDA) among other physiological measures to study stress levels in workplace settings. This research is committed to upholding the highest ethical standards in accordance with relevant laws and guidelines, including those set forth in the Law (2003:460) on Ethical Review of Research Involving Humans in Sweden.

All participants have been provided with detailed information about the nature of the research, what it entails, potential risks, and their rights. Participants has been required to give informed consent, affirming that they understand the research and agree to partake voluntarily. Measures have been taken to ensure confidentiality and anonymity of the data collected. Personal identifiers have not been collected in the dataset to prevent any possibility of the data being traced back to individual participants. The data collected from physiological measures are handled with utmost confidentiality. Only authorized personnel have access to the raw data, and it is stored securely in the SWEA server. Data protection compliance will be ensured as per the General Data Protection Regulation (GDPR) and the specific provisions mentioned in the ethical guidelines of the 2003:460 law. The research protocol has been submitted for review and approved by The Swedish Ethical Review Authority (Dnr 2020-03061). This ensures that the research meets the required ethical standards and addresses any potential risks to participants.

The research team is committed to minimizing any potential discomfort or risk to participants. The methodology has been carefully designed to ensure that the stress measurement does not harm or adversely affect the participants. Any potential adverse event or complaint will be addressed promptly, with procedures in place for participants to express their concerns and withdraw from the study at any time without any consequence.

References

1. Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological antecedents and implications. *MIS quarterly*, 831-858.

2. Caruelle, D., Shams, P., Gustafsson, A., & Lervik-Olsen, L. (2024). Emotional arousal in customer experience: A dynamic view. *Journal of Business Research*, *170*(C).

3. Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition*, 65(3), 209-237.

4. Kivimäki, M., & Steptoe, A. (2018). Effects of stress on the development and progression of cardiovascular disease. *Nature Reviews Cardiology*, 15(4), 215-229.

5. Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

6. Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions–A literature review and a call for action. *Journal of Business Research*, 104, 146-160.

7. Ashkanasy, N. M., & Dorris, A. D. (2017). Emotions in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, 4, 67-90.

8. Palm, K., Bergman, A., & Rosengren, C. (2022). Private ICT-Activities and Emotions at Work: A Swedish Diary Study. *Nordic Journal of Working Life Studies*, 12(1), 27-47.

9. Aqajari, S. A. H., Naeini, E. K., Mehrabadi, M. A., Labbaf, S., Dutt, N., & Rahmani, A. M. (2021). pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity. Procedia Computer Science, 184, 99-106.

From Bees to Prisons in Computer Vision: Lessons for Recording Behavior in the Real World

J. Lumetzberger[1] and M. Kampel[1]

[1]Computer Vision Lab, Vienna University of Technology, Vienna, Austria. jennifer.lumetzberger@tuwien.ac.at

This paper addresses computer vision methods for analyzing the behavior of humans and non-human animals in real-life conditions. It emphasizes the need to consider technical, ethical, and environmental challenges when applying these technologies in natural environments. Through three case studies— beehive observation, prison inmate behavior analysis, and dementia patient support—this research demonstrates effective field data collection strategies. By highlighting these approaches, the paper advances behavioral research techniques, moving beyond conventional manual observation methods.

1. Introduction

When analyzing the behavior of humans or non-human animals, relying only on manual observation can present a tedious challenge and might influence the subject's natural behavior. Computer vision techniques can pose a compelling alternative, offering the capacity to record behavior for subsequent image or video data analysis. Consider, for instance, the challenging task of counting bees in a beehive. This poses a significant challenge for humans. On the other hand, computer vision can be deployed to continuously monitor bee hives, thereby facilitating data collection with minimal or zero human intervention [1]. Furthermore, when applied to hive observation, automated systems exert minimal disruption on the natural behavior of bees compared to manual measurements [2].

Careful attention to camera setup is required when developing behavior models and algorithms using computer vision. While it may be tempting to use and adapt data and methods from existing research, we must acknowledge that simply working within the confines of our offices may lead to an oversimplified view of our research problems. We take the perspective that it is essential to go into the real world, engage with experts, and gain a firsthand understanding of the real-world dynamics to produce meaningful and relevant work.

For instance, imagine an attempt to determine whether cows in a stable are lying or standing using pre-existing image databases where cows are fully visible. In the controlled environment of the laboratory, this approach may seem promising. However, practical challenges, such as occlusions due to multiple cows, varying camera angles, or unfavorable lighting conditions, may make the model ineffective in real-world settings. Thus, thoroughly understanding each use case's specific conditions and possibilities becomes paramount.

Although transitioning from the laboratory to the field is crucial, it introduces a unique set of challenges. Prior research in this domain has primarily concentrated on issues related to computer vision and data collection. Jegham et al. [3] explored difficulties such as camera movement, varying backgrounds, and weather conditions, offering insights into the technical aspects of human action recognition. Nayar et al. [4] concentrated on the effects of adverse weather on computer vision, whereas Chandel and Vatta [5] delved into handling occlusions, touching upon challenges like low visibility, shadows, object occlusions, and complex backgrounds. O'Connor et al. [6] expanded the discussion to include practical IoT design strategies for health data collection, addressing consent levels and privacy concerns, albeit without delving deeply into the nuances of different health data types, like images.

Despite discussions on privacy and handling image data by several researchers [7,8,9], there appears to be a gap in comprehensive research addressing the technical, ethical, and environmental challenges of capturing image data in real-world scenarios.

This paper, therefore, aims to bridge this gap by presenting various factors essential for field-based vision data recording, such as environmental conditions, lighting, and privacy concerns. To illustrate these factors comprehensively, we will present three use cases from our research.

The significance of this paper becomes evident when researchers follow its guidelines to organize their recordings, leading to a substantial improvement in their research quality. By mapping actual behaviors accurately rather than depending on theoretical assumptions and synthetic data, the quality of the results can be significantly improved. Understanding and avoiding potential pitfalls through proper organization is crucial. There is no one-size-fits-all approach; this paper emphasizes the necessity to evaluate specific needs and adjust methods accordingly, providing a comprehensive perspective on the methodology.

This paper can benefit a diverse audience, including biologists, psychologists, and computer scientists, who seek to record and model behavior in humans and non-human animals using image data. We aim to empower researchers to tailor their approach to their unique circumstances, fostering a more nuanced and practical understanding of behavior.

We present the following contributions:

1. We provide a comprehensive overview of the methods for capturing behavior through image data in real-world environments. This includes not only technical elements such as managing occlusions and selecting appropriate hardware but also delves into ethical considerations like informed consent and privacy, as well as environmental factors, including indoor versus outdoor settings and the placement of cameras.
2. We illustrate our methodology through three use cases from our research: a) monitoring behavior in beehives, b) studying potentially violent behavior among prison inmates, and c) observing the behavior of residents in care homes.

Section 2 of this paper describes the three use cases in more detail. Section 3 presents various considerations essential for capturing image data in real-world settings, while Section 4 concludes the article.

2. Methodology

Before discussing the three use cases, it is essential to define the term “behavior” within the context of this paper. We adopt a technical viewpoint, characterizing behavior as observable postures, actions, and movements of individuals. This includes various activities, from dynamic actions like running and walking to static states such as sitting or lying down.

2.1 Use case 1: Bee Hive Monitoring

In one of our research projects, we used computer vision technology to observe beehives and detect varroa mites [10]. The system we developed was tailored to the automated identification of these mites. It employed cameras to observe the bees as they arrived and departed from the flight board, allowing for the immediate detection of mites in real time. Our method combined digital image processing with machine learning techniques, executed on single-board processors to ensure efficient and rapid mite detection in the image sequences. As a result, we could automatically create a profile of the colony's exposure to mites, facilitating timely actions for mite management. This project was grounded in real-world data and involved collaborative efforts with beekeepers and the Vienna University of Veterinary Medicine.

2.2 Use case 2: Prison Inmates

We used 3D image analysis to identify potentially critical behaviors in prisons, like suicidal attempts or violent behavior. The system was designed to reduce the risk of safety-related incidents in prisons and provide valuable support to correctional personnel. We successfully implemented these systems temporarily in two Austrian prisons.

2.3 Use case 3: Toilet Guidance in Dementia Patients

We employ computer vision in various research projects related to active assisted living. One example is our ToiletHelp system, which utilizes image data to help individuals with dementia during toileting. The primary objective is to enable these individuals to perform the task independently without needing a second person's assistance.

Given that individuals with dementia often encounter challenges while using the toilet, and offering support can be challenging for both the patient and the caregiver, our approach aims to promote greater autonomy. Our Artificial Intelligence (AI) system identifies the person's position and location within the room, providing corresponding instructions (e.g., “sit down,” “stand up,” “wash your hands”). The system combines action recognition and deviation detection with interactive features to achieve this. In cases where issues are detected, guidance through visual and auditory cues and pre-recorded messages from familiar individuals are offered. At the same time, reminders are sent if something has been overlooked [11].

The action recognition accuracy is critical for the interactive system to deliver appropriate assistance. To validate the effectiveness of our system, we conducted testing in real-world settings and semi-controlled environments, including a private toilet in a geriatric hospital and a semi-public toilet in a day center.

3. Results

We identified several key categories to consider while setting up a computer vision system for behavior recording, encompassing technical, ethical, and environmental aspects. It should be noted that the subsequent subsections may only constitute a partial list of all the factors that need to be considered in your case.

3.1 Ground Truth Data and Annotation

Creating ground truth data becomes vital when advancing data processing, such as crafting AI models or analyzing behavioral patterns. This is especially crucial in scenarios lacking pre-labeled databases or where existing data is overly specific or poorly categorized. Labeling, a tedious yet fundamental task, involves classifying images into categories like "sitting" or "standing" based on the postures depicted. Selecting skilled individuals for labeling, preferably with expertise in the subject matter, enhances accuracy. Utilizing multiple annotators to determine an average and choosing the correct annotation tools are also important considerations.

Take the bee behavior analysis as an example: despite having access to over a thousand hours of bee activity recordings, there was a lack of ground truth data. As a result, we had to manually annotate a variety of scenarios and particularly challenging video segments to create this data. The absence of an automated system for detecting and counting bees made this labor-intensive manual process necessary. For example, annotating a mere 13 seconds of video, which included 143 different bee tracks, demanded over 8 hours of work. In total, we annotated 197 seconds of bee hive footage. In contrast, the task of annotating 77,000 images from a prison study was accomplished by a single annotator in two months, highlighting the variable time investment required for annotation tasks.

In developing the action recognition algorithm for the ToiletHelp system, we engaged three annotators in a joint effort. This collaborative approach was designed to reduce potential biases and errors from individual annotations, thus ensuring greater accuracy and reliability in the ground truth data established.

3.2 Temporal Considerations: Duration and Frequency

Temporal factors encompass considerations such as the required data quantity, different seasons, times of the day, or particular behaviors like animal mating periods. There are also questions regarding the number of subjects and when they are available. For example, there were ongoing recordings in the prison project, one lasting two weeks and another a month, in various prisons. Frame rate is a vital consideration, particularly for subjects that move quickly. While higher frame rates provide more detail, they also result in larger volumes of data.

In the bee project, a longtime test was performed where video sequences of 7 days were recorded, with a frame rate of 30 fps at a resolution of 1920x1080 pixels [1]. The recording time window was between 06:00 and 21:00, processing over 73 hours of recordings.

The ToiletHelp system uses the maximum frame rate of the Orbbec Astra sensor with 30 fps. In contrast, the prison project utilized a lower frame rate of 7.5 fps to synchronize with thermal module capabilities, recording in 10 rooms over 3.5 months, resulting in over 72,000 recorded frames and a data size of 2,6 TB.

These examples highlight the importance of carefully planning and adapting data collection methods to meet project goals while accommodating technical limitations and temporal dynamics.

3.3 Environmental Factors

Environmental factors impact data gathering, necessitating protective measures for the equipment used. For outdoor bee tracking, specialized hardware enclosures were developed to shield against weather elements, encapsulating the camera in a robust box topped with acrylic glass (see Figure 1). This design facilitated a self-sufficient, no-external-computer-required setup crucial for outdoor environments.

Prisons, although indoor, also constitute a unique environment. The equipment needs to be designed in a safe way and not being able to be damaged or manipulated. In one room, aware of the occupant's history of damaging furniture, we placed a sensor encased in a sturdy, tamper-proof housing to ensure its safety and integrity (see Figure 2).

For the ToiletHelp system, we selected a sensor suitable for small rooms. The device, equipped with an Orbec Astra depth sensor, accurately captures depth maps up to 7 meters within a 1-meter range, offering a 60° horizontal and 49.5° vertical field of view, ideal for compact indoor areas. We primarily used bathrooms with power outlets, but alternative power sources like batteries could be used in different environments where outlets are unavailable.

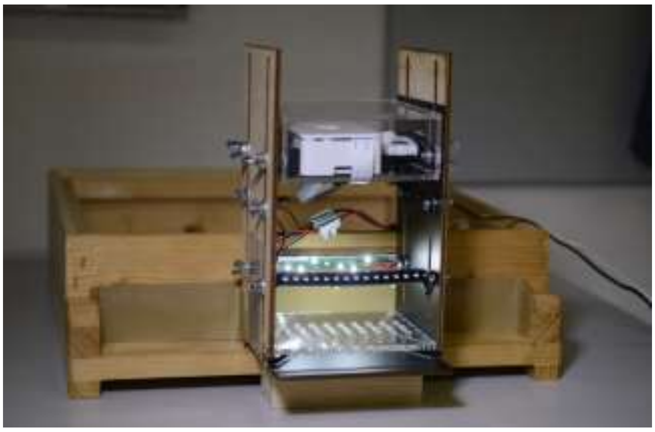


Fig. 1: A picture of the final prototype for bee mite monitoring. It is mounted at the entrance of a bee hive floor. The acrylic glass on the top protects the computing unit from rain. (Image from [10])

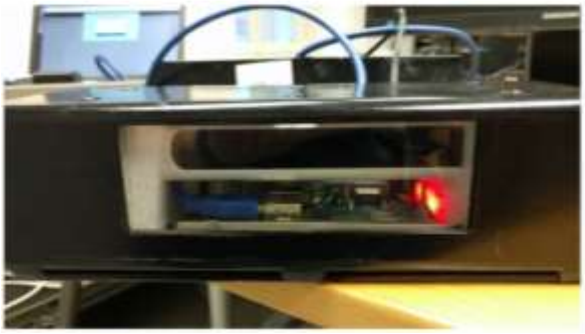


Fig. 2: A manipulation-proof sensor unit in a robust housing.

3.4 Hardware Selection

Choosing the proper hardware for behavior monitoring is crucial, balancing cost, precision, and user-friendliness. Hardware selection can be based on factors like the camera's field of view, resolution, and integrated AI capabilities. A compact computer like a Raspberry Pi might be practical for outdoor use or limited space as it eliminates the need for a separate computer. The choice between running AI algorithms locally or processing data on a remote server will also influence the hardware selection, especially considering the processing power limitations in space-constrained, outdoor environments.

In the prison and care scenarios, a combination of a Raspberry Pi and an Orbbec Astra RGBD sensor was initially used. Later, the Raspberry Pi was replaced with an NVIDIA Jetson Nano for the prison project to facilitate real-time processing of moderately sized neural networks. Only depth data was used for human recordings to ensure privacy, while RGB data was utilized for bee monitoring. The Raspberry Pi in the bee monitoring setup was equipped with the Raspberry Camera Module v2.1.

Sometimes, exploring commercial solutions rather than building something from scratch is advantageous. However, custom development may be necessary if these solutions don't meet specific needs.

3.5 Installation, Camera Positioning and Remote User Interface

For bee hive monitoring, we strategically positioned the camera above the hive's entrance for a top-down view, aligned with the hive's side walls to prevent bees from exiting or entering from the sides. The camera's height remained constant, ensuring consistent bee size in recordings. A transparent glass plate above the entrance area prevents bee overlap, simplifying the analysis to 2D. This setup, including a monotonous gray plate at the bottom, requires consideration of dirt accumulation over time. In the ToiletHelp system, the depth sensor is strategically installed on the wall or ceiling, providing a comprehensive view of the room. This placement ensures that key areas like the toilet and wash basin are within the sensor's field of view.

For the prison project, data recording strategies were planned through both online and in-person meetings with prison managers. This preparation included selecting specific rooms like recreation, TV, and workshop areas, where inmate interaction is most frequent. Room size and sensor placement were critical, avoiding places like long corridors due to the sensor's 7-meter range limit. Installation planning involves identifying power sources and creating room plans to position the sensor optimally for the best coverage.

Another consideration during installation is how to manage the location of the subjects. In care homes and prisons, installations were carried out under supervision from staff members of the respective organizations. For bee monitoring, bees were enclosed within the prototype to maximize footage. Different scenarios or species might require a species expert for handling during installation or transportation. Planning these aspects is crucial for a smooth setup process.

Moreover, consider implementing remote system control, especially if you can only sometimes be on-site for immediate problem-solving. Remote access allows for interventions like restarting devices from a distance. The ToiletHelp project demonstrates this capability through a sensor linked to a platform with Wi-Fi connectivity. This functionality is particularly beneficial when the system is moved to a different room, allowing for reconfiguring essential components like the toilet and basin. Such adaptability ensures that the sensor continues to function effectively.

3.6 Ethical Considerations and Subject Interaction

Whenever you engage with subjects, whether humans or animals, it's crucial to determine the best way to involve them effectively and ethically to gain their consent to participate in your study and use the recordings. In the ToiletHelp project, two end-user organizations recruited patients. Informative sessions about the system were conducted for carers and older adults, with a detailed informed consent form provided to each participant containing all the information about the testing, and all doubts were clarified at 1-1 talks. This was challenging with dementia patients, requiring legal representatives or family members to sign the consent. The process, including additional family member consent for those unable to fully comprehend, was ethically approved.

Ethical approval for both the prison and dementia projects was sought and obtained from an ethics committee, aiming to uphold high ethical standards. To address new and evolving ethical challenges, we established a project-internal ethics board dedicated to continuous dialogue and review. Additionally, we involved experts in sociology and law to ensure the proper implementation of the projects.

For the ToiletHelp project, patient recruitment was conducted by two end-user organizations. We organized informative sessions for caregivers and elderly participants to thoroughly understand the system, supported by a comprehensive informed consent form. This form detailed the testing procedures and ensured all participant queries were addressed through one-on-one discussions. The consent process for individuals with dementia was carefully managed, requiring legal representatives or family members to sign on their behalf.
In the case of the prison project, consent was secured from inmates for installations in single rooms. They were informed about the purpose of the recordings both verbally and through a written consent letter tailored to the project. On the other hand, the concept of informed consent does not apply to projects with non-human subjects, like bees. However, when human interaction occurs in these settings, such as during maintenance work, it's important to inform those involved about the project and take steps to protect their privacy.

3.7 Impact on Subjects and Influence on Natural Behavior

Two different setups that impacted the bees' natural behavior were considered for the beehive project. The 3D setup allowed unrestricted bee movement, using a camera near the hive entrance or above a landing pad. The 2D setup, in contrast, had an entrance structure, restricting bees from walking through and ensuring all movement was recorded. We opted for the 2D setup to avoid occlusion issues, as it limits bees to 2D movement and prevents them from flying over each other. This may have affected natural bee behavior, but it was suitable for our goals of counting bees and detecting varroa mites. However, this setup could impact results for studying social behavior or movement. In contrast to non-human subjects, where environmental changes due to hardware can influence behavior, humans might alter their behavior simply by being aware of sensors or the project itself.

In the prison setting, inmate awareness of monitoring could lead to less aggression and better behavior. In the ToiletHelp scenario, knowing about the sensor might cause stress, extra effort, avoidance of the monitored toilet, or a preference for it. It is essential to consider how subject awareness might alter behaviors.

3.8 Privacy, Security, and Data Management

Ensuring data from unauthorized access is crucial, as is deciding whether to process data on-site at the sensor or a remote location. For example, in the ToiletHelp project, only numerical data, such as the distance between a person and the toilet, is transmitted to a server without sending any images.

In the prison project, all processing of the potentially sensitive images captured by the 3D depth sensor occurs locally on the single-board computing unit NVIDIA Jetson Nano. The specially developed recording software was used and thoroughly tested in advance under laboratory conditions, which enables fail-safe data storage over more extended periods, automatically prevents data corruption, and restarts the recording automatically if problems are detected. This was important as access to the devices is only possible to a limited extent, and any failure would lead to a reduction in the amount of data that should be avoided as far as possible.

3.9 Handling Occlusions

In 3D bee movement, occlusions like shadows need consideration. Using a background plate can limit bee shadows directly below them. Placing a transparent plate above bees prevents them from overlapping, eliminating occlusions in a 2D setup. Challenges arose with occlusions among multiple people in one cell in the prison project. The system's performance in these scenarios improved through iterative enhancements in path modeling and location prediction for occluded subjects or those outside the sensor's view.

3.10 Light Conditions

Lighting conditions significantly impact data quality, necessitating careful selection of lighting setups. This varies with data types like RGB, depth, or infrared. In the bee project, overhead illumination ensures shadows fall directly below bees. A custom-built entrance allows for uniform artificial lighting, unlike 3D setups, which must account for varying natural light. The door is consistently lit by LED lights, minimizing illumination changes. Depth data is less sensitive to lighting, but variations can still affect quality, as seen in a prison cell where changing light from a window influences data quality. Testing in varying light conditions is essential for assessing data impact.

3.11 Image properties

The resolution and format of images significantly affect data analysis and storage requirements. In the prison project, recordings were performed at 640x480 resolution. For the bee monitoring, varying cameras produced resolutions of 1416x540 or 1920x1080 pixels. The 2D setup for bees offers better image quality, allowing for closer, more precise camera alignment, resulting in larger bee images suitable for detailed analysis, such as parasite detection. It is essential to consider the distance between the camera and the subject and the specific data requirements, such as high resolution to discern details or whether a lower resolution is adequate for purposes like counting subjects.

4. Conclusion

This paper underscores the significance of considering a range of factors when collecting image data in real-world scenarios. While manual observation is valuable, it can be laborious and potentially disrupt the natural behaviors being observed. Therefore, computer vision techniques offer an appealing alternative for behavior recording. By examining three practical use cases, we have demonstrated the critical technical, ethical, and environmental factors to consider for successful field-based vision data collection.

By carefully considering these factors and adjusting methodologies accordingly, researchers have the potential to greatly enhance the quality of their research. This strategy allows for a more accurate representation of real-world behaviors, reducing dependence on theoretical models or artificial data and strengthening the research outcomes. Whether collecting new data in the field or checking if existing datasets accurately reflect the behaviors being studied, it is essential to recognize that there is no universal solution. This paper highlights the importance of customizing research approaches to meet specific needs, providing a thorough overview of how to collect image data in field settings effectively.

References

1. Schurischuster, S., Zambanini, S., Kampel, M., & Lamp, B. (2016). Sensor study for monitoring varroa mites on honey bees (apis mellifera). *In Proceedings of Visual observation and analysis of Vertebrate And Insect Behavior Workshop (VAIB 2016)* (p. 4).

2. Tu, G. J., Hansen, M. K., Kryger, P., & Ahrendt, P. (2016). Automatic behaviour analysis system for honeybees using computer vision. *Computers and Electronics in Agriculture*, 122, 10-18.

3. Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32, 200901.

4. Nayar, S. K., & Narasimhan, S. G. (1999, September). Vision in bad weather. *In Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 820-827). IEEE.

5. Chandel, H., & Vatta, S. (2015). Occlusion detection and handling: a review. *International Journal of Computer Applications*, 120(10).

6. O'Connor, Y., Rowan, W., Lynch, L., & Heavin, C. (2017). Privacy by design: informed consent and internet of things for smart health. *Procedia computer science*, 113, 653-658.

7. Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE access*, 4, 2751-2763.

8. Noiret, S., Ravi, S., Kampel, M., & Florez-Revuelta, F. (2022, June). On The Nature of Misidentification With Privacy Preserving Algorithms. *In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 422-424).

9. Wu, Z., Wang, Z., Wang, Z., & Jin, H. (2018). Towards privacy-preserving visual recognition via adversarial training: A pilot study. *In Proceedings of the European conference on computer vision (ECCV)* (pp. 606-624).

10. Schurischuster, S., Remeseiro, B., Radeva, P., & Kampel, M. (2018). A preliminary study of image analysis for parasite detection on honey bees. In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018*, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15 (pp. 465-473). Springer International Publishing.

11. Ballester, I., & Kampel, M. (2022). Automated vision-based toilet assistance for people with dementia. *Human Factors in Accessibility and Assistive Technology*, 37, 21.

Implementing Behavioural Decision-Making Tasks into Immersive Virtual Reality Environments

J. Hadaschik[1,2], I. van Sintemaartensdijk[2], L. Rabago Mayer[2,], M. Stel[2], K. Massar[1]

[1] Maastricht University, Department of Work and Social Psychology, Maastricht, Netherlands, j.hadaschik@maastrichtunviersity.nl

[2] University of Twente, Department of Psychology of Conflict, Risk and Safety, Enschede, Netherlands

Short Abstract

This talk presents a proof-of-concept study which implemented two behavioural decision-making tasks in two immersive Virtual Reality (VR) environments to investigate risk-taking and delay discounting under the influence of socio-environmental cues of harshness. Implementing behavioural tasks in VR environments can allow to investigate the effect of cues on decision-making, can improve ecological validity and open up new research areas.

Introduction

This proof-of-concept study explored a new way to test a theoretical framework informed by behavioural ecology [1], [2], [3] and developmental plasticity [4], which states that cognition and decision-making can be influenced by environmental cues in a way that can be potentially adaptive, i.e. increase fitness [5]. It is theorised that in an environment that signals threats to health and survival it can increase fitness to take risks and invest time and effort in behaviour that aids current survival rather than future well-being [6]. Behaviours that are oftentimes labelled 'dysfunctional' or 'maladaptive', such as impulsivity or risk-taking, might aid short-term survival in adverse environments [7].

Behaviours that maximise current reward at the potential expense of future well-being are especially present in current or former residents of deprived neighbourhoods, e.g. poor health behaviours, high risk-taking, and high delay discounting [8], [9], [10], [11], [12]. Evolutionary-developmental psychology theory postulates that these behaviours are more prevalent in individuals who receive cues from their environment that signal their survival and well-being are in danger [5]. These socio-environmental cues of harshness (COH) are especially present in deprived neighbourhoods in the form of derelict housing and infrastructure, poverty, poor health and low life expectancy of other residents and signs of criminal activities [13], [14] The relationship between exposure to COH and short-term focussed decision-making might be moderated by Early Life Stress (ELS), with ELS leading to an increase in biased cognition in the presence of COH [15]. While there is ample correlational and longitudinal evidence for a positive association between exposure to COH and behavioural strategies that rely on short-term focussed decision-making, experimental evidence is scarce and the results are mixed [15], [16], [17].

Behavioural decision-making tasks that investigate risk-taking or the discounting of delayed rewards are usually administered on 2D computer interfaces in a lab environment in the absence of environmental cues, e.g. [18], [19]. While this approach allows for high experimental control, its ecological validity can be criticised due to the complete absence of stimuli compared to real world decision-making. Virtual Reality offers an opportunity to assess risk-taking and decision-making in the presence of multisensory and dynamic stimuli that are contextually embedded in immersive virtual environments [20], [21].

The current proof-of-concept study aims to fill this gap in the literature by using immersive Virtual Reality to expose participants to COH similar to those that are commonly experienced in a deprived neighbourhood (e.g. signs of poverty and deprivation, cues of higher morbidity and mortality in the community, cues of social conflict, violence and aggression). Risk-taking and decision-making is measured by the Balloon Analogue Risk Task [19] and a delay discounting task respectively, both of which are implemented in each of the two experimental VR environments (deprived neighbourhood vs. non-deprived neighbourhood).

Methods

Design

The study used an experimental between-subjects design. The dependent variables were “risk-taking” (measured by number of inflations on the VR version of the BART) and “discounting of future suffering”. The independent variable was “exposure to COH” with two levels (low versus high amount of cues of harshness). The two levels of the independent variable corresponded to two different VR environments: a deprived neighbourhood with a high amount of COH and a non-deprived neighbourhood with a low amount of COH. The covariate ELS was measured by three self-report questionnaire items on childhood neighbourhood quality.

Participants

Convenience sampling was used to recruit 58 student participants aged between 18 and 59 (M = 22.6, SD = 5.4). Participants were required to be at least 18 years old, to not have participated in a similar previous study, not have red-green colour blindness and not be pregnant (due to higher susceptibility to motion/cyber-sickness during pregnancy [22], [23]). The majority identified as women (53.4%), 46.6% identified as men. Most participants had attained high school level education (82.8%), fewer had an undergraduate degree (15.5%) or postgraduate degree (1.7%). The majority were German (60.3%) or Dutch (19%). Psychology students of University of Twente received 1.5 study participation credits. All participants had the chance to win one of five 20 Euro shopping vouchers, based on their performance on the BART.

The study was approved by the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at University of Twente, Netherlands (request number 220322). The study was conducted in accordance with the principles of the Declaration of Helsinki. Participants gave informed consent prior to taking part in the study and had the right to withdraw from the study at any moment without having to give a reason.

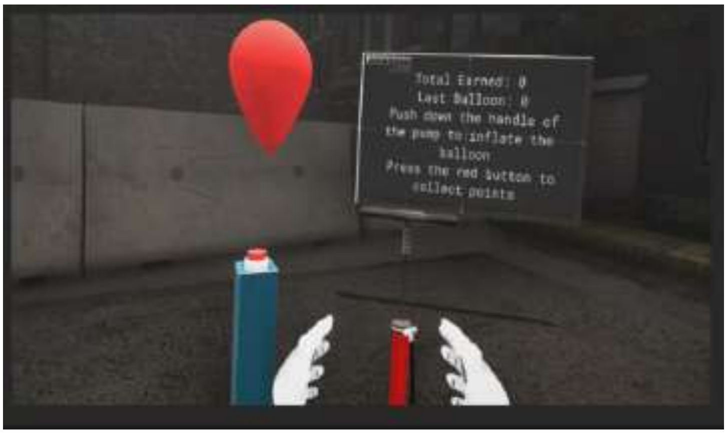
Materials

VR environments were created and presented to participants using Unity version 2020.3.28f1. In total, there were three VR environments: a practice scene, and the two experimental conditions. The practice scene contained both decision-making tasks in a neutral and calm VR environment to allow participants to practice interacting with the tasks before performing them in the experimental condition. Previous pilot testing had confirmed that too many participants needed support from the experimenters when first interacting with the tasks. The high COH condition was modelled after a neighbourhood that is affected by multiple deprivation. Examples of auditory stimuli embedded in this VR neighbourhood were voices of people arguing from inside houses, police sirens, and close traffic noise. Visual features of the environment showed derelict housing (e.g. broken windows), poor infrastructure (e.g. pot holes in the road, derelict bus stop, disused and broken basketball court) and disruption of public services (e.g. littered streets, signs of poor waste management). The control condition was modelled after a middle-class, suburban neighbourhood that is not affected by deprivation. Examples of auditory stimuli embedded in this VR neighbourhood were voices of people having friendly conversations inside houses, birds singing, and distant traffic noise. Visual features of the environment showed intact housing, well-kept roads and public transport infrastructure (e.g. functional and clean bus stop) and well-functioning public services (e.g. no litter on the streets, well-kept park and greenery at the centre of the neighbourhood).

The risk-taking task consisted of a VR version of the Balloon Analogue Risk Task [19]. Participants interacted with the task by using their hands to use a virtual pump to inflate the balloon. For each inflation, participants could win 50 points that were stored in a temporary bank. Each balloon had a random breaking point. If the balloon breaks, all points in the temporary bank are lost. At any point, the participant could decide to stop inflating, all points were stored in a permanent bank and a new balloon was presented. Participants therefore had to make a trade-off between inflating the balloon as much as possible to maximise their points, while at the same time being cautious to not risk that the balloon breaks. Figure 1 presents the VR version of the BART embedded in the high COH condition (deprived neighbourhood).

Figure 1

Participant's view of the Balloon Analogue Risk Task implemented in an immersive VR deprived neighbourhood that is high in cues of harshness

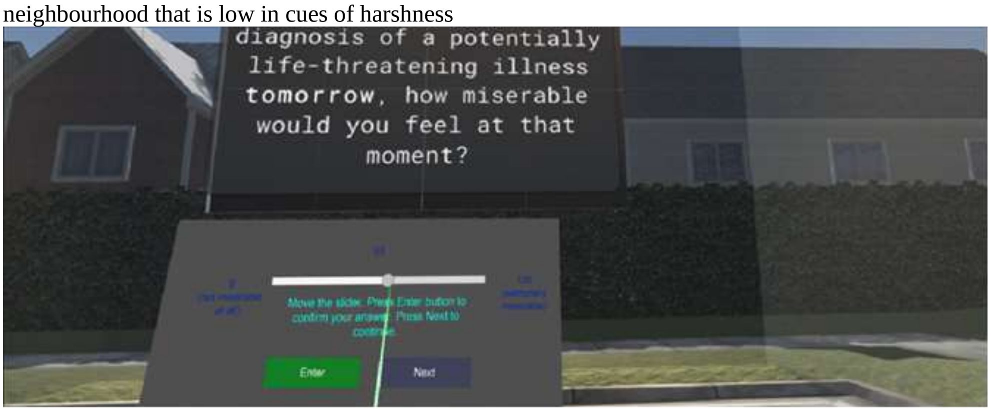


Total Earned: 0
Last Balloon: 0
Push down the handle of the pump to inflate the balloon
Press the red button to collect points

The delay discounting task consisted of a VR version of the 'discounting of future suffering' task [24] which measures to what extent participants would feel psychological distress if they were to receive the diagnosis of a potentially life-threatening illness at six time points in the future. For each delay period (i.e. tomorrow, in three months, in one year, in three years, in 10 years, in 30 years) they indicated how miserable they would feel on a scale from 0 (not miserable at all) to 100 (extremely miserable). Figure 2 shows the task embedded in the low COH condition (non-deprived neighbourhood).

A post-VR questionnaire asked participants how they experienced the VR neighbourhoods,. Using a sliding scale from 0 to 100, participants indicated to what extent they were immersed in the virtual environment ("While being in the virtual neighbourhood, did you have a sensation of 'being there' (in the virtual environment)"), to what extent they felt safe ("Please rate how safe you felt while being in the virtual neighbourhood"), and how stressed they felt ("Please rate how stressed you felt while being in the virtual neighbourhood"

Figure 2 Participant's view of the 'discounting of future suffering' task, implemented in an immersive VR non-deprived neighbourhood that is low in cues of harshness



diagnosis of a potentially life-threatening illness tomorrow, how miserable would you feel at that moment?

55

0 (not miserable at all)

100 (extremely miserable)

Move the slider. Press Enter button to confirm your answer. Press Next to continue.

Note: The text on the text panel above the sliding scale reads: "If you were to receive the diagnosis of a life-threating illness tomorrow, how miserable would you feel at that moment?" The text below the slider scale reads: "Move the slider. Press Enter button to confirm your answer. Press Next to continue." The left side of the scale is labelled " 0 (not miserable at all)". The right side of the scale is labelled "100 (extremely miserable)". The currently selected value is displayed above the slider scale (here: 55).

Procedure

Participants were randomly allocated to the high COH/deprived neighbourhood condition or the low COH/non-deprive neighbourhood condition. Before entering one of the two conditions, they practiced interacting with the two tasks in a neutral 'practice scene' until they understood how the tasks worked and they had no more questions. In the neighbourhood environments, they were instructed to follow a designated path (marked by arrows on the pavement) for about seven minutes. At the end, they completed both the VR version of the BART (15 balloons) as well as the VR version of the 'discounting of suffering' task. The order of the tasks was reversed for 50% of the sample to prevent order effects. After completing the tasks, they filled in the questionnaire assessing their experience of the VR environment.

Results

Figure 3 presents mean inflations across 15 balloons (excluding exploded balloons, following [19]) on the VR BART. A Wilcoxon rank sum test showed that participants did not take more risks (as indicated by a higher number of inflations) in the high COH condition/deprived neighbourhood, compared to the low COH condition/non-deprived neighbourhood (W = 280.5, p > .05)

Figure 3 Box-plot of mean inflations on the VR BART across low- and high COH conditions

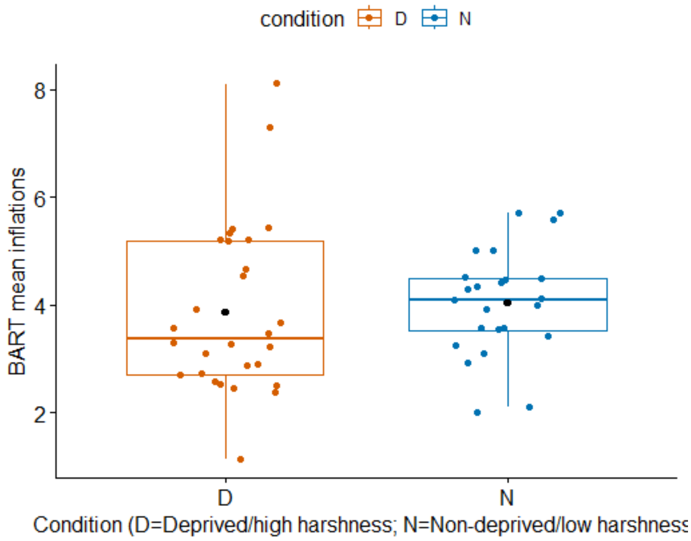
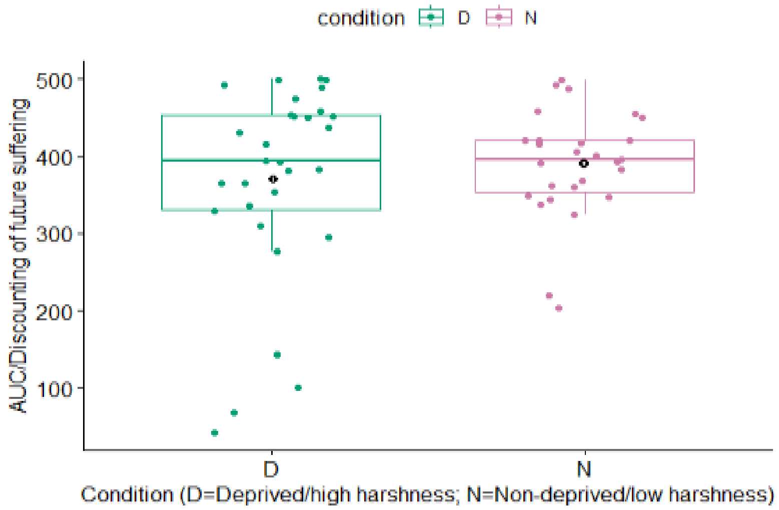


Figure 4 shows the area under the discounting curve (AUC) for the ‘discounting of future suffering’ task in both conditions. A higher AUC represents lower discounting. A Wilcoxon rank sum test showed that participants did not discount their future suffering more (as indicated by a lower AUC) in the high COH condition/deprived neighbourhood, compared to the low COH condition/non-deprived neighbourhood (W = 409, p > .05).

Figure 4 The area under the discounting curve (AUC) in the high COH and low COH condition



Results of the post-VR questionnaire indicated that participants felt significantly less safe (Wilcoxon rank sum test; W = 107, p<.001) and significantly more stressed (Wilcoxon rank sum test; W = 660.5, p<.001) in the COH condition/deprived neighbourhood. An independent samples t-test showed that there was no difference in the level of immersion between conditions (t(56) = 1.5673, p>.05).

Discussion

The aim of this proof-of-concept study was to implement behavioural decision-making- and risk-taking tasks in two immersive VR neighbourhood environments to investigate the effect of exposure to multisensory, embedded COH on risk-taking and delay discounting. Results showed that COH in the VR environments did not have an effect on risk-taking as measured by the BART or on the discounting of future suffering. However, participants felt less safe and more stressed in the high COH condition. There was no difference in immersion across conditions.

The study was limited by a small, non-representative student sample. There is evidence that effects of COH on decision-making might only be observable in individuals who grew up in high-stress environments [15] (but also see [16] for a replication study with divergent results). In the current student sample, exposure to early life stress (ELS) was extremely low. Future studies should attempt to recruit more diverse samples, including people with low SES and the experience of ELS.

Although there was no difference regarding risk-taking and delay-discounting, results showed that participants' levels of stress and feelings of safety were strongly affected by exposure to the high COH condition. This indicates that VR is a promising tool to measure decision-making in the presence of stimuli. Before meaningful comparisons can be made between behavioural tasks in 2D versus VR, the VR versions of tasks need to be validated against a set of self-report instruments and participants behaviour in VR versus with 2D administered tasks needs to be compared.

References

1. S. C. Stearns, 'Trade-Offs in Life-History Evolution', *Functional Ecology*, vol. 3, no. 3, pp. 259–268, Jan. 1989, doi: 10.2307/2389364.
2. B. J. Ellis, 'Timing of pubertal maturation in girls: An integrated life history approach', *Psychological Bulletin*, vol. 130, no. 6, pp. 920–958, Nov. 2004, doi: http://dx.doi.org/10.1037/0033-2909.130.6.920.
3. D. Nettle, 'Why Are There Social Gradients in Preventative Health Behavior? A Perspective from Behavioral Ecology', *PLOS ONE*, vol. 5, no. 10, p. e13371, Oct. 2010, doi: 10.1371/journal.pone.0013371.
4. M. J. West-Eberhard, *Developmental Plasticity and Evolution*. Oxford University Press, 2003.
5. W. E. Frankenhuis, K. Panchanathan, and D. Nettle, 'Cognition in harsh and unpredictable environments', *Current Opinion in Psychology*, vol. 7, no. Supplement C, pp. 76–80, Feb. 2016, doi: 10.1016/j.copsyc.2015.08.011.
6. J. Fenneman and W. E. Frankenhuis, 'Is impulsive behavior adaptive in harsh and unpredictable environments? A formal model', *Evolution and Human Behavior*, vol. 41, no. 4, pp. 261–273, Jul. 2020, doi: 10.1016/j.evolhumbehav.2020.02.005.
7. W. E. Frankenhuis, E. S. Young, and B. J. Ellis, 'The Hidden Talents Approach: Theoretical and Methodological Challenges', *Trends in Cognitive Sciences*, vol. 24, no. 7, pp. 569–581, Jul. 2020, doi: 10.1016/j.tics.2020.03.007.
8. M. H. Algren, C. K. Bak, G. Berg-Beckhoff, and P. T. Andersen, 'Health-Risk Behaviour in Deprived Neighbourhoods Compared with Non-Deprived Neighbourhoods: A Systematic Literature Review of Quantitative Observational Studies', *PLOS ONE*, vol. 10, no. 10, p. e0139297, okt 2015, doi: 10.1371/journal.pone.0139297.
9. N. Canale, A. Vieno, M. Lenzi, M. D. Griffiths, D. D. Perkins, and M. Santinello, 'Cross-national differences in risk preference and individual deprivation: A large-scale empirical study', *Personality and Individual Differences*, vol. 126, pp. 52–60, May 2018, doi: 10.1016/j.paid.2018.01.006.

10. F. I. Matheson, H. L. White, R. Moineddin, J. R. Dunn, and R. H. Glazier, 'Drinking in context: the influence of gender and neighbourhood deprivation on alcohol consumption', *J Epidemiol Community Health*, vol. 66, no. 6, pp. e4–e4, Jun. 2012, doi: 10.1136/jech.2010.112441.
11. R. J. Noonan, L. M. Boddy, Z. R. Knowles, and S. J. Fairclough, 'Cross-sectional associations between high-deprivation home and neighbourhood environments, and health-related variables among Liverpool children', *BMJ Open*, vol. 6, no. 1, p. e008693, Jan. 2016, doi: 10.1136/bmjopen-2015-008693.
12. R. J. Tunney and R. J. E. James, 'Individual differences in decision-making: evidence for the scarcity hypothesis from the English Longitudinal Study of Ageing', *Royal Society Open Science*, vol. 9, no. 10, p. 220102, Oct. 2022, doi: 10.1098/rsos.220102.
13. S. Jivraj, E. T. Murray, P. Norman, and O. Nicholas, 'The impact of life course exposures to neighbourhood deprivation on health and well-being: a review of the long-term neighbourhood effects literature', *European Journal of Public Health*, vol. 30, no. 5, pp. 922–928, Oct. 2020, doi: 10.1093/eurpub/ckz153.
14. T. Nakaya *et al.*, 'Associations of All-Cause Mortality with Census-Based Neighbourhood Deprivation and Population Density in Japan: A Multilevel Survival Analysis', *PLOS ONE*, vol. 9, no. 6, p. e97802, Jun. 2014, doi: 10.1371/journal.pone.0097802.
15. V. Griskevicius, J. M. Tybur, A. W. Delton, and T. E. Robertson, 'The influence of mortality and socioeconomic status on risk and delayed rewards: A life history theory approach', *Journal of Personality and Social Psychology*, vol. 100, no. 6, pp. 1015–1026, 2011, doi: 10.1037/a0022403.
16. G. V. Pepper, D. H. Corby, R. Bamber, H. Smith, N. Wong, and D. Nettle, 'The influence of mortality and socioeconomic status on risk and delayed rewards: a replication with British participants', *PeerJ*, vol. 5, Jul. 2017, doi: 10.7717/peerj.3580.
17. D. Nettle, G. V. Pepper, R. Jobling, and K. B. Schroeder, 'Being there: a brief visit to a neighbourhood induces the social attitudes of that neighbourhood', *PeerJ*, vol. 2, p. e236, Jan. 2014, doi: 10.7717/peerj.236.
18. W. K. Bickel, M. N. Koffarnus, L. Moody, and A. G. Wilson, 'The behavioral- and neuro-economic process of temporal discounting: A candidate behavioral marker of addiction', *Neuropharmacology*, vol. 76, Part B, pp. 518–527, Jan. 2014, doi: 10.1016/j.neuropharm.2013.06.013.
19. C. W. Lejuez *et al.*, 'The balloon analogue risk task (BART) differentiates smokers and nonsmokers', *Experimental and Clinical Psychopharmacology*, vol. 11, no. 1, pp. 26–33, Feb. 2003, doi: 10.1037/1064-1297.11.1.26.
20. T. D. Parsons, 'Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences', *Front. Hum. Neurosci.*, vol. 9, Dec. 2015, doi: 10.3389/fnhum.2015.00660.
21. L. R. Bruder, L. Scharer, and J. Peters, 'Reliability assessment of temporal discounting measures in virtual reality environments', *Sci Rep*, vol. 11, no. 1, p. 7015, Mar. 2021, doi: 10.1038/s41598-021-86388-8.
22. F. O. Black, 'Maternal susceptibility to nausea and vomiting of pregnancy: Is the vestibular system involved?', *American Journal of Obstetrics and Gynecology*, vol. 186, no. 5, Supplement 2, pp. S204–S209, May 2002, doi: 10.1067/mob.2002.122602.
23. K. Nesbitt, S. Davis, K. Blackmore, and E. Nalivaiko, 'Correlating reaction time and nausea measures with traditional measures of cybersickness', *Displays*, vol. 48, pp. 1–8, Jul. 2017, doi: 10.1016/j.displa.2017.01.002.
24. J. Hadaschik, I. van Sintemaartensdijk, R. A. C. Ruiter, K. Stel, and K. Massar, 'Associations Between Early Life Stress, Socio-Sexual Behaviors, Life History Traits and Adult Health-Behavior and Risk-Taking: A Structural Equation Modelling Approach', *in review*.
25. C. de-Juan-Ripoll, J. L. Soler-Domínguez, J. Guixeres, M. Contero, N. Álvarez Gutiérrez, and M. Alcañiz, 'Virtual Reality as a New Approach for Risk Taking Assessment', *Frontiers in Psychology*, vol. 9, 2018, Accessed: Feb. 17, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02532

26. C. de-Juan-Ripoll, J. Llanes-Jurado, I. A. C. Giglioli, J. Marín-Morales, and M. Alcañiz, ‘An immersive virtual reality game for predicting risk taking through the use of implicit measures’, *Applied Sciences*, vol. 11, no. 2, p. 825, 2021.

Rest activity pattern as digital behavioral marker of apathy in patients with dementia at home

I. Sezer[1], N. Saidi[1], S. Aflalo[1], E. Karpinski[1], V. Godefroy[1], R. Migliaccio[1], R. Levy[1] and B. Batrancourt[1]

1 Sorbonne Université, Institut du Cerveau - Paris Brain Institute – ICM, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France. benedicte.batrancourt@upmc.fr

Introduction

According to the World Health Organization (WHO), there is currently more than 55 million people diagnosed with dementia and 10 million of new diagnoses each year appear worldwide. With the demographic shifts in the global population, marked by both growth and aging, the number of new cases will keep increasing. Dementia has devastating impacts on patients and their families as well as a negative economic effect. The economic burden was accounted in 2019 to reach 1.3 trillion US dollars and was half paid by caregivers [1]. Dementia is responsible of major disabilities in every-day life [2] and negatively impacts patients' autonomy. This loss of autonomy largely stems from the Behavioral and Psychological Symptoms in Dementia (BPSD) including apathy, depression, anxiety, delusions, hallucinations, sexual or social disinhibition, sleep–wake cycle disturbances, aggression, agitation, and other inappropriate behaviors [3]. Apathy is a frequently observed neuropsychiatric symptom in neurodegenerative diseases. It is prevalent in 60% of Alzheimer's disease (AD) patients and present in 73-100% of behavioral variant frontotemporal dementia (bvFTD) cases [4, 5]. It is a significant diagnostic criterion for these diseases, along with symptoms of disinhibition which are also frequently observed in both bvFTD and AD [6].

Traditionally, apathy has been viewed as a symptom indicating loss of interest or emotions. In 2006, Levy and Dubois refined the definition of apathy to a quantitative reduction of self-generated voluntary and purposeful behaviors [5]. Consequently, apathy is an observable state that can subsequently be quantified, and a pathology of goal-directed behavior. Apathy is also indicative of cognitive decline and predicts a loss of autonomy in daily tasks [5, 7]. Disinhibition, on the other hand, encompasses a range of behaviors such as a patient's inability to suppress inappropriate responses, control behaviors, adapt their actions to environmental changes and suppress impulsions that violate social norms [6]. Ultimately, these BPSD increase caregiver burden [8]. Yet, diagnosing BPSD, such as apathy or disinhibition, is challenging, as it relies on patient introspection, through questionnaires including questions on their internal state, thoughts, feelings, and past activities in daily tasks [4, 5]. Yet, these patients commonly have anosognosia (5), rendering biases in these self-reported scales. Thus, a more reliable characterization of BPSD in AD and bvFTD is required.

In line with these considerations, we have developed the ECOCAPTURE research program (FRONTlab, ICM, 2014-2024) to objectively assess apathy [9, 10] and disinhibition [11] under ecological conditions, using video and sensors, in bvFTD and AD. In particular, the ECOCAPTURE@HOME protocol aims to bridge the gap between laboratory settings and natural settings. The first objective of the ECOCAPTURE@HOME study is to remotely measure behavioral markers of apathy, such as daytime activity, quality of sleep, emotional arousal, in everyday life conditions [12]. Based on previous studies [13-14], these markers seem relevant to assess apathy, and previous work using actigraphy has showed that sleep quality is a relevant indicator of apathy [15]. Other studies monitor stressors of BPSD such as disturbances in circadian rhythm and inadequate physical activity levels and use machine learning models to predict BPSD occurrences in individuals with dementia [16]. The second objective of the ECOCAPTURE@HOME study is to investigate the relationship between the psychological status of the patient-caregiver dyad and the patient's apathetic state, switching from the patient level to the dyadic level.

Experimental plan

Objectives

Rest-activity rhythms are one of the most prominent outputs of the circadian system [17]. Circadian disruption impacts sleep, health, and well-being. The assessment of the circadian rhythm and its degree of disruption can help in the clinical management of these disorders [17]. Numerous studies consider the accelerometer-measured rest–activity rhythm (RAR) amplitude as a clinical biomarker relying on the endogenous circadian system. In The Lancet Healthy Longevity, Feng and al. (2023) examined the associations between accelerometer-measured rest–activity rhythm amplitude and health risk in the general population [18]. Moreover, assessment of circadian rhythms in clinical practice can determine the misalignment of circadian rhythms with the external environment [17].

The overarching aim of this study is to define behavioral profile based on rest-activity rhythm and patterns, for each of the participant groups. In our study, we considered rest activity rhythm (RAR) and associated patterns and metrics (mean amplitude deviation, fast acceleration, mean electrodermal activity, heart rate) as behavioral markers of apathy in patients living with dementia at home, see Figure 1. **The primary objective** is to investigate accelerometer-measured rest–activity rhythm (RAR) amplitude and other sensor-based data to provide behavioral markers for monitoring apathy and disinhibition and differentiating patients from healthy controls. **The secondary objective** of this study is to investigate caregiver burden through questionnaire-based scores and ecological momentary assessment (EMA) and investigate relationship between caregiver's burden and quality of life (QoL) with patient's behavioral disorders. We focus on the caregiver because we consider the patient-caregiver dyad as an ecosystem and argue that the dyad carries the symptoms of the patient's pathology (such as apathy and disinhibition).

Methodology

The methods of this study allow for the observation of diverse patient populations exhibiting apathy and disinhibition symptoms, within their home and under ecological conditions. This method of observation is not only a tool to understand patient and caregiver behaviors in their own ecosystem but also allows for the analysis of daily dyadic interaction and dynamics.

The ECOCAPTURE@HOME study has started in 2022 and we plan to recruit a total of 60 dyads between 40 and 85 years old divided into three groups: bvFTD, AD, and healthy controls (HC). Today, we have included 16 dyads (5 bvFTD, 5 MA, 6 HC). We stratified the population in five participant groups including: bvFTD patients, AD patients, bvFTD-caregivers, AD-caregivers, and HC. We split the caregivers into two groups (bvFTD-caregivers, AD-caregivers) since there is some evidence indicating that being a caregiver for a patient with FTD is not the same as being a caregiver for a patient with AD. This project represents the opportunity to better characterize the caregiver burden in various pathological conditions, to identify commonalities and specificities.

We will collect data over 28 days per patient-caregiver dyad, using both passive measures (via sensors) and active measures (via questionnaires and EMA). Sensor-based data will be collected from both patient and their caregiver using Empatica wristband technology. The dyads will be expected to wear the sensor wristband on their non-dominant wrist, for 24 hours per day for 28 days. The Empatica wristband (Empatica Inc., Boston, MA, USA) is a wearable, wireless, multi-sensory signal acquisition device with four inbuilt sensors allowing real-time physiological data collection. It reports Galvanic Skin Response (i.e., electrodermal activity, EDA), peripheral Skin Temperature, and Tri-Axial Acceleration. The PPG (Photoplethysmography) data allows to determine blood volume pulse (BVP; sampling rate: 64 Hz) from which heart rate (HR), inter-beat intervals (IBIs) and heart rate variability (HRV) are derived. Participants wore the wristband on their distal forearm.

These sensor-based data measures will allow for the detection of rest-activity rhythms (RAR) of participants, see Figure 1. Rest-activity rhythm (RAR), namely magnitude, timing, and regularity of rest-activity patterns, is the most evident manifestation of the circadian rhythm and can be objectively quantified from accelerometry data. Accelerometry-derived metrics of RAR include amplitude (strength or magnitude of the rhythm) and acrophase (timing of peak activity).

In a first step, the raw data will be collected from the Empatica wristband sensors and will be available under three CSV files, one per sensor (ACC.csv, EDA.csv, HR.csv). Hence to visualize this data, see Figure 1, the raw signal will be segment in timeframes that indicate strong changes in the signal (e.g. rest indicates low levels of ACC and HR) and provide a set of ACTIVITY and REST sessions, see Table 1. This output will allow for the metric computation based on the signals, to investigate the RAR-based sessions, we are interested in. Our statistical plan will consist of performing statistical analyses on the following dependent variables of two types: sensor-based metrics, see Table 1, and questionnaire-based scores as well as EM data, see Table2.

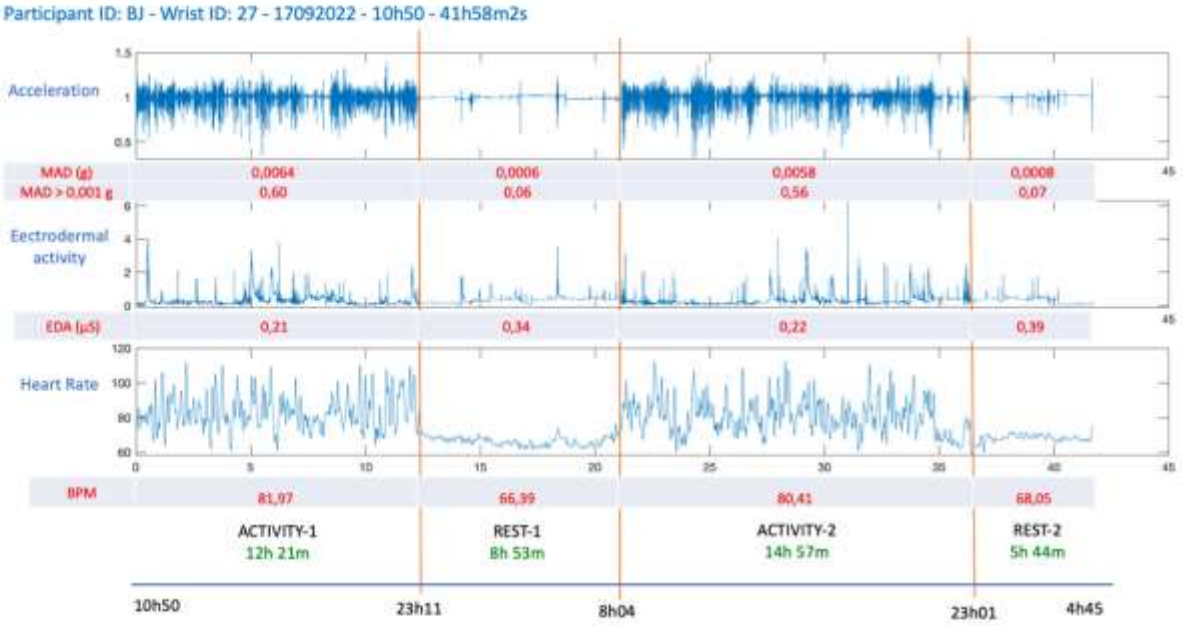


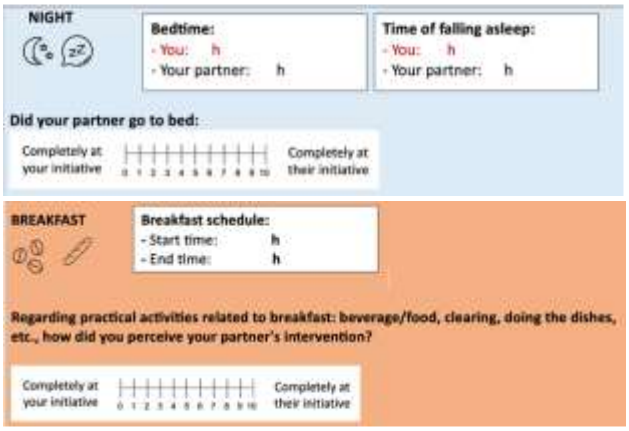
Figure 1. Graph of accelerometer-measured rest–activity rhythm amplitude. Raw data collected from the sensors over a 42-hour recording. The signals include Acceleration (ACC), Electrodermal Activity (EDA) and Heart Rate (HR). The x-axis represents time.

First, we setup two accelerometer-based metrics: 1/ ENMO: Euclidean norm minus one g of the raw accelerations, with resulting negative values rounded to zero and then averaged over 5 s epochs, 2/ MAD: mean amplitude deviation, Euclidean norm of each raw acceleration data point minus the mean of its correspondent 5 s epoch. Second, we setup one electrodermal activity-based metric: EDA: magnitude of electrodermal activity, and one blood volume pulse-based metric: HR: heart rate. Third, from these metrics, we computed dependent variables for each session (REST or ACTIVTY) or for the whole night. During an activity period, we computed: 1/ the mean of acceleration amplitude (nENMO), 2/ the mean of MAD (mMAD), 3/ the fast acceleration such as frequency of rapid acceleration calculated as MAD > 0,001 g (FAA), as well as the 4/ count activity (CA). During a rest period, we computed the mean of EA (EDA), and finally during the night, we computed the total time in bed (TTB), and the total sleep time (TST), see Table 1.

Table 1. Sensor-based variables computed with sensor-based metrics to describe patient’s behavior.

Sensor-based metrics and variables			
Session	Variable	Used Metric	Description
ACTIVITY	**mENMO**	ENMO	Mean of acceleration amplitude during an activity period
ACTIVITY	**mMAD**	MAD	Mean of MAD during an activity period
ACTIVITY	**CA**	Using a band-pass frequency filter to the raw signal	Count activity during an activity period
ACTIVITY	**FAA**	MAD > 0,001 g	Fast acceleration such as frequency of rapid acceleration during an activity period (MAD > 0,001 g).
NIGHT	**TTB**	ENMO, HR	total time in bed
NIGHT	**TST**	ENMO, HR, EDA	total sleep time
REST	**mEDA**	EDA	Mean of EDA during a rest period

Complementary to this, we will use ecological momentary assessment (EMA) and ethogram-based questionnaires for collecting qualitative data that the caregiver will fill out twice a week. These EMA and questionnaires aim to 1/ shed light and contextualize the passive data collected to better understand the level of apathy in patients as well as the dyad’s daily routine, and 2/ describe the caregiver’s feelings. The report on the chosen day begins with the evening before bedtime and ideally should be done throughout the day and at each moment when the caregiver’s feedback is requested: 1/ Bedtime, 2/ Wake-up, 3/ Breakfast, 4/ Lunch, 5/ Dinner, 6/ Snacking, 7/ Medications, 8/ Leisure, 9/ Outing, 10/ Social, 11/ Naps, 12/ Overall day. The questionnaire and EMA investigate: 1/patient’s behavior and their level of apathy as well as disinhibition at precise moments throughout the day (e.g. waking up, meals, social interactions, bedtime, etc.), see Figure 1, and 2/ the level of caregiver’s burden experienced and quality of life, see Figure 2.



NIGHT

Bedtime:
- You: h
- Your partner: h

Time of falling asleep:
- You: h
- Your partner: h

Did your partner go to bed:

Completely at your initiative 0 1 2 3 4 5 6 7 8 9 10 Completely at their initiative

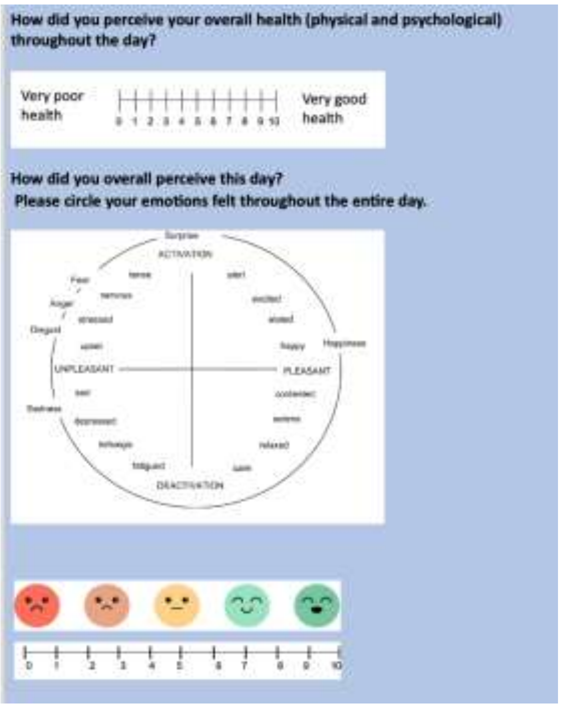
BREAKFAST

Breakfast schedule:
- Start time: h
- End time: h

Regarding practical activities related to breakfast: beverage/food, clearing, doing the dishes, etc., how did you perceive your partner’s intervention?

Completely at your initiative 0 1 2 3 4 5 6 7 8 9 10 Completely at their initiative

Figure 2. Questionnaire and EMA. How to describe the patient’s behavior.



How did you perceive your overall health (physical and psychological) throughout the day?

Very poor health 0 1 2 3 4 5 6 7 8 9 10 Very good health

How did you overall perceive this day?
Please circle your emotions felt throughout the entire day.

Surprise
ACTIVATION
tense
alert
Fear
nervous
excited
Anger
stressed
elated
Disgust
upset
happy
Happiness
UNPLEASANT
PLEASANT
sad
contented
Sadness
depressed
serene
lethargic
relaxed
fatigued
calm
DEACTIVATION
0
1
2
3
4
5
6
7
8
9
10

Figure 3. Questionnaire and EMA. How to describe caregiver' burden, feelings, and quality of life.

From these questionnaires and EMA-based data, we computed dependent variables (scores) for the patient: 1/ APATHY: apathy score, 2/ DIS: disinhibition score; and for the caregiver: 1/ QoL: Quality of life score, 2/ BURD: Caregiver's burden score, and 3/ FEEL: feeling score, see Table 2.

Table 2. Questionnaires and EMA-based score describe caregiver' burden, feelings, and quality of life.

Questionnaire and EMA-based scores		
PATIENT	**APATHY**	Apathy score
PATIENT	**DIS**	Disinhibition score
CAREGIVER	**QoL**	Quality of life score for the caregiver
CAREGIVER	**BURD**	Caregiver's burden score
CAREGIVER	**FEEL**	Feeling score

Analysis and expected results

Our statistical plan will consist of performing statistical analyses on the sensor-based metric and questionnaire and EMA-based scores, in each group. All statistical analyses on demographical, sensor-based metrics and questionnaire-EMA-based scores will be performed using Python software, where p-values <0.05 will be considered statistically significant. Data normality will be tested with the Shapiro-Wilk test, and Fischer's test will be used to check for variance equality. We will compare the mean questionnaire-EMA scores and sensor-based metrics between the five groups (AD, bvFTD, AD-caregivers, bvFTD-caregivers and HC). Depending on the normality of the distributed data, we will either use a student t-test for normally distributed data or the Mann-Whitney U test for non-normally distributed data. Based on the sensor and questionnaire data, this will help us determine whether there is a group effect. Next, we will explore the relationship between the sleep-wake cycles of patients (AD, bvFTD) and the quality of life (QoL) along with the experienced burden in caregiver groups. To achieve this, we will use Spearman's correlation to investigate the possible relationships between the sensor-based metrics and the questionnaire-based scores. Additionally, as a multiple comparison correction test, we will use the Holm-Bonferroni method.

We expect to show discrimination between the FTD, AD, AD-caregivers, FTD-caregivers, and HC based on the sensor-based metrics. By determining this discrimination, we hope to confirm differences in behavioral patterns relative to the RAR. Furthermore, we expect to observe a discrimination in the apathy disinhibition scores between the AD, FTD, and HC groups. Finally, we hope to show a relationship between the sensor-based metrics in patient groups (AD, FTD) with the QoL, feelings and burden in caregiver groups (AD-caregiver, FTD-caregivers).

Conclusion

Most laboratory studies typically examine dementia in mild stages and their methodology is divorced from everyday scenarios. On the other hand, at the hospital, neurologists evaluate and treat very severely impaired patients. Our study is an opportunity to monitor patient in middle stages of dementia – in opposite to lab and hospital settings – and follow evolution and precisely manage each patient's case. Our study could improve characterization to understand BPSD at the individual patient level. The realization of this study requires the identification of a comprehensive list of valuable information about the patient. As of today, there are no effective treatments for dementia for apathy nor disinhibition. Implementing treatment strategies at home to address these behavioral disturbances would be helpful to improve the QoL of patients and their caregivers. Ultimately, ECOCAPTURE@HOME will also have long-term implications with the setup of a healthcare data platform accessible for the global care community: patient's family, their care provider, the aged care facilities, various practitioners, and nurses. Online platform where BPSD information is housed to give access to families and each member of the care community, could help for monitoring and detecting changes across the disease trajectory.

Ethical statement

The ECOCAPTURE@HOME protocol, sponsored by Inserm, was granted approval by the "Comité de Protection des Personnes Ile-de-France VII" on 30/04/2021, and registered in a public registry (clinicaltrials.gov: NCT04865172). All study participants are informed about the research and provide consent to participate, in line with French legal guidelines.

References

1. Organization, World Health. WHO. WHO Dementia. <https://www.who.int/news-room/fact-sheets/detail/dementia>. Accessed 7 December 2023.
2. Piguet, O., Hornberger, M., Mioshi, E., Hodges J. R. (2011). Behavioural-variant frontotemporal dementia: Diagnosis, clinical staging, and management. *The Lancet Neurology* **10**, 162-172.
3. Cerejeira, J., Lagarto, L., Mukaetova-Ladinska, E. (2012), Behavioral and Psychological Symptoms of Dementia. *Frontiers in Neurology* **7**, 73.
4. Collins, J.R., Henley, S.M.D., Suarez-Gonzalez A. (2023). A systematic review of the prevalence of depression, anxiety, and apathy in frontotemporal dementia, atypical and young-onset Alzheimer's disease, and inherited dementia. *Int Psychogeriatr* **35(9)**, 457-76.
5. Levy, R., Dubois, B. (2006). Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits. *Cereb Cortex* **16(7)**, 916-28.
6. Migliaccio, R., Tanguy, D., Bouzigues, A., Sezer, I., Dubois, B., Le Ber, I., et al. (2020). Cognitive and behavioural inhibition deficits in neurodegenerative dementias. *Cortex* **131**, 265-83.
7. Cai, X., Zhao, H., Li, Z., Ding, Y., Huang, Y. (2022). Detecting apathy in patients with cerebral small vessel disease. *Front Aging Neurosci* **14**, 933-958.
8. Van Den Bossche, M.J.A., Van Vré, A.T.E., Van den Bulcke, L. et al. (2024). Clinical staging of behavioral and psychological symptoms of dementia. *Nat. Mental Health* **2**, 3–5.
9. Batrancourt, B., Lecouturier, K., Ferrand-Verdejo, J., Guillemot, V., Azuar, C., Bendetowicz, D., Migliaccio, R., Rametti-Lacroux, A., Dubois, B., Levy, R. (2019). Exploration Deficits Under Ecological Conditions as a Marker of Apathy in Frontotemporal Dementia. *Front Neurol* **10**, 941.
10. Godefroy, V., Batrancourt, B., Charron, S., Bouzigues, A., Sezer, I., Bendetowicz, D., Carle, G., Rametti-Lacroux, A., Bombois, S., Cognat, E., Migliaccio, R., Levy, R. (2022). Disentangling Clinical Profiles of Apathy in Behavioral Variant Frontotemporal Dementia. *J Alzheimers Dis.*
11. Tanguy, D., Batrancourt, B., Estudillo-Romero, A., Baxter, J.S.H., Le Ber, I., Bouzigues, A., Godefroy, V., Funkiewiez, A., Chamayou, C., Volle, E., Saracino, D., Rametti-Lacroux, A., Morandi, X., Jannin, P., Levy,

R., Migliaccio, R., ECOCAPTURE study group. (2022). An ecological approach to identify distinct neural correlates of disinhibition in frontotemporal dementia. *Neuroimage Clin* **35**, 103079.

12. Godefroy, V., Levy, R., Bouzigues, A., Rametti-Lacroux, A., Migliaccio, R., Batrancourt, B. (2021). ECOCAPTURE@HOME: Protocol for the Remote Assessment of Apathy and Its Everyday-Life Consequences. *Int. J. Environ. Res. Public. Health* (18), 7824.
13. König, A., Aalten, P., Verhey, F., Bensadoun, G., Petit, P.-D., Robert, P., David, R. (2014). A Review of Current Information and Communication Technologies: Can They Be Used to Assess Apathy?: Current and New Methods for the Assessment of Apathy. *Int. J. Geriatr. Psychiatry* **29**, 345-358.
14. Dorsey, E.R., Glidden, A.M., Holloway, M.R., Birbeck, G.L., Schwamm, L.H. (2018). Teleneurology and Mobile Technologies: The Future of Neurological Care. *Nat. Rev. Neurol*. **14**, 285–297.
15. Zeitzer, J.M., David, R., Friedman, L., Mulin, E., Garcia, R., Wang, J., Yesavage, J.A., Robert, P.H., Shannon, W. (2013). Phenotyping Apathy in Individuals With Alzheimer Disease Using Functional Principal Component Analysis. *Am. J. Geriatr. Psychiatry* **21**, 391-397.
16. Cho, E., Kim, S., Heo, S.J., Shin, J., Hwang, S., Kwon, E., et al. (2023). Machine learning-based predictive models for the occurrence of behavioral and psychological symptoms of dementia: model development and validation. *Sci Rep* **13(1)**,8073.
17. Reid, K.J. Assessment of Circadian Rhythms. (2019). *Neurol Clin* **37(3)**,505-526.
18. Feng, H., Yang, L., Ai, S., et al. (2023). Association between accelerometer-measured amplitude of rest–activity rhythm and future health risk: a prospective cohort study of the UK Biobank. *Lancet Healthy Longev* **4**, e200–210.

Effectiveness of zoom equipped drones for use in reading livestock ear tags for animal identification

John S. Church[1], Mathis Gegout[2], Paul J. Adams[3]

1 Department of Natural Resource Sciences, Thompson Rivers University, Kamloops, BC, V2C 0K8,

2 InstitutAgro Dijon, Dijon, 26 bd Dr Petitjean 21079, France

3 Applied Genomics Centre, Kwantlen Polytechnique University, Surrey, BC, V3W 2M8, Canada

jchurch@tru.ca

Introduction

The popularity and prevalence of breeds such as Angus over the last fifty years has resulted in homogeneous beef cattle herds across North America, which makes identification and subsequent observation of individual animals challenging. Many producers have already started adopting commercially available drones, from companies such as industry leader DJI, to find, monitor and even move animals while on pasture (Herlin et al. 2021). The development of relatively inexpensive multi-rotor drones with high quality zoom cameras has generated interest by farmers and ranchers in using these aerial systems for the purpose of locating, monitoring and identifying individual cattle while on pasture (Alanezi et al. 2022). In the last five years, a growing number of drones have become commercially available with integrated zoom cameras that are potentially capable of reading livestock ear tags, enabling identification of individual animals in the field. However, it is unclear what level of zoom capability is required, or under what conditions and distances the drones are effective at reading livestock ear tags. The purpose of this research was to determine which drone models are currently the most effective for detecting/reading various livestock ear tags to enable individual animal identification in the field.

Materials and Methods

To test ear tag identification from drones, and to ensure the livestock tags were hung as realistically as possible, we simulated a cow with ear tags using a short face bovine (Herford) veterinary head model (Figure 1. a.). This highly realistic head model was primarily designed for captive bolt training (Veterinary Simulator Industries; Calgary, AB Canada), and will henceforth be called the target. Cattle ear tags are made of synthetic material (nylon) and use high-visibility ink marking and/or laser engraving for readability of individual numbers for animal identification purposes. Tags come in a variety of colours and sizes depending on the age (i.e., calf vs. cow) and species (i.e., cattle vs. sheep), are ~ 10-11.5 cm high and 6.3 to 7.5 cm wide and typically held in the ear with a single reinforced pin. All tags tested were hung at 1.57 meters off the ground to the center ear tag pin with the head placement of the model measured with a protractor at precisely 90o. To accurately record the distance of the drone to the target, a Dewalt fiberglass long tape measure (100m) was used. The drone models selected for testing were all chosen because they provided some degree of zoom capability. As DJI currently dominates more than 70% of the global drone market, only DJI drones were tested.

The drone models tested, which were all released within the last 5 years, included: DJI Mavic 3T (56x hybrid zoom, released September, 2022); DJI Matrice 30 Series (200x zoom, 16x optical, released September, 2022); DJI Mini Pro (2x digital zoom, released May, 2022); DJI Mavic 2 Enterprise Advanced (32x digital zoom, released April, 2021); DJI Matrice 300 RTK with Zenmuse H20T Quad Sensor (200x zoom, 23x optical, released May, 2020); DJI Mavic 2 Enterprise Universal (2x optical, 3x digital zoom released August, 2018). All drone models were tested with the included controller with built-in screen (integrated), except for the Mavic 2 Enterprise Universal which was tested with a DJI CrystalSky monitor attached to the controller, as an integrated controller was not available at the time of purchase. No cell phones were used in lieu of an integrated drone controller screen, as cell phone screens are known to be notoriously difficult to read in bright sunlight. To improve the GPS positional accuracy of the DJI Matrice 300 RTK, the DNSS base station (DJI D-RTK 2) was also used. In addition, the laser range finder on two of the drone cameras (Zenmuse H20T onboard both the DJI Matrice 300 RTK and DJI Matrice 30 quadcopters) was also used to accurately measure drone distances to the target when the distance to the target exceeded 100 meters. The accuracy of the laser range finder on the drone was later verified by the fiberglass long tape measure.

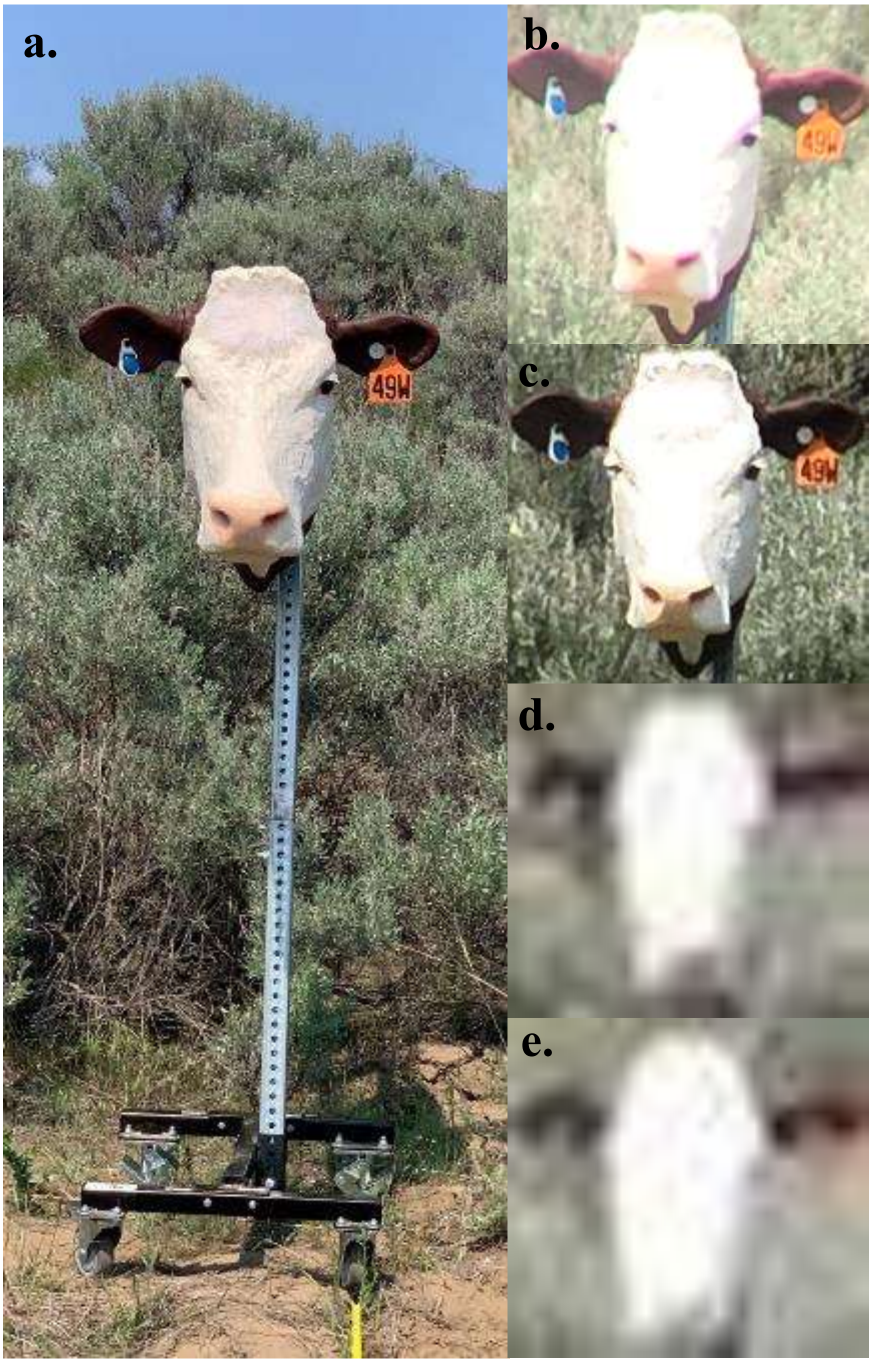


Figure 1. a. An anatomically correct soft silicone cattle head with flexible (and detachable) ears, provided by Veterinary Simulator Industries, Calgary, Alberta Canada, was used to mount the ear tags and served as the target to ensure realistic tag placement/hanging simulation; b. Bushnell H20 8x42 binoculars at 60m; c. Mavic Enterprise 3T 56x hybrid zoom at 60m; d. Mavic Enterprise Advanced 32 digital zoom at 60m; e. Mavic 2 Enterprise Zoom 2x optical 3x digital zoom at 60m

Horizontal maximum distance tests

All drones tested were initially launched from 60 meters horizontal to the target at approximately 2 m AGL as measured by the tape, except for the Matrice 300 and 30 with the more capable Zenmuse H20T zoom cameras. For safety reasons, the Matrice 300 and Matrice 30 drone models were flown backwards from 60 meters from the target at a safe height of 40 meters above ground level (AGL) until the drone reached 250 meters and 180 meters from the target as measured by the onboard integrated laser range finder, which was used to determine the limit in which the ear tag was no longer discernable by the observers on the controller screen. The distances measured by the integrated laser range finder on the M300 and M30 were thought to be at least as accurate as the tape measure as the laser was placed directly onto the ear tag via the controller. The accuracy of the laser range finder was later confirmed at a height of 2 meters AGL at 60 meters. This was performed with the fiberglass tape used to measure the distance to the target, and only the tape was used to measure the distance for the remaining Mavic and Mini drones that do not have an integrated laser range finder. The initial sixty-meter distance to the target was chosen to test the remaining drones as this was regarded as the likely upper limit for which the ear tags would be discernable, and it is the approximate distance that livestock ear tags can be read with conventional binoculars (Bushnell H20 8x42-mm) that producers often employ to read ear tags. When comparing the different ear tag size types, two distances to the head model target were selected, 60m as well as 40m, as we determined it was too challenging to read the smaller ear tags at 60m. These were the optimal distances that were determined easiest to read for both large and small livestock tags respectively, so that we could concentrate our efforts on the second test on the effects of different tag colors.

The different drone models were each launched and hovered above the extended tape at a flying height of ~2-2.5 meters which is the height that binoculars are used in conventional tag surveying; and flown towards the head model target with the camera facing forward on maximum zoom orientated on the ear tag target. Two independent observers (drone operators who were blind to the tag number) independently had to agree when the various assorted ear tags were readable on the integrated controller screen. In addition, a photo was taken by the camera on the drone to perform a post hoc accuracy assessment from the stored images to confirm accuracy. Once the tag was readable, the camera on board the drones was moved into nadir position to read the tape measure to determine the distance of the drone to the target.

Ear tag color tests

A variety of different ear tag colors (white, purple, yellow, orange, red, light green, light blue) and sizes (large for cattle, n=14; and small for sheep, n=12) were tested twice, with primarily black text factory printed on the ear tags, except for a couple of ear tags where the number was manually printed with a livestock marker. Color testing was done during midday using only the Mavic 3T drone, based on superior performance during the range test, between 10:00-13:00 hours; with the sun in front of the tag, shining directly into the tag, as opposed to the sun being orientated back behind the tags, which we determined to be more optimal to test the effects of sunlight reflection on tag readability. In all cases, once the initial observations were completed on the tags by the observers, the tags were randomized, and the observations were repeated on all tags a second time.

Altitude tests

Again, based on superior performance during the range test only the Mavic 3T drone was selected for a third test on the effect of altitude on the ability to read the large orange tag which is relatively easy to see from a horizontal distance of 60 meters to the target (the distance of the conventional binoculars) even though the maximum limit at ~2 meters AGL was determined to be 90 meters horizontally at both 28x and 56x zoom. The superior zoom capabilities of the Matrice 30 and 300RTK drones and the inferiority of both the Mavic 2 and Mini 3 Pro drone models as determined during the field limit range test precluded the need for testing these drones.

We started at 10 m AGL from 60 meters horizontally away from the target. We tested the limits of both 28x and 56x zoom on the Mavic 3T by increasing the height vertically by 10 m increments until we reached 120 m, the maximum AGL allowed by Transport Canada. In addition, using the protractor as our reference, we further tested the angle of the cow head model at 45, 90, and 135 degrees to determine the impact of head angle relative to the ground on the ability of the Mavic 3T drone to read the ear tag at 56x zoom at an AGL of 60 meters.

Results

For the field limit range test, the DJI Mavic 3, 3T and the Matrice 30 and 300 RTK all could discern the ear tag at between 40-60 meters. The field limits for the Mavic 3 and 3T were 70 and 90 m respectively, while the DJI Matrice 30 and 300 RTK limits were substantially greater at 180 and 250 meters. Neither of the two Mavic 2 Enterprise drones (Advanced or Universal) or the Mini 3 Pro drone enabled human viewers to read the tags at between 40-60 meters, and all were under 25 meters (15, 23 and 5 meters).

Testing the readability of the different colored tags with the DJI Mavic 3T drone, at 40 meters all tags were readable regardless of color or size. At 60 meters orange, purple, white and light blue colored tags were all observable regardless of size, but the small red and yellow tags were not.

The results from the altitude tests using the Mavic 3T using the two zoom settings indicated that at 60 meters horizontal to the target, we could ead the tag at 30 meters AGL at 28x zoom, but when we switched to 56x zoom, cattle ear tags were discernable at 50 meters AGL. The results between the vertical range test and the altitude test using the Mavic 3T drone were identical. Raising the cattle head beyond 90o (i.e., 110°) made the tag easier to read, and enabled us to increase the AGL distance by approximately 20 meters, such that we could still read the tag at an AGL of 70 meters. Additionally, all tag photos collected with the drone were readily discernable and agreed with the live observations.

Discussion

Predicting how the advertised zoom capabilities of the various commercially available drone models translates to their ability to resolve numbers on bovine ear tags can be complicated. In terms of accuracy and distance, the zoom capabilities as advertised on drone manufacturers' promotional material do not necessarily reflect performance as measured in our first two tests for reading ear tags. For example, one of the drones advertised as 32x zoom (Mavic 2 Enterprise Advanced) was incapable of reading a target ear tag at a modest 20 meters distance to the target. Traditionally, there have been two basic ways of zooming in photography, optical zoom and digital zoom, while an additional third way has recently emerged coined "hybrid" zoom (Fang et al. 2020). In addition, every camera sensor also has a different resolution and a different physical sensor size. Optical zoom relies on a physical camera lens movement, which changes the apparent closeness of an image subject by increasing the focal length (Lenk et al. 2019). A lens is then placed in front of the sensor, which is either a fixed focal length telephoto lens or a true optical zoom with variable focal length (Barreto et al. 2021). To zoom in, the glass lens moves further away from the image sensor, and the scene is magnified (Lenk et al. 2019). Optical zoom, first found on the DJI Mavic 2 Zoom offers the best results, with the possibility of zooming up to 2 times optically (Barreto et al. 2021), is the truest form of magnification. The target subject is enlarged by manipulating rays of light coming from the scene and offers lossless results (Fang et al. 2020), offering similar results to moving closer to your target. Factors that reduce the image results include glass quality, as well as the aperture, which can also be affected as you increase the focal length depending on the lens (Fang et al. 2020). In the case of the Matrice 300RTK equipped with an DJI H20 or H20T camera, the cameras provide for a true optical zoom, with a focal length from 6.83 mm to 119.94 mm. The recently released DJI Matrice 30 similarly provides for a true optical zoom with a focal length from 21mm to 75 mm, but they are far more expensive options. The Mavic 2 Enterprise Universal also provided a modest 2x optical zoom, which enabled the ear tag to be discernable at 20 m, outperforming the Mavic 2 Enterprise Advanced with 32x digital zoom. Optical zoom continues to be the best solution if drones can't physically get closer to the subject (Barreto et al. 2021). In contrast, digital zoom uses digital software magnification technology only to enlarge an area of an image (compromising the integrity of the picture by enlarging the actual pixels in the center of the photo and cropping out the rest), which gives the appearance of magnifying the subject, which also consequently reduces both image resolution and quality (Barreto et al. 2021). Unlike optical zoom, digital zoom is not lossless, meaning some information from the scene is discarded by the software in the process, that is why when using digital zoom, the ear tags often looked blurry or smudgy and the numbers were more difficult to read (Fang et al. 2020). As a result, we quickly determined that the zoom capability provided via drones with digital zoom is vastly inferior to zoom capabilities provided by drones utilizing some form of true optical zoom.

An important advancement that has emerged in the last couple of years for drones is hybrid zoom. Hybrid zoom, originally developed for use in cellular phones, does not perform at the true optical camera zoom level, but generally outperforms basic digital zoom for preserving fine details at a distance (Fang et al. 2020). Thanks to the development of computational photography, hybrid zoom uses a combination of optical zoom, digital zoom, and some form of AI-based software enhancement to dramatically improve results when zooming, which can far exceed the lens's physical capabilities within the drone camera (Fang et al. 2020). Even with identical hardware, computational photography methods can yield dramatically different results. That is why the Mavic 3T at either 28x zoom or 56x zoom (28x zoom is available on the standard Mavic 3, while 56x zoom is available on the Mavic 3 Enterprise series) is vastly superior to the Mavic 2 Enterprise Advanced drone released just a couple of years earlier that is 32x zoom, but digital only. We did not test a Mavic 3 drone per se, but we were informed by our drone supplier that there is no difference between the 28x zoom on a Mavic 3T vs a standard Mavic 3, so we tested our Mavic 3T at two zoom settings, 28x and 56x times zoom. While the numbers would imply that the zoom capability between the two zoom settings should be twice as much (i.e., 28x vs 56x zoom), in our real world tests the resolution differences were incremental. At 28x zoom, cattle ear tags were still discernable at 70 meters to the target and increasing the zoom to 56x resulted in the cattle ear tags still being discernable at 90 meters, a difference of only twenty meters. Using the Mavic 3T, yellow and light green tags were more challenging to read in general compared to white or orange tags, likely due to the impact of direct sunlight, while the black text on the red tag was likely hardest to discern due to a lack of contrast. Otherwise, the ability to read the rest of the tags (white, orange or light blue) was similar. The same twenty-meter differences observed between the two zoom settings on the Mavic 3T drone during our vertical range test was nearly identical to the results in our vertical altitude tests, but changing the head angle improved the ability to read the tag. The camera tilt angle proved to be efficient in monitoring other species like small cetaceans (Barreto et al. 2021). These altitudes are very important as they enable the Mavic 3 drones to fly above most tree canopies when attempting to identify individual animals. All of the other drone models are incapable of discerning ear tags at 25 meters or less, which would make their use in environments with trees impractical. During these same brief altitude tests, we determined that the head angle of the cattle head model had a pronounced effect on our ability to read the cattle ear tag. The head orientation needed to be at least ninety degrees relative to the ground to read the ear tag. Deviation below 90o is typical when cattle are grazing and results in the tag being obscured within the cavity of the ear. Interestingly, if the cattle head was raised beyond 90o (i.e., 110°), it was easier to read the tag, and enabled us to increase the AGL distance by approximately 20 meters, such that we could still read the tag at an AGL of 70 meters. The AGL distance is important when using drones to monitor livestock because it potentially reduces the observer effect (the animal being aware of the drone) as reported in Mulero-Pázmány et al.

In conclusion, while optical zoom is still the gold standard, the hybrid zoom capabilities of new Mavic drone models which have been released over the last couple of years (i.e. DJI Mavic 3, Mavic 3T etc.) are far superior to older drone models that utilize digital zoom only. As such, while the Matrice series drones are still unparalleled when reading ear tags, the new more affordable Mavic models using hybrid zoom would still be useful tools for livestock producers wishing to use drones to identify and monitor individual animals housed under extensive pasture conditions solely by their livestock identification ear tags. Besides the obvious advantage in cost, size differences between the Matrice and Mavic drones might be a further advantage as larger drones tend to disturb animals more than smaller drones (Mulero-Pázmány et al. 2017; Herlin et al. 2021). While we have flown as low as 10 m over cattle using Mavic drones in the past with few behavioral responses (Mufford et al. 2021), the benefits of using smaller Mavic models as opposed to larger drones on the observer effect should be investigated in the future.

References

1. Alanezi, M.A., Shahriar, M.S., Hasan, Md.B., Ahmed, S., Sha'aban, Y.A., Bouchekara, H.R.E.H. (2022). Livestock Management With Unmanned Aerial Vehicles: A Review. *IEEE Access* 10, 45001–45028. https://doi.org/10.1109/ACCESS.2022.3168295 >. Accessed 25 January 2024

2. Barreto, J., Cajaiba, L., Teixeira, J.B., Nascimento, L., Giacomo, A., Barcelos, N., Fettermann, T., Martins, A. (2021). Drone-monitoring: improving the detectability of threatened marine megafuana. *Drones* 5, 14. https://doi.org/10.3390/drones5010014 >. Accessed 30 January 2024

3. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z. (2020). Perceptual Quality Assessment of Smartphone Photography. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3677–3686.

4. Herlin, A., Brunberg, E., Hultgren, J., Högberg, N., Rydberg, A., Skarin, A. (2021). Animal Welfare Implications of Digital Tools for Monitoring and Management of Cattle and Sheep on Pasture. *Animals* 11, 829. https://doi.org/10.3390/ani11030829 >. Accessed 25 January 2024

5. Lenk, L., Mitschunas, B., Sinzinger, S. (2019). Zoom systems with tuneable lenses and linear lens movements. *Journal of the European Optical Society-Rapid Publications* 15, 9. https://doi.org/10.1186/s41476-019-0106-3 >. Accessed 28 January 2024

6. Mufford, J.T., Reudink, M.W., Rakowbowchuk, M., Carlyle, C.N., Church, J.S. (2021). Using unmanned aerial vehicles to record behavioral and physiological indicators of heat stress in cattle on feedlot and pasture. *Canadian Journal of Animal Science* 102, 1, https://doi.org/10.1139/cjas-2020-012 >. Accessed 30 January 2024

7. Mulero-Pázmány, M., Jenni-Eiermann, S., Strebel, N., Sattler, T., Negro, J.J., and Tablado, Z. (2017). Unmanned aircraft systems as a new source of disturbance for wildlife: A systematic review. *PLoS One* 12(6): e0178448. Public Library of Science. https://doi:10.1371/journal.pone.0178448. >. Accessed 28 January 2024

Towards deep physiology in home cage: Sensor cages and AI for multi-animal whole-body health testing

Michael. Florea[1,2*], Noah Weber[1*], Michael Kaca[1], Pratomo Alimsijah[1]

1 Olden Labs, PBC, South San Francisco California USA

2. Harvard Stem Cell and Regenerative Biology, Harvard University, Cambridge USA

michael@oldenlabs.com, noah@oldenlabs.com

Abstract

Traditional behavior and physiology testing is laborious and stressful. Current home-cage-based systems are limited to one animal or a few metrics. To achieve multi-modal health testing in home cages of group-housed mice, we engineered a smart cage (DOME), that integrates video tracking, sensors and AI to estimate 15 metrics of health. Prototype DOMEs enable precise tracking of individual animals and identification of age-related changes in at least 8 metrics, demonstrating feasibility of multi-modal home-cage testing.

Introduction

Animals often exhibit complex physiological responses to a given drug or treatment, but most of this data is not collected due to the high cost of comprehensive physiological testing (*1*, *2*). The current paradigm in animal health and behavior testing is to maintain animals in their home cages, move them to test setus, and conduct tests that are often stressful and assay animal’s capacity at their limit. However, this is manual, low-throughput and induces stress due to handling and context switching which can compromise results. On the other hand, options for home-cage monitoring now exist (*3*), but these systems provide limited data on health and behavior, or require single-housing of animals, and thus cannot replace specialized test setups.

We argue that solving these problems requires a fundamental paradigm shift in the field – a goal of measuring most physiological metrics directly in the home cage during the natural behavior of animals, rather than in a test setting at the limits of their abilities. Achieving this will require creative redesign of assays and the technical capacity to a) individually track animals in a group-housed setting, b) gather multi-modal data about their behavior and health, c) conduct perturbations in home cage while d) being compatible with biweekly cage exchanges, microisolation and other necessities of daily animal care.

Methods

To build a system that meets the above requirements, we have engineered DOME cages (Digital Online Monitoring Equipment). The DOME system consists of three parts: 1) new home cages containing multiple sensors and actuators, 2) a new animal ID method and 3) and new animal tracking AI. The sensor home cage contains hardware capable of capturing video, sound, weight, bite strength, pull strength and temperature data and produce light and audio stimuli, while being compatible disposable plastics, biweekly cage exchanges and sterilization needs. The animal identification method uses clearly visible single or double digits/letter tags, or striped tail markings optimized for computer vision detection. The AI consists of a new deep convolutional neural net (DCNN), optimized for identification of animals and their tags from low-resolution video data in a cost-effective manner.

To arrive at multi-modal health metrics for individual animals, the DOME system’s DCNN first analyses video data to estimate positions of each animal in the cage. This data is then used to derive four sets of metrics. First, positional information is syncrhonized to video-based behavioral metrics that primarily examine location (such as activity, rearing, eating, drinking, climbing, etc.) to identify the first set of metrics. Secondly, positional information is syncronized to sensor data to deconvolute sensor-based metrics (weight, temperature, bite and pull strength). Third, experimenter-programmed perturbations (such as generation of light or sound pulses, or placement of run wheel or novel objects in the cage) are synchronized to location-based metrics to assay animal response to perturbation (hearing and vision, voluntary run wheel activity and novel object recognition/displacement tests). Finally, data from the first three metrics, along with audio recordings, are used to estimate complex behaviors (fighting, mating, grooming, nesting). Overall, DOME cages are designed to capture 31 behavioral/physiological metrics.

While the development of the DOME system is on-going, we have technically established measurements of 15 metrics and have tested the first set of metrics (8 location-based metrics) with young, middle-aged and aged mice (3 month old, 11 month old and 26 month old mice) over a 48h recording period. All mice were male C57BL/6J with ad libitum access to food and water, housed in a 12h light and dark cycle. Animal experiments were reviewed and authorized by the ethical review committee at Olden Labs, PBC. No animals were subjected to invasive or painful procedures during testing.

Results

The DOME cage (Figure 1A) achieves individual animal identification in group-housed mice, along with head and tail position tracking (Figure 1B). The current identification accuracy of the DCNN is 99.9% (estimated as the number of frames where the animal was correctly identified by the DCNN compared to human-labelled ground truth data). Tracking is achieved at high efficiency, with a computational cost of ~1h GPU time for 24h analysis video (not shown). Estimation of 8 position and activity-based metrics revealed a sharp decline in most metrics for aged mice compared to young, and many metrics between young and middle-aged (Fig. 1C). We anecdotally also observe low variation between individual mice of the same cage, which may be due to the fact that mice tend to engage in group behaviors, synchronizing their activities (*4*).

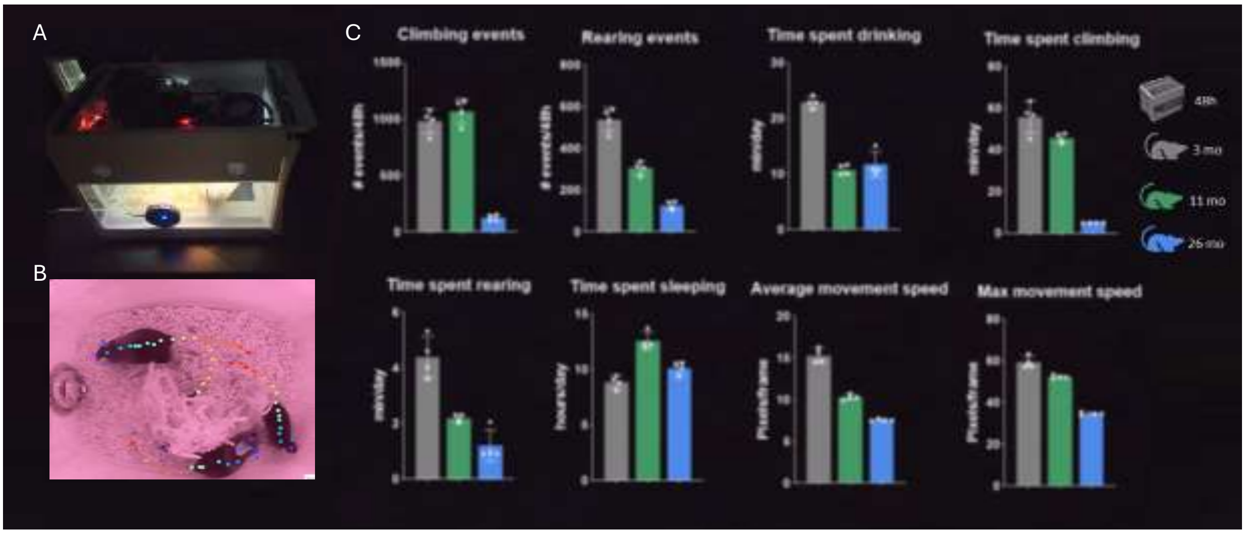


Figure 1. DOME cages for deep home-cage physiology. (A) DOME cage hardware. (B) Identification of individual mice and head and tail positions in a group-housed setting by the deep convolutional neural network. (C) Eight location-based health metrics estimated from young, middle-aged and aged mice.

Conclusions

Here we present pilot data on DOME cages, engineered to capture multi-dimensional physiological data directly from the home cage. Through the DOME system, we aim to provide the community with a tool to gather richer behavioral and physiological data while reducing costs and improving animal welfare through reduced stress.

References

1. R. Hoehndorf *et al.*, Mouse model phenotypes provide information about human drug targets. *Bioinformatics* **30**, 719-725 (2013).
2. S. D. M. Brown *et al.*, High-throughput mouse phenomics for characterizing mammalian gene function. *Nature Reviews Genetics* **19**, 357-370 (2018).
3. P. Kahnau *et al.*, A systematic review of the development and application of home cage monitoring in laboratory mice and rats. *BMC Biology* **21**, 256 (2023).
4. M. I. Sotelo *et al.*, Neurophysiological and behavioral synchronization in group-living and sleeping mice. *Current Biology* **34**, 132-146.e135 (2024).

Multi-modal measurements

The constructive effect of positive encouragement on preschool children

Wenhao Lv[1], Qiongfang Cao[1], Haiqi Xiang[1], Fangfang Liu[2], Xi Yang[3], Fan Xu[1]

1 Department of Public Health, Chengdu Medical College, Sichuan, 610500, China

2 Art College, Southwest Minzu University, Chengdu, Sichuan, 610041, China

3 Department of psychology, Chengdu Medical College, Sichuan, 610500, China

Correspondent author: Fan Xu, email: xufan@cmc.edu.cn

Abstract

The invisible positive encouragement from parents on the preschool children exist extensively, however the neural modulation mechanism behind it remains unclear. Here we recruited 9 children, mean age 5 years old, to stack the block, while randomly invited the parent to be positively encourage on child or not, during the stack process. During this process, we captured the face expressions and life signs from the participants. Furthermore, the functional Near Infrared Spectroscopy, fNIRS, was used to detected brain activity. The results disclosed that the children under the positive encouragement presented much more silent, more attention in the task, including more happy and surprised face expression, stable heart rate variability, and less entropy of brain function connectivity. This study may increase our understanding in child developmental psychology and how to setup a more congenial and close relationship between parents and children.

Key words: positive encouragement, fNIRS

Abbreviation list

fNIRS: Functional Near-Infrared Spectroscopy; **HBO:** Hemoglobin Oxygen; **SPO2:** Peripheral Capillary Oxygen Saturation; **DBP:** Diastolic Blood Pressure; **SBP:** Systolic Blood Pressure; **SDNN:** Standard Deviation of Normal-to-Normal heartbeat

Introduction

The preschool children are undergoing considerable the golden neuroplasticity. Their brain can absorb the undifferentiated information from outside, including the positive encouragement from the parents[1]. Previously studies demonstrated that the encouragement from the parents to child can release the anxiety, build the self confidence in facing of upcoming challenges. For instance, gentle encouragement provides initial support to facilitate adaptive emotional regulation for shy toddlers in social environments. Specifically, by offering positive emotions, enthusiastic responses, and acknowledging the autonomy of toddlers, parents appear to assist toddlers in regulating emotions and cultivating their ability and confidence to actively participate in social interactions[2]. These findings highlighted the potential importance of gentle encouragement in promoting the psychological well-being and social adaptation of young children. If the degree of encouragement is not mild, the expected effects may not necessarily occur[3].

However, little is known about the science behind the encouragement from neural modulation perspective. Exactly, the neural modulation of positive encouragement remains unclear. Here we invited 9 pairs of child-and-parent to participated in the experiments. The child invited to play the game, to stack the blocks. When the blocks stacked enough high, it may fall. Then the parent was randomly invited to make the verbal encourage on his/ her child. The camera was setup in front of the blocks to recording child’s face expression, while the fNIRS used to measure the brain activity during the task.

Methods

Ethical statements

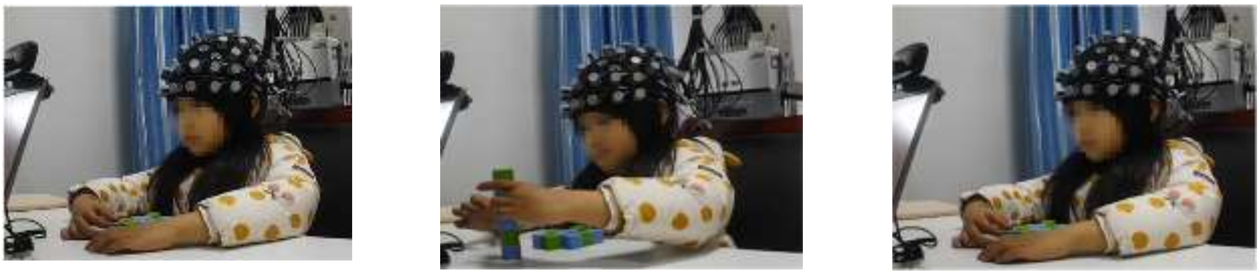
This study was priorly approved by Chengdu medical college ethic committee. All measure procedure was non-invasive and non-contact, while the professional staff surveillance the whole process and ensure the process running smoothly.

Participants

We invited the child and her/his parent to participate in this test together. The inclusive criteria included agreed to capture the physiological data, and no mental or other psychological disease.

Test procedure

At the beginning, the child wore the headcap, in which fNIRS opcode detectors and receivers embedded. The child was instructed to stack the blocks, from bottom to top, during which process, the camera was used to capture the full-face expression process from beginning to the ending. The parents were randomly invited to make the positive encourage in verbal or not, see Figure 1 ABC. The stack task was not less than 3 minutes, to ensure to capture the enough amount of fNIRS data.



A. Before B. In experiment C. After

Figure 1. Experimental process.

Measurement of face expression and life sign

The camera was used to capture the full-face expression. The face reader Noldus 7.0 was used to measure the face expressions; Moreover, we develop a specific algorithm to measure the life signs including the heart rate, SpO2, SDNN, SBP and DBP.

fNIRS

Data acquisition

All fNIRS signals were acquired by a multichannel fNIRS system (NirSmartII-3000A, Danyang Huichuang Medical Equipment Co., Ltd., China) with two wavelengths (730 and 850 nm) at a sampling rate of 11 Hz. The stretchable head cap covered the whole brain area including frontal lobe, bilateral temporal lobe, parietal lobe, and occipital lobe. There are 24 sources and 16 detectors (source-detector separation: 3 cm) on the cap, which form 48 measurement channels. The spatial locations of sources, detectors and anchor points (located at Nz, Cz, Al, Ar, Iz referring to the standard international 10 – 20 system of electrode placement) were measured by an electromagnetic 3D digitizer device (Patriot, Polhemus, USA) on headform (Figure 2). The anatomical regions corresponding to each channel and their respective coverage percentages, see Table1.The acquired coordinates were transformed into MNI coordinates and further projected to the MNI standard brain template using spatial registration approach in NirSpace (Danyang Huichuang Medical Equipment Co., Ltd., China).

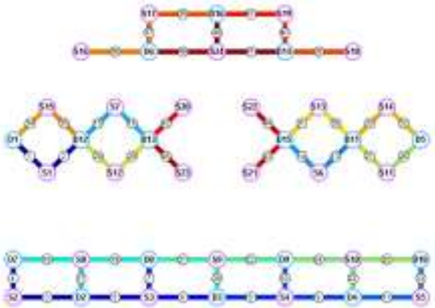


Figure 2. fNIRS Layout Schematic.

Data pre-processing

The fNIRS signals were preprocessed via NirSpark V1.8.1 (Danyang Huichuang Medical Instrument Co., Ltd.). Firstly, head motion correction was performed, followed by digital bandpass filtering of the raw optical density signals within the 0.01 to 0.2 Hz range. Secondly the relative concentration curves of oxygenated hemoglobin, deoxygenated hemoglobin, and total hemoglobin were obtained using the modified Beer-Lambert law[4]. The pathlength factor for each wavelength was set to 6, resulting in relative concentration signals for deoxygenated hemoglobin, oxygenated hemoglobin, and total hemoglobin. Finally, to check the completeness of channels after preprocessing, if any data missing then excluded.

Table 1 The 3D MNI coordinates, anatomic partition and coverage percentage of fNIRS channels.

Channel	MIN			Anatomic partition	Percentage of Overlap（%）
	X	Y	Z		
CH01	63.90	51.74	42.69	Pre-Motor and Supplementary Motor Cortex	0.70
CH02	54.61	-10.23	54.99	Pre-Motor and Supplementary Motor Cortex	0.42
CH03	58.57	32.77	3.09	Inferior prefrontal gyrus	0.45
CH04	62.17	16.12	18.31	pars triangularis Broca's area	0.48
CH05	47.36	55.72	0.58	Frontopolar area	0.86
CH06	19.13	72.12	-0.30	Frontopolar area	0.96
CH07	32.39	61.89	18.38	Frontopolar area	1.00
CH08	-13.38	73.21	-0.96	Frontopolar area	0.90
CH09	-43.08	60.31	-1.91	Frontopolar area	0.92
CH10	-27.20	65.58	18.58	Frontopolar area	1.00
CH11	-57.25	37.36	-1.49	Inferior prefrontal gyrus	0.70
CH12	-61.17	21.81	15.39	pars triangularis Broca's area	0.69
CH13	-42.37	-2.50	62.43	Pre-Motor and Supplementary Motor Cortex	0.99
CH14	-31.24	-2.99	68.43	Pre-Motor and Supplementary Motor Cortex	1.00
CH15	45.13	-27.19	68.12	Primary Somatosensory Cortex	0.65
CH16	33.71	-26.87	72.76	Pre-Motor and Supplementary Motor Cortex	0.31
CH17	52.15	42.16	17.52	Dorsolateral prefrontal cortex	0.86
CH18	54.55	28.74	31.82	Dorsolateral prefrontal cortex	0.96
CH19	39.49	47.69	31.93	Dorsolateral prefrontal cortex	0.69
CH20	2.18	67.38	17.97	Frontopolar area	1.00
CH21	16.24	59.51	34.23	Dorsolateral prefrontal cortex	0.62
CH22	-11.60	61.39	35.03	Dorsolateral prefrontal cortex	0.56
CH23	-49.17	47.31	14.95	Dorsolateral prefrontal cortex	0.77
CH24	-36.06	51.04	30.70	Dorsolateral prefrontal cortex	0.5
CH25	-52.89	33.22	30.41	Dorsolateral prefrontal cortex	0.97

CH26	-62.11	-5.49	41.45	Pre-Motor and Supplementary Motor Cortex	0.98
CH27	-54.02	-4.11	54.55	Pre-Motor and Supplementary Motor Cortex	0.75
CH28	44.68	-7.38	64.03	Pre-Motor and Supplementary Motor Cortex	0.90
CH29	32.84	-7.31	68.68	Pre-Motor and Supplementary Motor Cortex	1.00
CH30	-45.53	-24.91	67.54	Primary Somatosensory Cortex	0.79
CH31	-34.04	-23.27	73.41	Pre-Motor and Supplementary Motor Cortex	0.50
CH32	-65.98	-27.13	46.04	Primary Somatosensory Cortex	0.57
CH33	-55.58	-24.46	58.22	Primary Somatosensory Cortex	0.89
CH34	65.82	-32.26	46.98	Supramarginal gyrus part of Wernicke's area	0.64
CH35	55.54	-29.51	57.04	Primary Somatosensory Cortex	0.57
CH36	36.08	-89.67	32.19	V'	1.00
CH37	24.36	-102.14	15.64	Visual Association Cortex (V)	0.61
CH38	15.91	-107.38	2.82	Visual Association Cortex (V)	0.99
CH39	-38.47	-88.28	32.52	V'	0.97
CH40	-26.61	-101.37	15.02	V'	0.54
CH41	-16.17	-107.34	0.90	Visual Association Cortex (V)	0.99
CH42	21.56	-26.57	77.10	Pre-Motor and Supplementary Motor Cortex	0.37
CH43	-19.11	-3.34	77.18	Pre-Motor and Supplementary Motor Cortex	1.00
CH44	-20.76	-22.13	78.04	Pre-Motor and Supplementary Motor Cortex	0.59
CH45	21.41	-5.24	76.64	Pre-Motor and Supplementary Motor Cortex	1.00
CH46	14.62	-96.96	30.55	V'	1.00
CH47	-14.58	-96.96	30.54	V'	0.99
CH48	-5.63	-103.47	14.11	Visual Association Cortex (V)	0.99

Statistical Analysis

All data were stored and managed in Microsoft office 365.Measurement data were expressed as mean ± standard deviation. Life signs and face expressions data compared via two sides Student’s T test, while the fNIRS data in double group also compared via two sides Student’s T test . Stata 18 software was used for statistical analysis. A p-value < 0.05 was deemed as the statistically significant.

Results

Demographic:

In total, 9 children included, 6 boys and 3 girls, averages 5 years old. Good relationship between parents-child remains. While they were randomly divided into encouragement group (N=7) and no-encouragement group (N=2).

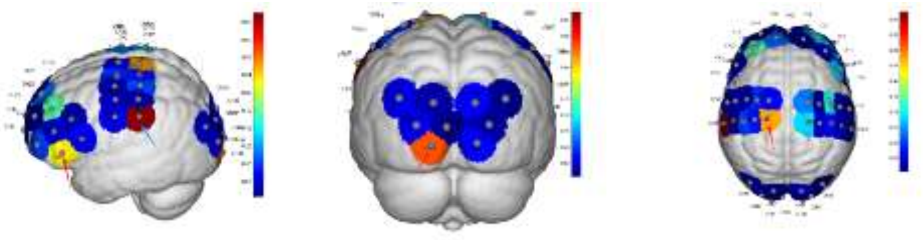
HBO (Hemoglobin Oxygen) Levels changed in double groups

A two side student’s t-test was conducted to analyze the differences of blood oxygen concentrations between the two groups. Group 1 defined as the encouragement group, and Group 2 defined as the non-encouragement group. The results showed statistically significant differences in blood oxygen concentrations between double groups in CH11 (T=−2.4138, P=0.045837), CH32 (T=3.6658, P=0.008009), CH41 (T=−2.9461, P=0.021526), and CH44 (T=2.6962, P=0.03081), details see Table 1. The channels' locations in the brain were illustrated in the Figure 3. No statistically significant differences were found between the two groups after FDR correction.

Table 2 Analysis of Differences in Mean Blood Oxygen Concentration Indicators

Channels	S-D	T	degree of freedom	p
11	S5-D4	-2.4238	7	0.045837
32	S14-D5	3.6658	7	0.008009
41	S19-D16	-2.9461	7	0.021526
44	S22-D15	2.6962	7	0.03081

P＜0.05, The results differ significantly in terms of statistical analysis.



Channel 11+32 CH41 CH44

Figure 3 Differential Brain Map of Mean Blood Oxygen Concentration, with the red arrow.

Time-Dependent Effects on Blood Oxygen Concentration

The following line graph presented the dynamic changes of blood oxygen concentration between double groups, it reveals that the encouragement group exhibited a comparatively lower significantly overall magnitude of changes of HBO than the non-encouragement group throughout the entire task duration, see Figure 4. No significant difference of average HBO in double groups were found, see figure 5.

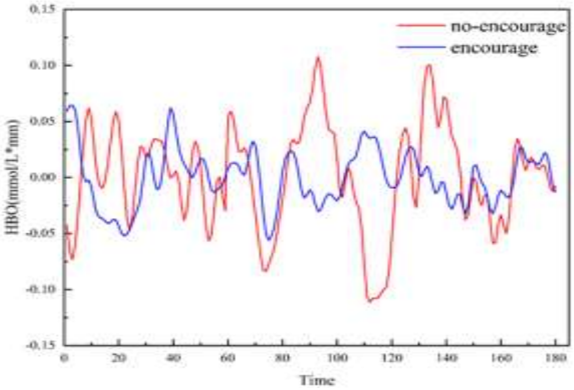


Figure 4. Changes in blood oxygen concentration

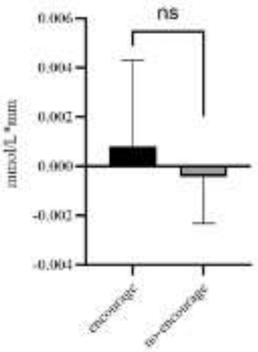
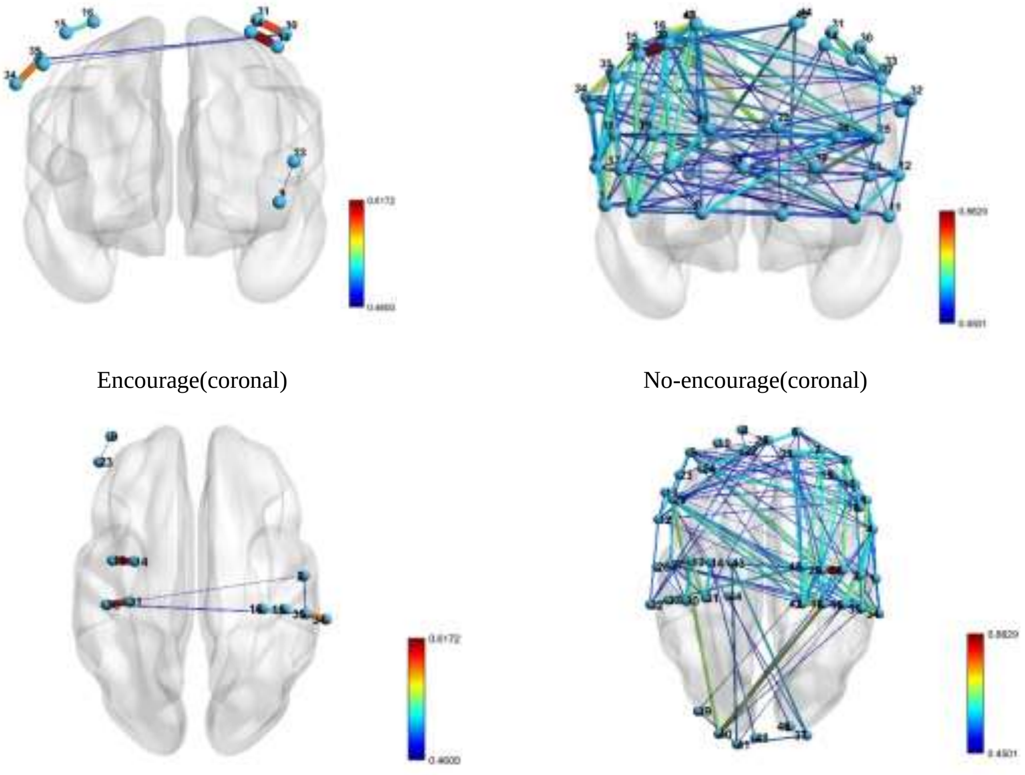


Figure 5. Inter-group Differences in Mean HBO

Visualization of Functional Connectivity during the task period in two groups

With respecting the functional connectivity of double groups during the task, we further visualized the differential functional connectivity strength during task under encourage and not, our data demonstrated that the encourage group exhibited the significant less functional connectivity strength during task execution, whereas the non-encourage group present more entropy and stronger functional connectivity. The threshold for both group connectivity strengths was set at 0.45, see Figure 6.



Encourage(coronal)

No-encourage(coronal)

Encourage(axial)

No-encourage(axial)

Figure 6 Visualization of Functional Connectivity during the task period

Inter-Group Analysis of Differential Functional Connectivity

A student’s T test was used to analyze the differential brain functional connectivity strength between the double groups, the results revealed statistically significant differences in 12 channel pairs: CH4 and 15 (T=−5.23022, P=0.001435), CH6 and 18 (T=−3.56155, P=0.0092), CH6 and 44 (T=−3.66242, P=0.008044), CH7 and 42 (T=−4.17134, P=0.004182), CH11 and 48 (T=−3.54775, P=0.009372), CH14 and 48 (T=−5.85854, P=0.000625), CH19 and 28 (T=3.504201, P=0.009937), CH19 and 29 (T=3.765882, P=0.009221), CH26 and 38 (T=3.951806, P=0.005519), CH29 and 32 (T=7.07418, P=0.000396), CH32 and 45 (T=5.030033, P=0.001513), see Table 3. These differences were failed to pass after FDR correction, this may cause by limited sample size, see table 3 and figure 7.

Table 3 Results of Inter-Group Differences in Functional Connectivity Strength

Ch Name	1 Mean	2 Mean	1 Std	2 Std	1 nSubj	2 nSubj	p	p (FDR corrected)	T-test
4~15	0.163831	0.574836	0.201518	0.027346	7	2	0.001435	0.426753012	-5.23022
6~18	-0.00163	0.482378	0.149391	0.259218	7	2	0.0092	0.945621754	-3.56155
6~44	-0.19887	0.364586	0.189566	0.205221	7	2	0.008044	0.945621754	-3.66242
7~42	0.116371	0.63155	0.156576	0.137827	7	2	0.004182	0.943489515	-4.17134
11~48	0.001606	0.428047	0.146015	0.17147	7	2	0.009372	0.945621754	-3.54775
14~48	-0.11074	0.508552	0.142393	0.004523	7	2	0.000625	0.352591265	-5.85854

19~28	0.291576	-0.17354	0.178691	0.015984	7	2	0.009937	0.945621754	3.504201
19~29	0.100772	-0.31144	0.289127	0.008906	7	2	0.009221	0.945621754	3.765882
26~38	0.172938	-0.35775	0.176814	0.093729	7	2	0.005519	0.945621754	3.951806
29~32	0.170436	-0.26607	0.163164	0.002899	7	2	0.000396	0.352591265	7.074180
32~45	0.037282	-0.6147	0.17443	0.019713	7	2	0.001513	0.426753012	5.030033

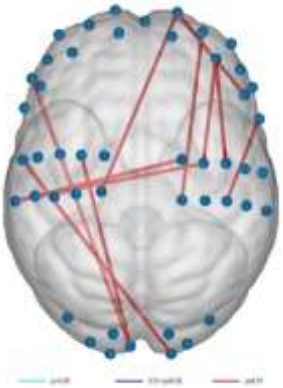


Figure 7 Visualization of Functionally Differentiated Connectivity

Face expression and life signs

Interestingly, we found the face expressions including, neutral happy sad angry surprised scared and disgusted, presented the significant different between double groups. Moreover, SPO2 exhibited significant different between two groups. SDNN presented significant higher in no-encourage group, details see Figure 8.

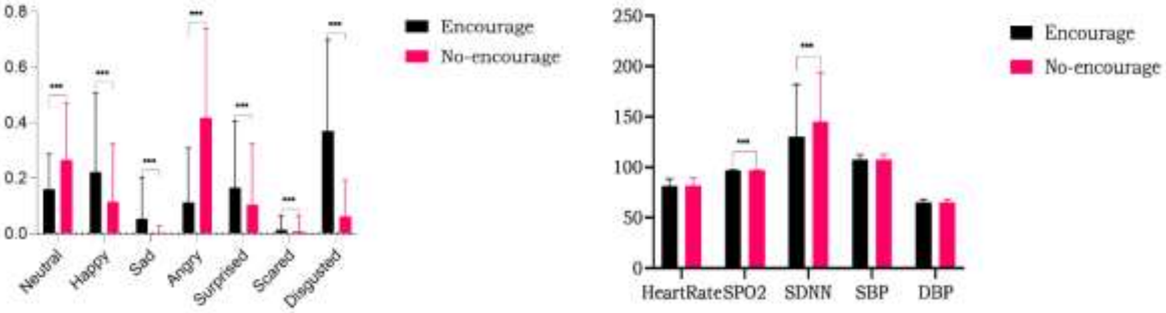


Figure 8. Face expressions (left) and life signs (right) ***p<0.01

Discussion

The whole experiment contains 9 pairs of child-and-parent. The encourage group presented more silent, happy and surprised face expression, less SDNN and functional connectivity strength during the task than the controls. In total, there were four channels, represent the Inferior Prefrontal Gyrus, Primary Somatosensory Cortex, Visual Association Cortex (V), and Pre-Motor and Supplementary Motor Cortex. With regarding to the blood oxygen concentration level in the brain during the task, the encourage group displayed the significantly lower than controls, this signifying the sedative effect work. We infer the reasons as followed, see figure 9.

Figure 9. The effect positive encouragement from parent on child.

Firstly, the emotional regulation effect. The encouragement may lead to a heightened sense of joy and relaxation in children, thereby to release emotional stress and tension. A pattern characterized by high levels of warmth and acceptance, coupled with parents avoiding the use of pressure tactics, may lead to better performance in various domains for children. This encompasses academic achievements, peer relationships, and overall psychological well-being. This emphasizes the positive influence of the family environment on children achieving better functional outcomes in different aspects. [5]. The regulation of emotional states could potentially impact blood oxygen concentration, resulting in a smaller magnitude of overall changes. Secondly, to reduce the cognition load. The encouragement may assist children in better understanding tasks and increasing their enthusiasm for participation. This could be significant reduced cognitive load during task execution, influencing the magnitude of changes in blood oxygen concentration, according to the entropy changes in functional connectivity. Interestingly, the research findings by Tiansheng Xia et al. indicate that parental encouragement not only directly influences children's reading motivation but also exerts an indirect effect through the mediation of reading self-concept. The impact of parental encouragement is more positive for boys than girls, while the influence of reading self-concept is more positive for girls than boys. The results underscore the significance of parental encouragement in enhancing children's reading motivation[6]. This aligns with our conjecture that parental encouragement can enhance children's understanding and enthusiasm for tasks.

Thirdly, to enhance the attention: The Encouragement may significantly to improved concentration and reduced distraction in children. Some research suggests that parents' emotion-related socialization behaviors may shape the children's social-emotional functioning, providing children with more emotional strategies and encouragement. Specifically, parental encouragement is significant for children with Attention Deficit Hyperactivity Disorder (ADHD)[7]. Stable attentional focus may result in relatively stable blood oxygen concentrations, diminishing the amplitude of fluctuations.

Finally, the physiological sedative effects. In general, positive encouragement and affirmative recognition typically have positive effects on children's physiological indicators. The Positive encouragement may induce physiological sedative effects, such as a decrease in heart rate and more stable frequency in breathing. These physiological changes may be associated with positive encouragement from parents. Research indicates that the effect of encouragement on emotional regulation in preschool children, a matter of significant importance in the fields of medicine and psychology. When parents provide positive encouragement to their children, it can trigger positive changes in the children's internal emotional experiences, leading to a cascade of physiological and psychological effects. The emotional regulation effect of encouragement primarily manifests as a positive transformation in the emotional states of children. Through verbal support from parents, children may undergo a heightened sense of joy and relaxation. This positive emotional change contributes to the reduction of negative emotions experienced by children, such as anxiety and tension[8]. This is attributed to the positive encouragement stimulating neural circuits in the children's brains associated with pleasure and emotional regulation, prompting the release of beneficial physiological signals, such as endogenous neurotransmitters and hormones. Interestingly, the encourage group presented significant lower brain functional connectivity strength than the controls during task execution according to the functional connectivity. In the encourage group, children presented more confident in facing of challenges. The encourage may strength the reward effect, which may reduce the anxiety and nervous mood. Our participants presented more happy face expression than the controls, and less SDNN than the controls. Higher SDNN was normally understood as low stress, the lower SDNN in the encouragement group suggested that there was some pressure to complete the task under the encouragement of parents[9], in consistent with Francis's study that moderate pressure can have a positive impact on task completion[10]. These physiological parameters validated the stable mood of participants in the task. All these factors can help them in attention and manipulate the detail task.

Taken together, the child present more relaxed mood and confident under the encouragement. It can not only to strength the bond of parent-child, but also enhance the child's confidence in facing of upcoming challenges.

References

1. Jasińska, K.K., et al., *Functional connectivity in the developing language network in 4-year-old children predicts future reading ability.* Dev Sci, 2021. **24**(2): p. e13041.
2. Grady, J.S., *Parental gentle encouragement promotes shy toddlers' regulation in social contexts.* J Exp Child Psychol, 2019. **186**: p. 83-98.
3. Kiel, E.J., J.E. Premo, and K.A. Buss, *Maternal Encouragement to Approach Novelty: A Curvilinear Relation to Change in Anxiety for Inhibited Toddlers.* J Abnorm Child Psychol, 2016. **44**(3): p. 433-44.
4. Wang, L., et al., *Evaluation of light detector surface area for functional Near Infrared Spectroscopy.* Computers in biology and medicine, 2017. **89**: p. 68-75.
5. Lessard, J., E. Greenberger, and C. Chen, *Adolescents' response to parental efforts to influence eating habits: when parental warmth matters.* J Youth Adolesc, 2010. **39**(1): p. 73-83.
6. Xia, T., H. Gu, and W. Li, *Effect of Parents' Encouragement on Reading Motivation: The Mediating Effect of Reading Self-Concept and the Moderating Effect of Gender.* Front Psychol, 2019. **10**: p. 609.
7. Smit, S., A.Y. Mikami, and S. Normand, *Effects of the Parental Friendship Coaching Intervention on Parental Emotion Socialization of Children with ADHD.* Res Child Adolesc Psychopathol, 2022. **50**(1): p. 101-115.
8. Shalev, I., et al., *Parental guilt and children's internalizing and externalizing behavior: The moderating role of parental reflective functioning.* Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43), 2023. **37**(8): p. 1241-1252.
9. Damapong, P. and P. Damapong, *Short-Term Effects of Aroma Therapeutic Herbal Steam on Heart Rate Variability and Stress.* American Journal of Applied Sciences, 2018.
10. Francis, A. *The Effects of Positive and Negative Stress in the Workplace*. 2018.

Classifying arousal and valence from facial expressions and physiological responses evoked by multiple stressors

Ivo Stuldreher[1], Juliette Bruin[1], Anne-Marie Brouwer[1,2]

[1] Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, The Netherlands; [2] Artificial Intelligence, Donders Centre, Radboud University, Nijmegen, Netherlands; ivo.stuldreher@tno.nl

Introduction

High-risk professionals, such as police officers or military personnel, are frequently exposed to stressful circumstances and therefore need to be stress resilient to recover rapidly from these situations. Screenings and assessments for high-risk professionals therefore often include interviews and questionnaires to assess candidates' stress resilience. However, both self-assessment questionnaires and assessments of recruiters can be biased, unreliable or incorrect.

Automatically detecting stress from video images of the face could support evaluating stress responses in applicants for high-risk jobs, or could even contribute to timely stress detection in challenging operational settings. Various studies have shown that facial expressions, analyzed using machine learning models, are informative of mental states. Challenges in automatically estimating mental state include the generalization of models across contexts and across participants. We here aim to create robust models by training them using data from different contexts and including physiological features.

Methods

The study was approved by the Internal Review Board at TNO (reference number 2022-093). Fifty-one participants (25 male, mean age = 38, SD = 13.49) were exposed to different types of stressors and corresponding baseline variants. The stressors included a mental capacity test of 40 multiple choice questions that had to be answered under time pressure (cognitive stressor), the instruction to sing a song out loud after a 60-second countdown (social evaluative stressor), and a public speaking task in which participants were instructed to verbally reply with a thoughtful moral judgement on a moral dilemma (social evaluative stressor). Corresponding baseline variants were a test of 40 multiple choice questions in which participants were guided to the answer in the question, a neutral sentence followed by a 60-second countdown, and a public speaking task in which participants were asked to tell about their day.

Video, electrocardiogram (ECG), electrodermal activity (EDA) and self-reports (arousal and valence) were recorded. For video data, features were extracted using OpenFace (facial action units, gaze, head movement). Features from ECG and EDA included heart rate, heart rate variability, and phasic skin conductance. Logistic regression models aimed to classify between high and low arousal and valence across participants, where 'high' and 'low' were defined relative to the center of the rating scale. Accuracy scores of different models were evaluated: models trained and tested within a specific context (either a baseline or stressor variant of a task), intermediate context (baseline and stressor variant of a task) or general context (all conditions together). Furthermore, for these different model variants, only the video data was included, only the physiological data, or both video and physiological data.

Classification accuracies were compared to assessment by three selection psychologists. Each of these three observers rated the arousal and valence of each participant during each of the tasks, based on 10-second fragments of video-data.

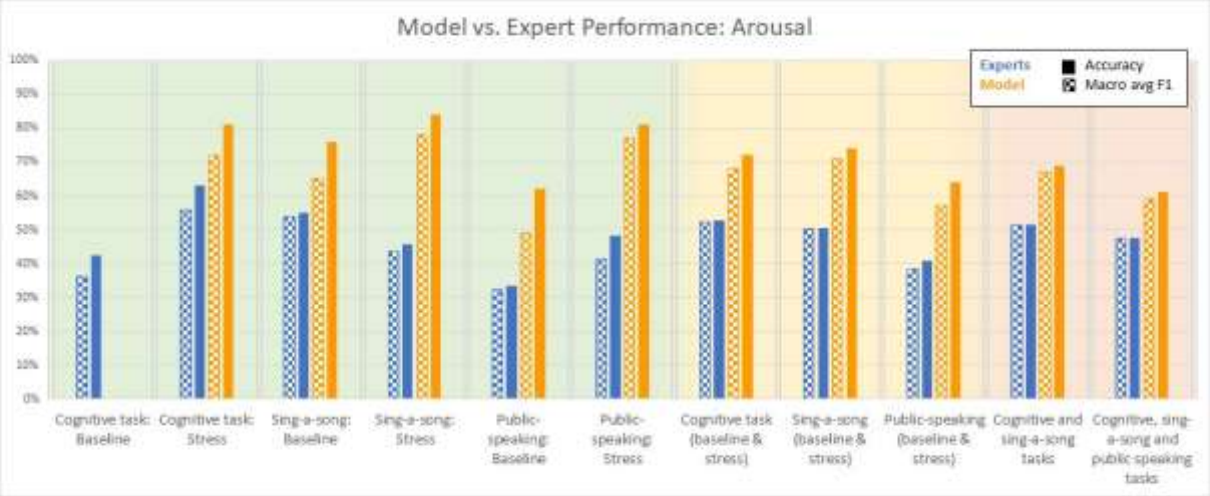
Results

Figure 1 summarizes the mean accuracies for predicting high vs. low arousal and valence based on models trained on video-physio features and based on assessment by the human observers. We found that all (video, physiological and video-physio) models could successfully distinguish between high- and low-rated arousal and valence, though performance tended to be better for 1) arousal than valence, 2) specific context than intermediate and general contexts, 3) video-physio data than video or physiological data alone. Automatic feature selection resulted in inclusion of 3 to 20 features, where the models based on video-physio data usually included features from all modalities: video, ECG and EDA. Still, performance of video-only models approached the performance of video-physio models.

Arousal and valence ratings by the three experienced human observers did not match with self-reports.

Conclusion

In sum, we showed that it is possible to automatically monitor arousal and valence even in relatively general contexts and better than humans can (in the given circumstances), and that non-contact video images of faces capture an important part of the information, which has practical advantages.



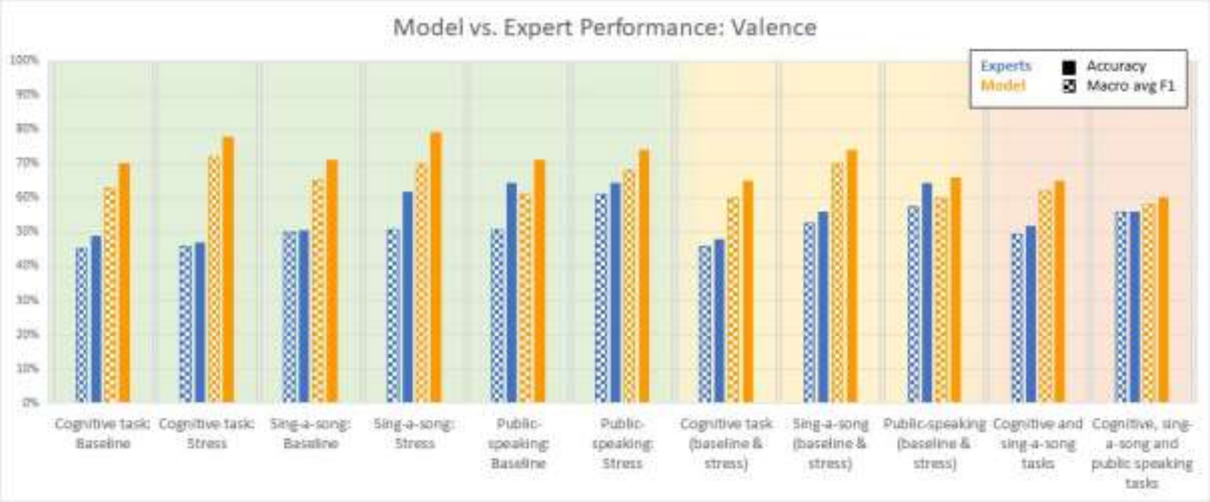


Figure 1. Mean accuracies and macro average F1 scores for predicting high vs. low arousal (top) and valence (bottom) based on models trained on video and physiological features or based on expert ratings for different sets of tasks: specific context (green background), intermediate (yellow), and general context (red).

Non-invasive ways to measure sleep behavior in family dogs

Anna Kis, József Topál

Institute of Cognitive Neuroscience and Pschology, HUN-REN, Budapest, Hungary

Introduction

Sleep is a fundamental part of the mammalian behavior repertoire, as most animal species spend a considerable time of their life sleeping. This is also true for dogs, an important study species both from the applied perspective (commonly kept as pets or for working purposes) and as models for human behavior. Furthermore, sleep and related physiological processes are strongly interrelated with both awake behaviors (including social and cognitive aspects) as well as individual parameters such as ageing. It is, however, not possible to gain detailed information about an animal's sleep (e.g. time spent in the different sleep stages) using classical behavioral observation only. In this paper I will present a fully non-invasive method that allows for the parallel monitoring of different parameters during sleep (including brain activity), thus providing detailed information on dogs' sleep behavior.

Method

Sleep in dogs (and wolves) is monitored by polysomnograhy (PSG), simultaneously recording neural oscillations (EEG), electrooculogram (EOG), electrocardiogram (ECG), respiration, and electromyography (EMG). Electrode placement (Figure 1) involves attachment of scalp electrodes over the anteroposterior midline of the skull (Fz, Cz, Pz) and on the left zygomatic arch (os zygomaticum; F7). The Fz-Cz derivation serves as the EEG signal, the F7-Cz derivation serves as the EOG signal. The ground electrode (G) is placed on the left musculus temporalis. Gold-coated Ag|AgCl cup electrodes fixed with EC2 Grass Electrode Cream (Grass Technologies, USA) are used. All scalp electrodes are placed on a bone to minimize muscle tone and movement artifacts. ECG electrodes are placed bilaterally over the second rib and EMG electrodes are placed bilaterally on the musculus iliocostalis dorsi. Respiration is recorded via a chest respiratory belt. The electrode placement does not require to shave the dogs' fur, the electrode cream used is water-soluble and can simply be washed off after the completion of the measurement.

Signals are collected, prefiltered, amplified and digitized at a sampling rate of 1024 Hz/channel by using the SAM 25 R style MicroMed Headbox (MicroMed Inc, Houston, TX, USA), with hardware passband at 0.5–256 Hz, sampling rate of 512 HZ, anti-aliasing filter with cut-off frequency at 1 kHz, and 12-bit resolution covering a voltage range of±2 mV as well as second-order software filters at 0.016 Hz (high pass) and 70 Hz (low pass) using System Plus Evolution software (MicroMed Inc, Houston, TX, USA). Impedances for the EEG electrodes were kept below 20 kΩ.

Participation in the PSG research does not require any prior training from the dogs, even dogs without a basic obedience training can be involved in the study. Furthermore no sedatives are used to induce sleepiness; measurements last for a duration of three hour (per occasion) which allows enough time for dogs to spontaneously fall asleep and thus enables the recording of natural sleep behavior and physiology. It is, however, advised that data from the second and third occasions are used to test behavioral treatment effects (e.g. pre-sleep learning versus control), discarding the first sleeping occasion due to what is known in the human literature as the so called "first night effect".

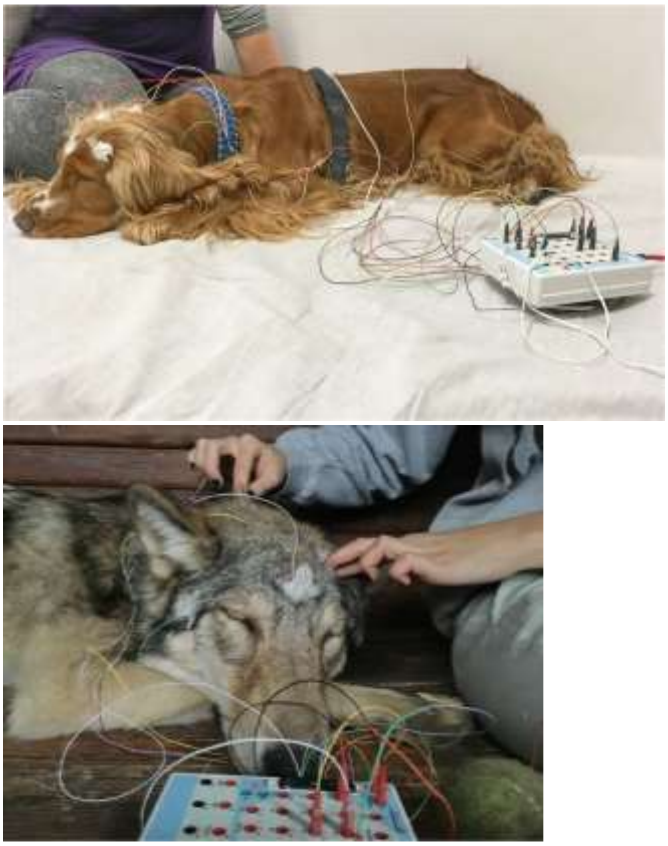


Figure 1. Photographs showing polysomnography measurement setup for a family dog (left) and a hand-raised wolf (right)

Results

Over the past 10 years, me and my colleagues have conducted several experiments using dog polysomnography. These included studies on learning and memory consolidation, the effects of positive and negative social interactions on dogs' sleep, the effect of selective rem- and non-rem sleep deprivation, as well as individual sleep fingerprints of behavioral variation (e.g. susceptibility of being observed or ADHD-like symptoms), ageing and cognitive decline, skull length, etc.

The proposed talk will focus on methodological research with a two-fold aim: 1) to present data on the validity of the method (both in terms of the measurement procedure as well as the data analysis) and 2) to highlight factors that crucially influence the measurement and thus need to be taken into consideration during study design.

Using the exact same electrode placement on human subjects and family dogs we showed that dogs' sleep EEG resembled that of human subjects and was generally in accordance with previous literature using invasive technology [1].

Following an active (compared to a passive) day, dogs slept more, were more likely to have an earlier drowsiness and NREM, and spent less time in drowsiness and more time in NREM and REM. At nighttime (compared to during the afternoon), dogs slept more and spent less time in drowsiness and awake after first drowsiness, and more time in NREM and in REM. When not at home (compared to at home), REM sleep following a first NREM was less likely. [2]

Detailed reliability analysis of sleep structure scoring showed that the data analysis method results in sustainable inter-rater agreement and works better than neural learning algorithms [3].

In addition to the most commonly used sleep macrostructure and sleep EEG spectrum data, other kinds of information can also be extracted from the PSG recordings. These include sleep spindles via using an automated algorithm [4], rapid eye movement density during REM sleep [5] as well as heart rate and heart rate variability [6].

Conclusion

Canine polysomnography (PSG) is a fully non-invasive method that is easy to carry out and allows for detailed data collection during dogs' (and wolves') sleep. The PSG procedure has been used in the past decade on N > 200 dogs to answer several scientific questions and also many methodological experiments have also been carried out that inform us about its validity as well as serve as guidelines for designing future research.

Ethical statement

All experimental protocols were approved by the Scientific Ethics Committee for Animal Experimentation (Állatkísérleti Tudományos Etikai Tanács) of Budapest, Hungary (number of ethical permission: PE/EA/853-2/2016).

References

1. Kis, A., Szakadát, S., Kovács, E., Gácsi, M., Simor, P., Gombos, F., ... & Bódizs, R. (2014). Development of a non-invasive polysomnography technique for dogs *(Canis familiaris). Physiology & behavior, 130,* 149-156.

2. Bunford, N., Reicher, V., Kis, A., Pogány, Á., Gombos, F., Bódizs, R., & Gácsi, M. (2018). Differences in pre-sleep activity and sleep location are associated with variability in daytime/nighttime sleep electrophysiology in the domestic dog. *Scientific Reports, 8(1),* 7109.

3. Gergely, A., Kiss, O., Reicher, V., Iotchev, I., Kovács, E., Gombos, F., ... & Kis, A. (2020). Reliability of family dogs' sleep structure scoring based on manual and automated sleep stage identification. *Animals, 10(6),* 927.

4. Iotchev, I. B., Kis, A., Bódizs, R., van Luijtelaar, G., & Kubinyi, E. (2017). EEG transients in the sigma range during non-REM sleep predict learning in dogs. *Scientific Reports, 7(1),* 12936.

5. Kovács, E., Kosztolányi, A., & Kis, A. (2018). Rapid eye movement density during REM sleep in dogs *(Canis familiaris). Learning & Behavior, 46,* 554-560.

6. Varga, B., Gergely, A., Galambos, Á., & Kis, A. (2018). Heart rate and heart rate variability during sleep in family dogs *(Canis familiaris).* Moderate effect of pre-sleep emotions. *Animals, 8(7),* 107.

Towards a Multi-Modal Human Digital Twin for Nutrition and Wellbeing

N. Thammasan[1], T. Teichmann[1], A. van Kraaij[1], S. Gaitan[1], and R. van Stiphout[1]

1OnePlanet Research Center, Wageningen, Netherlands. nattapong.thammasan@imec.nl

Introduction

In the current context of an aging population and increasing healthcare costs, preventive health is becoming increasingly important. Diet is one of the major factors for prevention and management of chronic diseases, and it is crucial in the maintenance and improvement of overall health.[1] Dietary effects on wellbeing can also occur on a daily basis; for example, a bi-directional relationship has been found between over- and undereating and mood or psychological state.[2] Furthermore, there is accumulating evidence that dehydration [3] affects mood and cognitive performances negatively [4] and that long-term dehydration may result in increased mortality, morbidity, and disabilities in elderly [5, 6], who have a decreased gene transcription response to dehydration and are more vulnerable to the related neurogenerative impacts [7]. Dietary advice could help individuals to avoid the short-term negative effects of suboptimal diet choices, and therefore also prevent more severe long-term effects. However, one-size fits all diet advice is known to be ineffective because of the high variability in responses to nutritional intake between individuals, as in for example the glycemic response [8].

A personalized approach to dietary advice is thus needed. However, the high complexity of the biological and physiological processes involved and the intricate interplay with an individual's behavior and activities requires a data collection approach capable of considering all these factors. This asks for continuous measurements from multiple data sources and data integration to acquire insights and deliver recommendations that can be provided continuously in ambulant environments. Capturing the required physiological and behavioral data, processing relevant insights, and delivering timely and actionable recommendations for multiple individuals at a personal level poses a technical challenge. A solution to tackle this requires a multiplicity of knowledge domains and technologies that must be integrated and coordinated continuously, involving wearables, mobile apps, smart nutritional intake logging, cloud infrastructure, and software and Machine Learning development and operations.

Digital Twin technology has high potential to tackle such challenges, delivering continuous and personalized lifestyle advice. Digital Twins have already been used for a few decades in the context of manufacturing and engineering. [9, 10]. A Digital Twin is a virtual representation of a physical entity [11]. Digital Twinning enables the optimization of processes or interventions by simulating the effects of certain contexts or interventions in the twin before applying it on the physical entity. It also facilitates the validation of the (short-term) effects such measures may have on the subject. The same applies to Digital Twin for human health; a digital twin of a person, which is trained and updated continuously through continuous data capture using sensors, can model and simulate health processes by state-of-the-art artificial intelligence (AI). Based on its predictions, it can provide just-in-time feedback to the user to improve or maintain health. In order to operationalize Digital Twins in both healthcare and preventive health settings, it is important to take the following aspects into account: sufficient data quality that is continuously validated, integration of data from different devices, interoperability with other health platforms by having standardized data formats, a digital infrastructure and standardized data pipelines to execute the required real-time data processing, data security through compliance with regulations and robust access controls, and ethical guidelines for responsible use of digital twins [12].

In this short paper, we introduce proof of concept of such technology in the context of nutrition and wellbeing based on a pilot experiment. It covers continuous data capture, continuous data processing, integration in the cloud and a set of preliminary health-related insights. In synthesis, the aim of this proof of concept is to assess the technological feasibility of a Digital Twin to capture, process, and integrate the data required to enable the delivery of diet recommendations to improve the subject wellbeing.

Methods

Study

Data was collected from three healthy participants over five full working days. The study was exempted from being subject to the Medical Research Involving Human Subjects Act (WMO) by a Dutch Medical Ethical Committee. Participants have signed informed consent forms. During the study, the participants were instructed to log their food intake using the Mijn Eetmeter app [13], and their drinks were continuously logged on a research device called SnackBox [14]. This device automatically detects drinking moments by tracking weight changes over time in RFID-tagged drinking cups that are placed on one of the three weighing stations on the device. Each participant was also wearing a Garmin Vivosmart 5 to capture heart rate (variability) and physical activity. Five questionnaires were sent to the in-house developed OnePlanet Research App [15, 16] at random times between 09:00h and 21:00h to collect mood and physical states from experience sampling (EMA). The 6 EMA items on a Visual Analogue Scale (VAS) scale (0-100) were: dehydrated – hydrated, low energy – high energy, stressed – relaxed, hungry – satiated, dry mouth – moist mouth, and headache – no headache.

The data capture was automated using an in-house developed Azure cloud environment (Figure 1). Both SnackBox data and Garmin wearable data were uploaded every 15 min into the cloud environment, but for the wearable it was done via the Research App. Participants were also notified of new questionnaires in this mobile app, after which they could fill in and submit them in the same app. Information regarding nutrition, including energy (kCal), water content, and sugar content, can be extracted from the Mijn Eetmeter food log data via their platform based on Voedingscentrum database, which was then regularly exported as a CSV file and uploaded to the same cloud environment. The Microsoft service AzureML was used to set up a data pipeline that processes all data in a real-time manner. AzureML provides a single environment to train, validate, store, and deploy models and algorithms. It uses a standardized framework to generate data pipelines and log the metadata of all activities. The data is stored in a medallion architecture, in which raw data is stored in the bronze layer; processed data, such as features in time, are stored in the silver layer; and relevant insights, such as model outputs, are stored in the gold layer. This tiered system aims to properly maintain the structure and quality of the data.

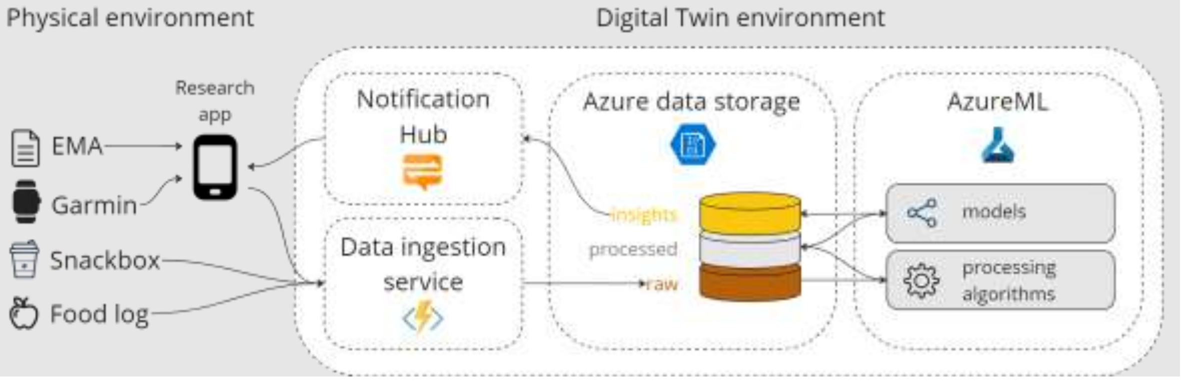


Figure 1. The implemented bi-directional link between the physical environment (healthy person) and the digital twin environment in the cloud.

Data Processing

Physiological features are calculated from Garmin data. Heart rates were averaged within a 5-minute sliding window, without overlapping between consecutive windows. Heart rate variability (HRV) was derived from inter-beat interval using the root mean square of successive differences (RMSSD) method, with a window size of 5 minutes. Features were collected from 3 hours before EMA time to investigate the change of physiology before the reported wellbeing scores [17]. In addition, features within 1 hour after the EMA time were collected to investigate the change of physiological signals after the moment.

Likewise, features regarding nutritional intake from foods and drinks were collected by calculating the cumulative amounts of energy (kCal) and water content (ml), using the same window of 3-hour before and 1-hour after EMA. Features from physiological signals were averaged within those windows, while features regarding the extracted nutritional intake are summed over the window.

Results

Overall, 51 EMA surveys were completed, 314.35 hours of Garmin data was recorded, and 476 times of food and drink intakes were registered. On average, participants consumed an energy intake of 2456.6 (SD = 782.1) kCal and water of 2611.5 (SD = 625.6) ml from food and drinks per day. Figure 2 depicts the raw heart rate data, the reported questionnaire answers, and cumulative energy from food and drinks (in kCal) and water intake (in ml) over the course of time for one of the participants. In this example it is noticeable that the participant had a higher heart rate during lunch time, and after working hours. Also, consistent headaches throughout the day can be observed. Low drinking and food intake in the afternoon resulted in reported tiredness (low energy), stress (low relaxation), hunger (low satiation), and dehydration, which were all mitigated after water drinking and dinner in the evening.

EMA scores were categorized into two categories, at the threshold of 50, which is the middle point in the VAS scale (0-100). The statistics of features accumulated from all participants were calculated within each category. Initial insights are shown in Figure 3. Figure 3A reveals a relationship between caloric intake and perceived energy levels. Participants experienced tiredness after consuming more calories, namely food energy dip. Figure 3B showcases the association between fluid intake and satiety. It indicates a noticeable trend where increased water consumption aligns with a greater sense of satiation. In Figure 3C, it might also be logical to infer that after reporting dehydration, the participants moved to take more water or food, resulting in higher heart rate due to the entailed physical activity. This may reflect an immediate action to dehydration constituting the change in the physiological state. Before experiencing a headache, participants had low heart rate variability, resulting in low RMSSD value (Figure 3D); a further analysis also showed that the participants had constantly high heart rates during that time.

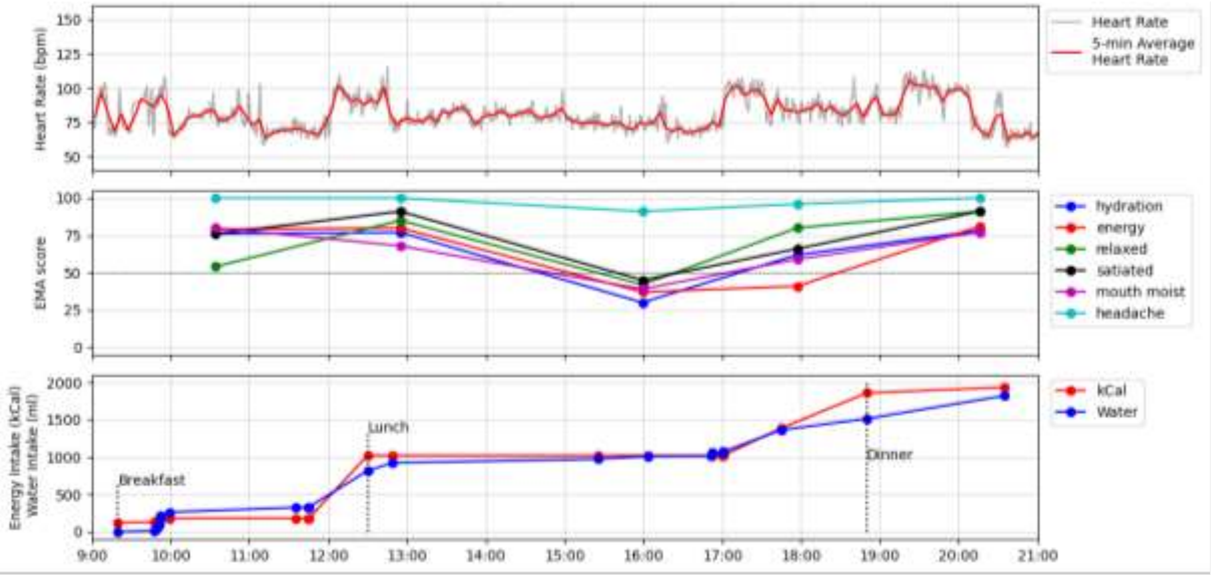


Figure 2. Example of integrated daily time series data of a single participant in this pilot study, displaying heart rates over the day, answered scores in five questionnaires regarding hydration (dehydrated-hydrated), energy (tired-energized), relaxation (stressed-relaxed), satiation (hungry-satiated), mouth moisture (dry-moist), and presence of headache (no-yes), where intake amounts of energy and water content from foods and drinks are also shown.

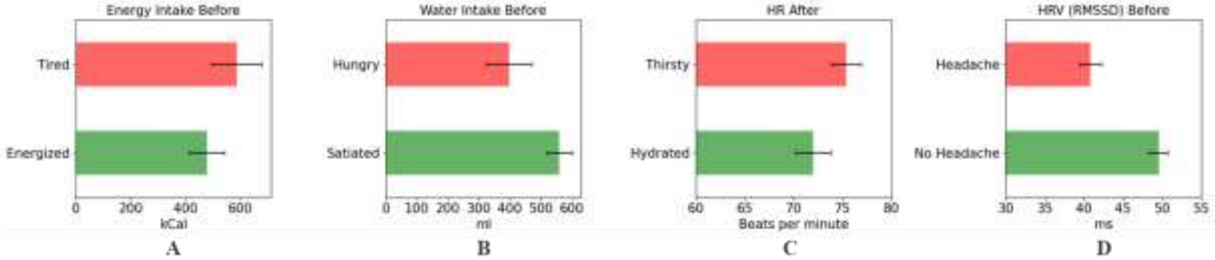


Figure 3. Initial insights on linking nutritional intake and cardiac measures from the wearable to self-reported mood and physical states, 3-hour before (A, B and D) and 1-hour after (C) the EMA reported time.

Discussion

In this short paper we introduced the concept of digital twins in health and presented preliminary results of a first iteration in the operationalization of a Digital Twin cloud environment with the use of real-time data captured in a study pilot. The use of two continuous, connected sensing devices, i.e., the Garmin watch and the SnackBox, and an automated experience sampling through questionnaires in a smartphone app, resulted in a continuous data stream into the Digital Twin cloud environment. The environment was able to successfully preprocess this data in real-time and generate insights. Making the Digital Twin cloud-ready, guaranteed scalability and security, as well as a higher level of reusability.

Logging food items still required a manual step of uploading data into the environment. This could be improved with an app that communicates with the environment directly through an SDK or with emerging technologies for automated food tracking [18].

In our pilot study we did not close the Digital Twin loop; specifically, no data was sent back to the participants and no interventions on diet changes were recommended. Although the implemented system already has the connectivity or notification capabilities to perform those actions, additional improvement iterations are required to properly deliver the content of a notification so that it becomes an actionable, positive recommendation.

Future work in this direction implies improving the data collection strategy to achieve a more accurate representation of the health processes. Personalizing Digital Twins demands rich data about individuals over time to properly calibrate existing models at a personal level. Also, additional sensors that can directly measure physiological variables such as urine and gut conditions would allow us to better manage the inherent subjectivity of information from questionnaires. Developing gut and urine biomarkers, while keeping track of behavior, can enrich the availability of key measurements. Secondly, multivariate modeling approaches need to be in place to capture the complexity of the physiological processes involved. Approaches that would facilitate making predictions in the future on health status can be the development of fully data-driven models using machine learning, mechanistic simulation models where prior knowledge of the involved physiological mechanism can be embedded, or a combination of the two to better estimate the parameters and their uncertainties in a simulation model [19].

The Digital Twin technology shows great potential in the health domain to help patients and healthy people to improve and maintain their health, given the continuously learning virtual representation of them that leverages their monitored and predicted health status and behavior to give tailored advice. On the other hand, Digital Twins for health can also facilitate research and development by providing a platform that has standardized and continuous data capturing, real-time preprocessing and validated simulation models that can be leveraged to generate data, test interventions virtually and run intervention studies.

References

1. Neuhouser M. L. (2019). The importance of healthy dietary patterns in chronic disease prevention. *Nutrition research*, **70**, 3–6.

2. Polivy, J., & Herman, C. P. (2005). Mental Health and Eating Behaviours: A Bi-directional Relation. *Canadian Journal of Public Health*, **96** (Suppl 3), S49–S53.

3. Lacey, J., Corbett, J., Forni, L., Hooper, L., Hughes, F., Minto, G., Moss, C., Price, S., Whyte, G., Woodcock, T., Mythen, M., & Montgomery, H. (2019). A multidisciplinary consensus on dehydration: definitions, diagnostic methods and clinical implications. *Annals of medicine*, **51**, 3-4

4. Masento, N. A., Golightly, M., Field, D. T., Butler, L. T., & van Reekum, C. M. (2014). Effects of hydration status on cognitive performance and mood. *The British journal of nutrition*, **111**(10), 1841–1852.

5. Miller H. J. (2015). Dehydration in the Older Adult. *Journal of gerontological nursing*, **41**(9), 8–13.

6. Edmonds, C. J., Foglia, E., Booth, P., Fu, C. H. Y., & Gardner, M. (2021). Dehydration in older people: A systematic review of the effects of dehydration on health outcomes, healthcare costs and cognitive performance. *Archives of gerontology and geriatrics*, **95**, 104380.

7. Elsamad, G., Mecawi, A. S., Pauža, A. G., Gillard, B., Paterson, A., Duque, V. J., Šarenac, O., Žigon, N. J., Greenwood, M., Greenwood, M. P., & Murphy, D. (2023). Ageing restructures the transcriptome of the hypothalamic supraoptic nucleus and alters the response to dehydration. *npj aging*, **9**(1), 12.

8. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., … Segal, E. (2015). Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, **163**(5), 1079–1094.

9. Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W. (2018), Digital Twin in manufacturing: A categorical literature review and classification, *IFAC-PapersOnLine*, **51**(11), 016-1022

10. https://www.iso.org/obp/ui/en/#iso:std:iso:23247:-1:ed-1:v1:en. Accessed 12 January 2024.

11. Wright, L., Davidson, S. (2020) How to tell the difference between a model and a digital twin. *Adv. Model. and Simul. in Eng. Sci.*, **7**, 13

12. Kamel Boulos, M. N., & Zhang, P. (2021). Digital Twins: From Personalised Medicine to Precision Public Health. Journal of personalized medicine, **11**(8), 745.

13. https://mijn.voedingscentrum.nl/nl/eetmeter/. Accessed 12 January 2024

14. de Gooijer, F.J., van Kraaij, A., Fabius, J., Hermsen, S., Feskens, E.J.M., & Camps, G. (2023), Assessing snacking and drinking behavior in Real-Life Settings: Validation of the SnackBox technology, *Food Quality and Preference*, **112**

15. https://apps.apple.com/nl/app/oneplanet-research/id1644843854. Accessed 13 March 2024

16. https://play.google.com/store/apps/details?id=com.imecint.oneplanet. Accessed 13 March 2024

17. Pérez-Idárraga, A. & Aragón-Vargas, L. (2010). Post-Exercise Rehydration: No Change in Diuresis from Water Ingested at Different Temperatures. *Medicina Sportiva*, **14**(2), 77-82.

18. Allegra, D., Battiato, S., Ortis, A., Urso, S., & Polosa, R. (2020). A review on food recognition technology for health applications. *Health psychology research*, **8**(3), 9297.

19. Procopio, A., Cesarelli, G., Donisi, L., Merola, A., Amato, F., & Cosentino, C. (2023). Combined mechanistic modeling and machine-learning approaches in systems biology - A systematic literature review. *Computer methods and programs in biomedicine*, **240**, 107681.

Measuring Farm Animal Behaviour

Method comparison to analyse the activity rhythm of dairy cows during early lactation

Marie Schneider[1,2], Kerstin Barth[1], Joanna Stachowicz[3], Eva Gallman[2] and Christina Umstätter[3]

[1] Johann Heinrich von Thünen Institute, Federal Research Institute for Rural Areas, Forestry and Fisheries, Institute of Organic Farming, 23847 Westerau, Germany, marie.schneider@thuenen.de

[2] University of Hohenheim, Center for Livestock Technology, 70599 Stuttgart, Germany

[3] Johann Heinrich von Thünen Institute, Federal Research Institute for Rural Areas, Forestry and Fisheries, Institute of Agricultural Technology, 38116 Braunschweig, Germany

Abstract

Several methods for analysing activity rhythms have been developed and compared in the past. However, the Degree of Functional Coupling (DFC) was never included in comparative studies. Therefore, we set out to compare the DFC with two other methods of analysing the circadian activity rhythm, Interdaily Stability (IS) and Spectral Entropy (SE), regarding the impact of oestrus and regrouping. Both oestrus and regrouping are known to affect the circadian activity rhythm of dairy cows. The study used data from two herds of German Holstein cows housed in a mirror imaged free stall barn. The days on which oestrus and regrouping occurred were documented. Accelerometers recorded cow activity from two weeks before expected calving until 82 days after calving. The DFC, IS and SE were calculated from 15-minute intervals of the activity using a sliding 7-day window. A linear mixed effect model was used to calculate the correlation between each pair of rhythmicity measures. Furthermore, the influence of oestrus and regrouping on each rhythmicity measure was analysed visually using boxplots. The correlation between the measures ranged from low to moderate, with the highest correlation found between DFC and IS, which were similarly interpreted as the alignment of the activity rhythm with the natural 24-hour cycle. The visual analysis indicated that all rhythmicity measures decreased during oestrus. Additionally, the DFC and the IS decreased during regrouping, while SE remained unaffected. In conclusion, DFC seems to be a reliable and comparable but also sensitive method for analysing the influence of oestrus and regrouping on the activity rhythms of dairy cows.

Introduction

In recent years, various measures of rhythmicity have been developed and used in different areas of biological rhythm research. Refinetti et al. [1] provided a guide for selecting a method for circadian rhythms analysis. They explained, when to use a Fourier Analysis, a Lomb-Scargle Periodogram, the Cosinor method or other methods [1]. However, they did not include non-parametric measures (other than the Kruskal-Wallis-Test) such as the Interdaily Stability (IS) or other measures, that include Fourier Analysis or the Lomb-Scargle Periodogram, but go further to classify the circadian rhythm, such as the Degree of Functional Coupling (DFC) or the Spectral Entropy (SE).

The DFC, which was developed by Sinz and Scheibe [2], analyses the alignment of animal activity rhythms with the natural 24-hour cycle by calculating the proportion of harmonic periods in all expressed periods. To date, the DFC has been used to analyse the circadian rhythms of various animals, including alpacas, Prezwalski horses, red deer and moufflons under free-ranging conditions [3] [4] as well as ewes under extensive conditions [5]. A pilot study also analysed the circadian activity rhythm of dairy cows milked by automatic milking systems using the DFC [6]. The SE is another rhythmicity measure, calculating the irregularity and complexity of activity rhythms [7]. It can be used as a frequency domain feature to improve the automatic classification of locomotor-associated sickness behaviour [8] or automatically classify cows' behaviour by comparing leg- and neck-mounted accelerometers [9]. A pilot study by McPherson et al. [10] used the SE to analyse the circadian rhythms of cows and how they are affected by cow-calf contact. Finally, the IS [11] which is a non-parametric indicator that measures the stability of activity rhythms between different days is a well-known measure of rhythmicity in human science, but has had less impact in animal sciences.

Various factors can affect the circadian activity rhythm of animals. For instance, changes in dairy cows' circadian activity rhythm were reported during oestrus, calving, or several diseases such as lameness, mastitis, or ruminal acidosis [12] [13]. Additionally, Wagner et al. [13] reported an effect of regrouping on cows' circadian activity rhythm. However, the effect of these influencing factors on the DFC, IS or SE, measured in dairy cows has not been previously studied.

The aim of this study was to compare three different rhythmicity measures (DFC, IS, SE) with respect the potential effects of oestrus and regrouping in dairy cows.

Animals, Housing and methods

The study was conducted on the research farm of the Thünen Institute of Organic Farming in Germany. Two dynamic herds were kept in the mirror imaged free stall barn. One herd consisted of an average of 40 polled German Holstein cows (28-48), while the other consisted of an average of 43 horned German Holstein cows (37-47). The cows were milked twice daily and fed with a total mixed ration, which was provided at the feeding table during the milking times. Cows were kept on pasture both day and night during the vegetative period. Around calving, they were held in single maternity pens until 4 ± 1 d after calving. As the cows were kept under their normal living conditions, without performing procedures deviating from standard husbandry and commercially available sensors were used, no ethical approval of the study was required.

Data collection and editing

The experiment was conducted from August 2020 to April 2021 and from August 2021 to June 2022. The study included all cows that calved between August and January or August and March. Since oestrus and regrouping can potentially affect the activity rhythms of cows they were recorded during the experimental period. The farm staff noted the date the cow was regrouped with the main herd, and the days of oestrus were recorded by either farm staff or the management system. The cows' activity data was collected at a frequency of 16 Hz using IceTags (Peacock Technology, Stirling, UK). The data was recorded from two weeks before expected calving until day 82 after calving to analyse the effect of regrouping in the herd after calving and the occurrence of oestrus events. As being on pasture affects the cows' activity, these data were excluded from the dataset.

The IceManager Software (Peacock Technology, Stirling, UK) was used to calculate the Motion Index and Steps from the activity data. These variables were used to identify incorrect data caused by technical issues. A technical issue was identified if either Steps or Motion Index were zero for more than 12 hours, as the cows had to move within that time when walking to the milking parlour twice a day. If a technical issue was identified, the entire day was excluded from the dataset.

Calculation of Rhythmicity

The Motion Index was used to analyse the activity rhythms, using the DFC, IS and SE. As the data were collected during European summer and winter time, they were converted from CEST and CET to GMT. Afterwards, the Motion Index was summed up to 15-minute intervals, following the method of [6] . To analyse a circadian activity rhythm, multiple consecutive days are necessary [4]. In a similar way to previous studies using the DFC, IS and SE, a sliding window of 7 days (today and the following 6 days) was used to calculate each of the rhythmicity measures [11][7][6].

The DFC analyses the alignment of the cows' activity rhythms with the natural 24-hour cycle by calculating the proportion of harmonic periods in all expressed periods. Harmonic periods are defined as those that can be obtained by dividing 24 h by an integer, resulting in 24 h, 12 h, 8 h, 6 h, 4.8 h etc. (all periods, that fit in with 24 h). The DFC can take on values between 0 and 1 and was calculated using the digiRhythm package [14]. First, the activity rhythm frequencies were extracted using a Lomb-Scargle-Periodogram [15][16], which is more suitable for uneven data than Fourier Transformation [17] as used by Sinz and Scheibe [2]. Subsequently, the Baluev method [18] was used to identify significant frequencies, as it is one of the most effective methods for calculating the false alarm probability [19]. Finally, the sum of significant harmonic periods was divided by the sum of all significant periods (harmonic and non-harmonic).

The IS is a non-parametric indicator that measures the stability of activity rhythms between different days [11][20]. It compares the activity pattern of each day with the average pattern across days. The higher the IS (ranging between 0 and 1), the more aligned the expressed rhythms are with the natural 24-hour cycle [20] . In this study, the IS was calculated using the nparACT package [21].

The SE measures the irregularity and the complexity of the activity rhythms [7]. A higher SE value (theoretical max. is infinite) indicates a more random and less predictable rhythm due to increased irregularity and complexity. In this study, the spectrum was extracted using an autocorrelation function. Additionally, the SE was calculated based on a Fourier Transformation. However, as the autocorrelation and Fourier Transformation-based SE did not differ, the autocorrelation-based SE was used. It was calculated using the R package tsfeatures [22] .

Data analysis

After calculating the rhythmicity measures, we excluded cows with less than 15 data points per lactation. The dataset included 68 cows, with 13 of which had two lactations, resulting in a dataset of 4208 datapoints spread over 81 cow datasets.

A linear mixed effects model (R package lme4, [23]) was used to analyse the linear correlation between each pair of rhythmicity measures. This method accounts for the data structure of repeated measurements. Following Christensen [24], one rhythmicity measure was set as the target and the other one as the fixed variable. The lactation number nested in the cow was used as the random effect. Afterwards, the coefficient of determination was calculated using the R package MuMIn [25]. The square root of this coefficient was used to analyse the correlation and was interpreted using the definition by Hinkle et al. [26]. Additionally, the effects of oestrus and regrouping on each rhythmicity measure were analysed visually using boxplots (using the R package ggplot, [27].

Results

The pairwise comparison of the rhythmicity measures showed correlation coefficients ranging from low to moderate. The correlation between DFC and IS was the highest at 0.60, followed by the correlation between IS and SE at 0.53. The correlation between DFC and SE was classified as low at 0.43.

The visual analysis indicated that oestrus and regrouping affected DFC and IS, whereas only oestrus had a visible effect on SE. The median DFC value for the entire dataset was 1.00 (IQR = 0.00, 1.00). During oestrus, it decreased to 0.18 (IQR = 0.00, 0.52), and during regrouping, it decreased to 0.58 (IQR = 0.00, 1.00). Similarly, the IS showed a median of 0.37 (IQR = 0.27, 0.44), which decreased to 0.24 (IQR = 0.18, 0.29) during oestrus and to 0.33 (IQR = 0.27, 0.39) during regrouping. In contrast, the SE decreased from a median of 0.98 (IQR = 0.96, 0.99) to 0.94 (IQR = 0.90, 0.97) during oestrus but was not affected by regrouping.

Discussion and conclusion

Among the three pairs of rhythmicity measures compared, DFC and IS had the highest correlation. They additionally showed a greater response to oestrus and regrouping compared to SE. However, the interpretation of DFC and IS, both of which show an alignment with the natural 24-hour cycle, is very similar. SE, on the other hand, deviates from this interpretation, as it refers to the irregularity and the complexity of a circadian rhythm. DFC showed the lowest median during oestrus, the second lowest median during regrouping. The DFC additionally had the highest median in the entire dataset, which indicates the highest alignment with the natural 24- hour cycle. However, the DFC is the only rhythmicity measure used in this study that includes ultradian rhythms and thus uses more information in its calculation process, which might explain the higher sensitivity to changes in the rhythm. Nonetheless, IS as a non-parametric rhythmicity measure, also seems to be a suitable method for detecting changes in the circadian activity rhythm of dairy cows caused by oestrus and regrouping. However, the SE measure demonstrated a weaker correlation with both methods. It is important to note that the SE measure has a different value range and interpretation compared to the other measures. An increase in DFC and IS indicates a higher alignment with the natural 24-hour cycle, while an increase in SE indicates greater irregularity and complexity of the rhythm.

The lower values in DFC and IS during oestrus and regrouping suggests a lower alignment with the natural 24-hour cycle on these days. However, the lower value in SE indicates a lower variability of the rhythm. This reduction could be explained by a lower behavioural variability, which animals are known to display under several conditions, such as during sickness or when exhibiting stereotypic behaviour (Miller et al., 2020). Given that during oestrus, certain behaviours such as mounting or chin resting on other cows and being mounted are frequently repeated [28], a decrease in behavioural diversity is reasonable. In contrast, in our visual analysis, SE was not affected by regrouping. This indicates that the variability of the cows' behaviour was not affected by regrouping, although their activity increases [29]. However, as DFC and IS had lower values during regrouping than in the entire dataset, the cows' circadian activity rhythm was influenced by this effect, even though the impact was lower than during oestrus.

In conclusion, oestrus altered all tested rhythmicity measures. Furthermore, DFC and IS were also affected by regrouping in our visual analysis. However, DFC appears to be more sensitive in detecting changes in the circadian activity rhythm of dairy cows, caused by oestrus and regrouping.

Acknowledgements

A great thank you to the farm and technical staff for their help in collecting the data.

References

1. Refinetti, R., Cornélissen, G., Halberg, F., 2007. Procedures for numerical analysis of circadian rhythms. Biological Rhythm Research, 275–325.
2. Sinz, R., Scheibe, K.M., 1976. Systemanalyse der multioszillatorischen Funktionsordnung im zirkadianen und ultradianen Frequenzbereich und ihr Indikationswert für Belastungswirkungen, dargestellt am Beispiel verschiedener Licht-Dunkel-Verhältnisse bei der Intensivhaltung von Schafen. Acta Biologica et Medica Germaniae 35, 465 - 414.
3. Scheibe, K.M., Berger, A., Langbein, J., Streich, W.J., Eichhorn, K., 1999. Comparative Analysis of Ultradian and Circadian Behavioural Rhythms for Diagnosis of Biorhythmic State of Animals. Biological Rhythm Research 30, 216–233.
4. Berger, A., Scheibe, K.-M., Michaelis, S., Streich, W.J., 2003. Evaluation of living conditions of free-ranging animals by automated chronobiological analysis of behavior. Behavior Research Methods, Instruments, & Computers 35, 458–466.
5. Nunes Marsiglio Sarout, B., Waterhouse, A., Duthie, C.-A., Candal Poli, C.H.E., Haskell, M.J., Berger, A., Umstatter, C., 2018. Assessment of circadian rhythm of activity combined with random regression model as a novel approach to monitoring sheep in an extensive system. Applied Animal Behaviour Science, 26–38.
6. Fuchs, P., Adrion, F., Shafiullah, A.Z.M., Bruckmaier, R.M., Umstätter, C., 2022. Detecting Ultra- and Circadian Activity Rhythms of Dairy Cows in Automatic Milking Systems Using the Degree of Functional Coupling—A Pilot Study. Frontiers in Animal Science 3, 839906.
7. Dowse, H.B., 2013. Maximum entropy spectral analysis for circadian rhythms: theory, history and practice. Journal of Circadian Rhythm 11.
8. Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., Krieter, J., 2020. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. Computers and Electronics in Agriculture 173, 105404.
9. Benaissa, S., Tuyttens, F.A.M., Plets, D., Pessemier, T. de, Trogh, J., Tanghe, E., Martens, L., Vandaele, L., van Nuffel, A., Joseph, W., Sonck, B., 2019. On the use of on-cow accelerometers for the classification of behaviours in dairy barns. Research in veterinary science 125, 425–433.
10. McPherson, S.E., Riaboff, L., Dissanayake, O., Sinnott, A., Cunningham, P., Kennedy, E. (Eds.), 2022. Effect of separation at weaning on the activity of cows and calves reared in a cow-calf contact system measured with accelerometer sensors, 8 pp.
11. Witting, W., Kwa, I.H., Eikelenboom, P., Miriam, M., Swaab, D.F., 1990. Alterations in the Circadian Rest-Activity Rhythm in Aging and Alzheimer's Disease. Biological Psychiatry 27, 563–572

12. Veissier, I., Mialon, M.-M., Sloth, K.H., 2017. Short communication: Early modification of the circadian organization of cow activity in relation to disease or estrus. Journal of dairy science, 3969–3974.
13. Wagner, N., Mialon, M.-M., Sloth, K.H., Lardy, R., Ledoux, D., Silberberg, M., Des Boyer Roches, A.d., Veissier, I., 2021. Detection of changes in the circadian rhythm of cattle in relation to disease, stress and reproductive events. Methods 186, 14–21.
14. Nasser, H.R., Schneider, M., Stachowicz, J., Umstätter, C., 2023. digiRhythm: Analyzing Animal's Rhythmicity. R package.
15. Lomb, N.R., 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and space science 39, 447–462.
16. Scargle, J.D., 1982. Studies in Astronomical time series analysis II. Statistical aspects of Spectral Analysis of unevenly spaed data. The Astronomical Journal 263, 835–853.
17. VanderPlas, J.T., 2018. Understanding the Lomb–Scargle Periodogram. The Astrophysical Journal Supplement Series, 1–28.
18. Baluev, R.V., 2008. Assessing the statistical significance of periodogram peaks. Monthly Notices of the Royal Astronomical Society 385, 1279–1285.
19. Süveges, M., Guy, L.P., Eyer, L., Cuypers, J., Holl, B., Lecoeur-Taïbi, I., Mowlavi, N., Nienartowicz, K., Blanco, D.O., Rimoldini, L., Ruiz, I., 2015. A comparative study of four significance measures for periodicity detection in astronomical surveys. Monthly Notices of the Royal Astronomical Society 450, 2052–2066.
20. Blume, C., Santhi, N., Schabus, M., 2016. 'nparACT' package for R: A free software tool for the non-parametric analysis of actigraphy data. MethodsX 3, 430–435.
21. Blume, C., Santhi, N., Schabus, M., 2017. nparACT: Non-Parametric Measures of Actigraphy Data. R package.
22. Hyndman, R., Kang, Y., Montero-Manso, P., O'Hara-Wild, M., Talagala, T., Wang, E., Yang, Y., 2023. tsfeatures: Time Series Feature Extraction. R package.
23. Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67, 1–48.
24. Christensen, R., 1996. Plane answers to complex questions: The Theory of Linear Models. Springer, New York.
25. Bartoń, K., 2023. MuMIn: Multi-Model Inference. R package.
26. Hinkle, D.E., Wiersma, W., Jurs, S.G., 2003. Applied Statistics for the behavioral sciences, 5th ed. Houghton Mifflin, Boston, MA, USA, p.756.
27. Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer Verlag New York.
28. Kerbrat, S., Disenhaus, C., 2004. A proposition for an updated behavioural characterisation of the oestrus period in dairy cows. Applied Animal Behaviour Science 87, 223–238.
29. Torres-Cardona, M.G., Ortega-Cerrilla, M.E., Alejos-de la Fuente, J.I., Herrera-Haro, J., Peralta Ortíz, j.G., 2014. Effect of regrouping Holstein cows on milk production and physical activity. Journal of Animal & Plant Sciences 22, 3433–3438.

Equine social proximity according to space availability using ultra-wideband technology

L. Torres Borda[1], U. Auer[2],[†] and F. Jenner[1],[†,*]

[1] **Equine Surgery Unit, Department of Companion Animals and Horses, University of Veterinary Medicine Vienna, Vienna, Austria, Florien.Jenner@vetmeduni.ac.at, Laura.Torres-Borda@vetmeduni.ac.at**

[2] **Anaesthesiology and Perioperative Intensive Care Medicine Unit, Department of Companion Animals and Horses, University of Veterinary Medicine Vienna, Austria, Ulrike.Auer@vetmeduni.ac.at**

[†]shared **last author**

Keywords

Social behaviour, distance monitoring, applied animal sciences, horse husbandry, ultra-wideband sensor

Background

In their natural habitat, horses form stable social groups characterised by enduring bonds between individuals and consistent dyadic interaction patterns, resulting in infrequent and typically mild, ritualised aggressive behaviours [1]–[5]. However, only close affiliative partners are permitted within a horse's personal space, while intrusion by non-affiliates triggers aggressive behaviour. Thus, proximity associations, encompassing both the closeness and duration of interactions, can serve as indicators of equine social relationships. While affiliative interactions are defined by close proximity of longer duration, agonistic interactions prompt immediate separation following proximity.

Equine social behaviour remains largely unchanged by domestication [6]. However, the environment of domestic horses significantly differs from naturalistic conditions. Domesticated horses face social challenges, including changing group compositions, inadequate enclosure sizes and high population densities. As equine social tolerance is influenced by space and resource availability, reduced space per animal is linked to increased aggression. In contrast, previous studies show that enlarging enclosure sizes and reducing population density decreases the frequency of aggressive and submissive behaviours [7]–[10], mitigating horses' stress and the risk of injuries [10], [11].

Therefore, the present study utilises ultrawide-band sensors to assess how space availability influences a horse group's proximity behaviour with the aim to use interindividual distance measurements, with the corresponding spatial relationships and dynamics, as a tool for studying equine social behaviour, optimising group composition, and enhancing welfare.

Methods

Horses and horse management conditions

Nine mixed-breed horses, four geldings and five mares, aged 9-32 years (mean age 22.1 years), were included in this study. The horses were located at an equine sanctuary and housed in individual box stalls, bedded with shavings, with daily paddock (450m2, 30m x 15m, sand and gravel surface) or pasture (2682m2, (40x65), grass surface) turn-out for 4-6 hours (appr. 07:30 am to 12:45 pm). Horses had ad libitum access to water and were fed a hay-based diet. In addition to the ad libitum access to hay provided in one hay feeder with eight feeding places on the paddock, all horses received two additional servings in the stable in the afternoon and at night. In the pasture, they had access to grass but no hay.

The group was tracked twice, once in the paddock (10 days) and once, 6 months later, in the field (6 days). A new horse was added to the group between the first and second tracking periods. Before each period tracking period, the group composition has been stable for at least two months. Horses were tracked for the entire duration of their turn-out, resulting in a total tracking duration of 52h in the paddock and 36h in the field.

Wearable ultra-wideband (UWB) sensor and proximity measurements

An ultra-wideband (UWB) wireless real-time location system (RTLOC®) with a resolution of 1 measurement per second and a measurement range of a minimum of 5 centimetres and a maximum of approximately 120 meters was used for proximity tracking in this study. The UWB radiofrequency technology does not require fixed infrastructure. UWB devices join an ad-hoc network and start ranging. By measuring the time-of-flight (TOF) of UWB signals, the distance between two transceivers is obtained using the two-way-ranging (TWR) method. A gateway device was associated with a computer during tracking. The UWB sensor (8cm x 7cm x 2cm, 95g) is adapted for field experiments due to its small size, lightweight and waterproof case.

UWB sensor – validation under laboratory conditions

The accuracy of UWB distance measurements was first tested under laboratory conditions by placing two sensors at set reference distances (1, 2, 3, 5 and 10 meters) apart, recording the distance measurements for 15 minutes for each distance (75 minutes in total, x5 set distances, 15minutes/distance) and calculating the spatial accuracy compared to the reference distances and the coefficient of variation (temporal stability) over the 15 min. In addition, the influence of the spatial arrangement of the UWB-tags, especially the placement of the transceiver, on distance measurements was assessed by placing 7 sensors 1 metre distant from an 8th sensor for 15 minutes and then turning the 8th sensor 90 degree every 15 minutes and measuring the inter-sensor distances for 15 minutes each for the 4 different orientations of the transceiver relative to the other sensors (toward, 90 degree to the left/right, 180 degree in the other direction).

UWB sensor – validation under field conditions

The wearable UWB sensors were easily affixed to the horses' halters and were worn for an acclimatisation period of at least 10 days prior the study without evident impact on their social behaviour.

The measurement accuracy of the UWB technology under field conditions in a horse group was validated by comparing timestamped UWB sensor measurements with proximity data extracted from corresponding recordings [12]. To conduct visual measurements, the individual proximity sensors were positioned within the calibrated images [13], allowing estimation of real-world distances through a two-stage process. Initially, the distance in pixels between these sensors was assessed. Subsequently, pixel distances were translated into real-world distances. This involves utilising a homography-based approach, assuming the movement of sensors within a plane situated 1.8 meters above the ground, to compute the actual distances in the real world [14].

Social proximity measurements between horses

The average (mean +/- s.d.) distance between all horses and between horse dyads while they were turned out together in a paddock and a field was calculated for the two observation periods. To evaluate the dispersion of horse groups according to the available space and the number of individuals within that space, the metric of the average distance between horses in meters divided by the available space per horse was calculated (=mean interindividual distance of the horse group in metres/(total space in m2/total horse number). In addition, the percentage of time spent at less than 3 meters was calculated by counting the total measurements recorded at that distance range and dividing this count by the total number of distance measurements. Based on the literature, this cut-off distance was used to define social proximity ([15], [16]). The nearest neighbours were determined by the percentage of time spent within a distance of less than 3 meters.

Statistical analysis

The Wilcoxon signed-rank test was used for the sensor validation under laboratory and field conditions.

To exclude any erroneous measurements caused by momentaneous poor connection quality, interference by another horse moving between two transceivers or similar obstructions, we excluded all measurements lower to 5 centimetres (impossible due to the sensor location behind and between the ears) and greater than the theoretical possible distance between two horses according to the paddock dimensions. All measurements recorded between horses at a distance lower than 3 meters from the paddock hay feeder were excluded in order to avoid taking into consideration proximity caused by the limited resources available for eating.

Statistical analyses were performed in R, v. 4.2.2 [17]. Measurements of each dyadic tag pair were averaged for downstream analysis. Pearson correlations were conducted. A non-parametric approach was chosen because the data did not follow normality of distribution [18]. Comparisons of means were done with Fisher-Pitman permutation test using the package coin [19]. P-values were calculated via Monte Carlo sampling with 1000.

Results

UWB sensor validation under laboratory and field conditions

Results showed that the average difference between the measured and the actual distances was 34.5 centimetres. The coefficient of variation over the 15 min measurement periods was < 3.3%, confirming the temporal stability of the measurements.

The spatial arrangement of the sensor's transceiver relative to the other sensors had a significant influence on distance measurements (p <0.001). The distance recorded was closest to the actual set distance (one metre) when the transceiver faced the other sensor (median: 95cm) and farthest when the transceiver was turned away (median: 117cm).

Validation under field conditions based on the comparison of sensor measurements and proximity data extracted from video surveillance revealed a correlation of 0.83, p <0.001, thus confirming the convergent validity.

Paddock versus field - Descriptive measurements

The average distance between horse pairs in the paddock was 7.27 meters (± 4.90 meters s.d.) and 13.47 meters (± 9.85 meters s.d.) in the field. Considering that the field is 6.4 times bigger than the paddock, the mean distance between horses measured in the field is only 1.85 times larger than the mean distance between horses measured in the paddock. The average distances (overall and between each horse pair) between the paddock and field groups were significantly different ($Z = 575.44$, $p < 0.001$). The ratio of the average distance between horses in meters divided by the available space per horse was higher in the paddock (0.13) compared to the field (0.05).

Nearest neighbours

Average distances in meters between horse pairs and percentage of time spent at less than 3 meters from each other allowed the identification of affiliative pairs. Affiliative partners (e.g. horse pairs 1-3 and 5-8) that stayed close to each other and spent a large percentage of time at a distance <3m from each other on the paddock, remained also close for large proportion of the time in the field. In contrast, horse dyads (e.g. horses 3-8) that remained further apart than the average interindividual distance of the group and rarely spent time at a distance <3m on the paddock, stayed even further apart and spent less time at a distance <3 m in the field.

Overall, the comparison between the time that horses spent at distances exceeding 3 meters from each other in the field and paddock areas showed not significant difference ($Z = 1.82$, $p = 0.074$). However, the percentage of time that horses spent closer than 3 metres from a conspecific was significantly different in the field and the paddock ($Z = 5.07$, $p < 0.001$).

Centrality analyses revealed differences in closeness centrality. Closeness centrality was higher in the paddock than in the field, implying that horses are, on average, more directly connected in the paddock. This do not imply actual closeness but indicates a potential for more direct or efficient pathways of interaction within the paddock group. All descriptive measurements obtained through the use of ultra-wideband sensors offer a global view of close associates and the group dynamic. The nearest-neighbour-based grouping (Figure. 5) identified from these results matched with onsite observations from caretakers.

Figure 1. Adjacency matrix of average distances (in meters) between horse pairs in the paddock.

	Horse 1	Hose 2	Horse 3	Horse 4	Horse 5	Horse 6	Horse 7	Horse 8
Horse 1		8.3	4.2	8.1	9.4	9.7	8.4	9.6
Horse 2			7.4	7	7.4	8.1	6.3	7.3
Horse 3				7.6	8.2	8.7	7.9	8.5
Horse 4					7.1	7.4	6.8	7.1
Horse 5						4.6	6.9	3.6
Horse 6							7.1	4.6

Horse 7								6
Horse 8								

Figure 2. Adjacency matrix of average distances (in meters) between horse pairs in the field.

	Horse 1	Hose 2	Horse 3	Horse 4	Horse 5	Horse 6	Horse 7	Horse 8	Horse 9
Horse 1		14.2	6.6	10.8	12.6	15.8	13	13.2	17.8
Horse 2			14.7	12.2	14.9	17.1	11.4	16.1	14.5
Horse 3				12.8	16	19.3	13.9	16.3	16.5
Horse 4					9.5	12.3	11.6	10.4	15.6
Horse 5						10	12.4	5.5	16.6
Horse 6							14.1	10	17.5
Horse 7								11.6	16.4
Horse 8									19
Horse 9									

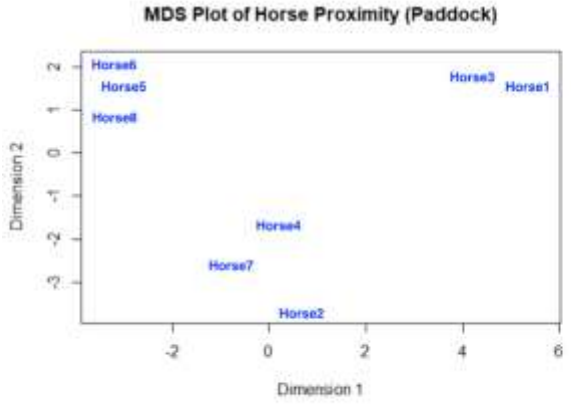
Figure 3. Percentage of time spent at less than 3 meters between horse pairs in the paddock.

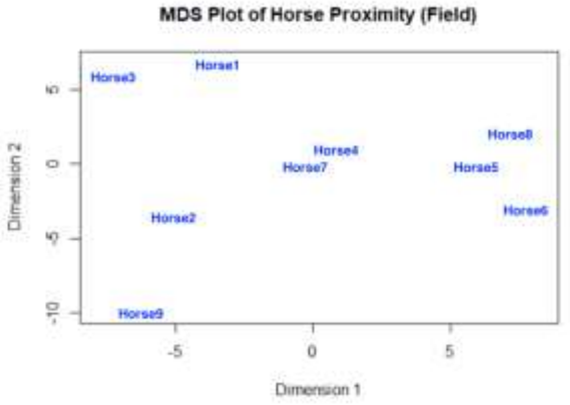
	Horse 1	Hose 2	Horse 3	Horse 4	Horse 5	Horse 6	Horse 7	Horse 8
Horse 1		18.4	58.9	7	13.3	3.2	8.4	5.7
Horse 2			20.6	15.8	22.6	10.1	21.7	17.1
Horse 3				11.9	20.9	7.3	11.4	11.4
Horse 4					16.5	7.5	16.5	13.3
Horse 5						41.4	16.9	57.3
Horse 6							14.4	45.9
Horse 7								28.1
Horse 8								

Figure 4. Percentage of time spent at less than 3 meters between horse pairs in the field.

	Horse 1	Horse 2	Horse 3	Horse 4	Horse 5	Horse 6	Horse 7	Horse 8	Horse 9
Horse 1		7.8	52	4.6	1.7	3.4	1.4	0.7	0
Horse 2			3.8	4.4	1.3	0.8	17.2	1.2	1.5
Horse 3				5.4	1.2	0.5	4.8	1.6	0.4
Horse 4					7	4.2	10.9	7.3	1.9
Horse 5						18.2	2	37.9	3
Horse 6							5.4	14	5.5
Horse 7								14.3	0.9
Horse 8									0.2
Horse 9									

Figure 5. Multidimensional scaling plot of the observed group in both paddock and field settings





Discussion

The present study presents the reliability and precision of an ultra-wideband sensor under both controlled laboratory and real-world field conditions in the context of animal behaviour research. This is evidenced through tests checking for consistency in measurements and limitations of the system and by the correlation between UWB-proximity measurements and video analysis.

Preliminary analysis first showed a tendency for horses in a stable group to increase inter-individual distances in field enclosures compared to their proximity in paddocks. However, the group dispersion was not observed as being proportional to the much larger field dimensions. A difference was found by calculating the ratio of the mean distance between horses per available square meter per horse. Indeed, this revealed a lower ratio in the field, meaning that horses are c on average loser together relative to the available space.

In addition, consistency in social dynamics was observed across enclosure settings. Similar patterns were noted in affiliative partners remaining close to each other for a high percentage of the time, both in the paddock and in the field.

Previous investigations regarding the importance of space availability for horses' groups have resulted in a minimum recommended space allowance of 331-477 m2 per horse [8], [10], [20] to minimize stress, agonistic interactions and the risk of injuries [10], [11]. In the current study, space availability per horse was 56 square metres in the paddock and 298 square metres in the field, however the rate of agonistic interactions was low.

Conclusions

The application of UWB technology in this study facilitated continuous measurement of distances between horse dyads in a group for extended periods, overcoming numerous constraints associated with traditional observational approaches. The UWB sensor provided objective quantitative data on interindividual distances, enabling comparability and repeatability within and between different labs and thus assessing the influence of environmental and management conditions on equine spatial relationships and dynamics as potential indicators of altered welfare and quality of life.

References

1. K. Krueger and J. Heinze, "Horse sense: social status of horses (Equus caballus) affects their likelihood of copying other horses' behavior," *Anim. Cogn.*, vol. 11, no. 3, pp. 431–439, 2008, doi: 10.1007/s10071-007-0133-0.

2. C. Sankey, M.-A. Richard-Yris, H. Leroy, S. Henry, and M. Hausberger, "Positive interactions lead to lasting positive memories in horses, Equus caballus," *Anim. Behav.*, vol. 79, no. 4, pp. 869–875, 2010, doi: 10.1016/j.anbehav.2009.12.037.

3. M. Hannan, I. Draganova, and L. Dumbell, “Factors affecting mutual grooming and play behaviour in a group of domestic horses (Equus caballus),” *BSAP Occas. Publ.*, vol. 35, pp. 193–197, 2006, doi: 10.1017/s0263967x00042701.

4. H. Sigurjónsdóttir and H. Haraldsson, “Significance of Group Composition for the Welfare of Pastured Horses,” *Animals*, vol. 9, no. 1, p. 14, 2019, doi: 10.3390/ani9010014.

5. K. A. HOUPT and T. R. WOLSKI, “Stability of equine hierarchies and the prevention of dominance related aggression,” *Equine Vet. J.*, vol. 12, no. 1, pp. 15–18, 1980, doi: 10.1111/j.2042-3306.1980.tb02288.x.

6. J. W. Christensen, T. Zharkikh, J. Ladewig, and N. Yasinetskaya, “Social behaviour in stallion groups (Equus przewalskii and Equus caballus) kept under natural and domestic conditions,” *Appl. Anim. Behav. Sci.*, vol. 76, no. 1, pp. 11–20, 2002, doi: 10.1016/s0168-1591(01)00208-8.

7. K. Majecka and A. Klawe, “Influence of Paddock Size on Social Relationships in Domestic Horses,” *J. Appl. Anim. Welf. Sci.*, vol. 21, no. 1, pp. 8–16, 2018, doi: 10.1080/10888705.2017.1360773.

8. B. Flauger and K. Krueger, “Aggression level and enclosure size in horses (Equus caballus),” *Pferdeheilkunde Equine Med.*, vol. 29, no. 4, pp. 495–504–495–504, 2013, doi: 10.21836/pem20130404.

9. M. Pierard, P. McGreevy, and R. Geers, “Effect of density and relative aggressiveness on agonistic and affiliative interactions in a newly formed group of horses,” *J. Vet. Behav.*, vol. 29, pp. 61–69, 2019, doi: 10.1016/j.jveb.2018.03.008.

10. J. K. Suagee-Bedore, D. R. Linden, and K. Bennett-Wimbush, “Effect of Pen Size on Stress Responses of Stall-Housed Horses Receiving One Hour of Daily Turnout,” *J. Equine Vet. Sci.*, vol. 98, p. 103366, 2021, doi: 10.1016/j.jevs.2020.103366.

11. N. Morgan and H. Randle, “Personal space requirements of mares versus geldings (Equus caballus): welfare implications and visual representation of spatial data via Spatial Web diagrams),” *BSAP Occas. Publ.*, vol. 35, pp. 203–206, 2006, doi: 10.1017/s0263967x00042725.

12. P. Düking, F. K. Fuss, H.-C. Holmberg, and B. Sperlich, “Recommendations for Assessment of the Reliability, Sensitivity, and Validity of Data Provided by Wearable Sensors Designed for Monitoring Physical Activity,” *JMIR mHealth uHealth*, vol. 6, no. 4, p. e102, 2018, doi: 10.2196/mhealth.9341.

13. Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000, doi: 10.1109/34.888718.

14. R. Hartley and A. Zisserman, “Multiple View Geometry in Computer Vision,” 2004, doi: 10.1017/cbo9780511811685.

15. F. Hildebrandt, K. Büttner, J. Salau, J. Krieter, and I. Czycholl, “Proximity between horses in large groups in an open stable system – Analysis of spatial and temporal proximity definitions,” *Appl. Anim. Behav. Sci.*, vol. 242, p. 105418, 2021, doi: 10.1016/j.applanim.2021.105418.

16. K. Krueger, B. Flauger, K. Farmer, and C. Hemelrijk, “Movement initiation in groups of feral horses,” *Behav. Process.*, vol. 103, pp. 91–101, 2014, doi: 10.1016/j.beproc.2013.10.007.

17. *RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL*. 2020. [Online]. Available: http://www.rstudio.com/

18. P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive Statistics and Normality Tests for Statistical Data," *Ann. Card. Anaesth.*, vol. 22, no. 1, pp. 67–72, 2019, doi: 10.4103/aca.aca_157_18.

19. T. Hothorn, K. Hornik, M. A. van de Wiel, and A. Zeileis, "A Lego System for Conditional Inference," *Am. Stat.*, vol. 60, no. 3, pp. 257–263, 2006, doi: 10.1198/000313006x118430.

20. F. Hildebrandt, K. Büttner, J. Salau, J. Krieter, and I. Czycholl, "Area and Resource Utilization of Group-Housed Horses in an Active Stable," *Animals*, vol. 11, no. 10, p. 2777, 2021, doi: 10.3390/ani11102777.

Olfactory Conditioning to Reduce Stress in Farm Animals – A Possible Experimental Set-up

J. Stenfelt[1], V. Bombail[2], H. Sassner[1], B. Forkman[3], A.M. Pálsdóttir[4] and M.V. Rørvang[1]

[1]Department of Biosystems and Technology, Swedish University of Agricultural Sciences, Alnarp, Sweden. [2]Animal and Veterinary Sciences, Scotland's Rural College, Edinburgh, UK. [3]Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Department of People and Society, Swedish University of Agricultural Sciences, Alnarp, Sweden. mariav.rorvang@slu.se

Background

Unlike other sensory modalities, the olfactory system projects directly onto key areas of the limbic system [1] involved in, for example, emotional processing, memory formation and retrieval, behavioural modulation, and initiation of the stress response. Research in human psychology shows that the hedonic perception of odours can modulate affective states [2,3]. Odour hedonics can be influenced by learning mechanisms such as associative learning, and olfactory conditioning is well-studied across species, usually by pairing an odour with an aversive stimulus such as electric shocks [4]. However, it has also been shown that odours can be successfully paired with pleasant experiences [5,6,7]. To our knowledge, the influence of appetitive olfactory conditioning on affective states of non-human animals has so far only been studied in rats [5]. Still, it is plausible that conditioned odours can modulate affective states in other mammals with well-developed olfactory systems, for example, farm animals like cattle and horses.

When deciding on an unconditioned stimulus, careful consideration should be given to the unconditioned response of the animals and how well it aligns with the goal of the conditioning. If the goal is to alleviate stress from a negative experience eliciting fear, the unconditioned stimulus should ideally evoke a positive and calm response in the animal. While food is a reliably positive stimulus, it is generally associated with high arousal [8], and thus possibly more appropriate for counteracting negative low-arousal mood states. Conversely, tactile stimulation facilitated by livestock brushes has been proposed as a positive low-arousal stimulus in cattle [8] and, if given the opportunity, both cattle and horses engage in and enjoy brushing [9]. Brush use has, at least in cattle, been proposed to be a low-resilience behaviour [10,11] associated with improved welfare and, thus, increasingly expressed by individuals who are more likely to be in a positive affective state. Moreover, the tactile stimulation from livestock brushes can be provided without human interference, allowing animals to interact with the brushes when motivated and without the added stress of handling or being removed from the herd. Thus, livestock brushes hold many qualities that could make them a suitable unconditioned stimulus in the context of olfactory conditioning with the aim of stress relief. In this comparative study, we aim to investigate if horses and dairy cattle can learn to associate the scent of lavender with the pleasant experience of tactile stimulation from voluntary interactions with livestock brushes. If so, we aim to answer whether this positive odour association can modulate affective states in stressful situations such as handling and transport.

Method

In two parallel experiments, 42 horses and cattle are divided into three treatment groups (T1-3, see Figure 1) balanced for sex and age. We start by conducting a temperament test battery [12] with the primary goal of assessing the emotional reactivity of individual animals, as this may be used to explain some of the individual variations in later tests. As part of the temperament test, we measure tactile sensitivity with monofilaments [13], and we further aim to investigate whether tactile sensitivity affects if and how individual animals interact with the livestock brushes (the unconditioned stimulus), which, in turn, is likely to affect the conditioning. Whether an individual likes or dislikes the scent of lavender (the neutral stimulus), is also expected to affect the conditioning and the direction in which the odour might modulate the affective state (positive or negative). We perform an odour preference test containing lavender and orange to account for individual preference. The two odours have been chosen as previous research indicates that at least horses voluntarily approach and sniff both odours and that no species-specific preference for one over the other exists [14]. The brushes are then installed in the home environment of the three treatment groups at a brush-to-horse/cattle ratio of 1:10. Two of the treatment groups (T1-2) are given access to lavender-scented brushes, and the third (T3) is given access to unscented brushes. The brushes remain available to the animals throughout the experimental period (approx. four months), but the first four weeks are considered the conditioning period (see Figure 2). To test whether the animals' interactions with scented brushes have created a positive odour association, we repeat the odour preference test to see if their preference has changed post-conditioning. Odour preference has previously been used to measure appetitive olfactory conditioning in rats [5] and we argue that if individuals with access to lavender-scented brushes to a greater degree change their preference to lavender over orange, this suggests that the lavender odour has been positively charged through conditioning. Based on the same study in rats [5], we further plan to test the effect of the conditioned odour on the affective state of the animals through short-term exposures to the experimental odours (lavender, orange) and an odourless control, for which qualitative behavioural assessments (QBA) will be performed.

To test whether a positively associated odour can modulate affective states, we subject the animals to a handling (low-stress) test and a transport (high-stress) test while exposed to their conditioned odour (see Figure 1). The tests are designed to reflect potentially stressful but commonly occuring situations that most horses and cattle are subjected to and to measure the effect of the conditioned odour on mild as well as intense stressors. Of the two treatment groups that are given access to lavender-scented brushes and, thus, have the opportunity to form a positive odour association, one (T1) undergoes the stress tests while exposed to the conditioned odour. The other group (T2) undergoes the stress tests without the conditioned odour to control for the potential effects of the conditioning. The third group (T3) that is given access to unscented brushes and, thus, has no opportunity to form an odour association, will also be tested while exposed to the conditioned (or in this case, unconditioned) odour to control for potential innate effects of lavender odour or odour presence in general (e.g., through odour masking). Following a four-week reconditioning period, we perform a judgement bias task (JBT) to assess if the conditioned odour can induce a positive affective state in the absence of stressors, and conclude whether a positive odour association can be restored following stressful events.

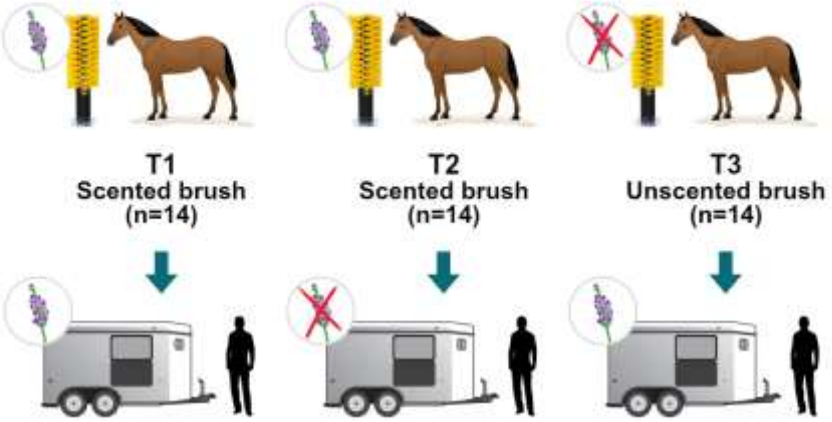
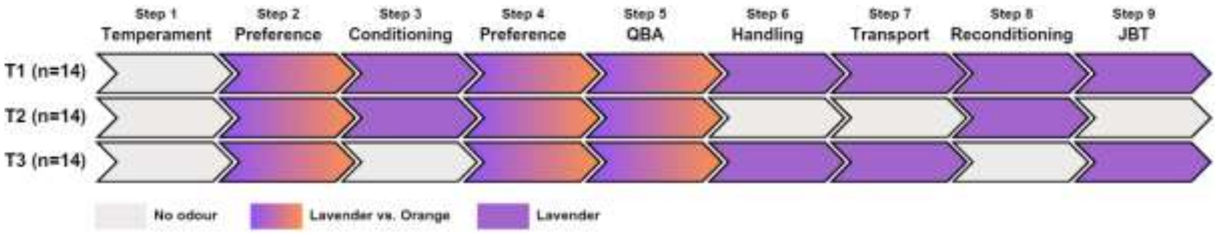


Figure 1. Horses in the three treatment groups, two with lavender-scented brushes (T1 and T2) and one with unscented brushes (T3), and their respective treatments in the following stress tests. Adapted from [15].



	Step 1 Temperament	Step 2 Preference	Step 3 Conditioning	Step 4 Preference	Step 5 QBA	Step 6 Handling	Step 7 Transport	Step 8 Reconditioning	Step 9 JBT
T1 (n=14)	⬜	▨	🟪	▨	▨	🟪	🟪	🟪	🟪
T2 (n=14)	⬜	▨	🟪	▨	▨	⬜	⬜	🟪	⬜
T3 (n=14)	⬜	▨	⬜	▨	▨	🟪	🟪	⬜	🟪

⬜ No odour ▨ Lavender vs. Orange 🟪 Lavender

Figure 2. The sequential order and the odour exposure for each treatment group during the experimental phases.

Practical implications

If horses and cattle form positive odour associations, the same may be true for other mammalian species whether farm, zoo, or companion animals. Moreover, if positive odour associations can be used for mitigating fear and stress, the application of olfactory conditioning can be extended beyond handling and transportation to include other stressful situations and management procedures. This includes, for example, separation and early weaning, social isolation, hoof trimming or farriery, veterinary procedures, and slaughter. Finally, if the positive odour association can be maintained or restored even following negative experiences during odour exposure, olfactory conditioning may prove an inexpensive and sustainable route towards improving the welfare of animals kept in large commercial systems, in a range of welfare-compromising situations.

Ethical statement

The experiments connected to this project are conducted at private farms in Sweden. Prior to the experimental start, all procedures and details of the experiments are evaluated and ethical permits are obtained from the Swedish Board of Agriculture, Linköping, Sweden: ID 5880 is approved and ID 6201 is under evaluation.

References

1. Lledo P. M., Gheusi G., Vincent J. D. (2005). Information processing in the mammalian olfactory system. *Physiological Reviews,* **85**(1):281–317.
2. Alaoui-Ismaïli, O., Vernet-Maury, E., Dittmar, A., Delhomme, G. & Chanel, J. (1997). Odor Hedonics: Connection With Emotional Response Estimated by Autonomic Parameters. *Chemical Senses,* **22**(3):237–248.
3. Knasko, S.C. (1992). Ambient odor's effect on creativity, mood, and perceived health. *Chemical Senses,* **17**(1):27–35.
4. Johnson, L.R., McGuire, J., Lazarus, R. & Palmer, A.A. (2012). Pavlovian fear memory circuits and phenotype models of PTSD. *Neuropharmacology,* **62**(2):638-646.
5. Bombail, V., Jerôme, N., Lam, H., Muszlak, S., Meddle, S.L., Lawrence, A.B. & Nielsen, B.L. (2019). Odour conditioning of positive affective states: Rats can learn to associate an odour with being tickled. *PLoS ONE,* **14**(6):e0212829.
6. Kippin, T.E. & Pfaus, J.G. (2001). The development of olfactory conditioned ejaculatory preferences in the male rat: I. Nature of the unconditioned stimulus. *Physiology & Behavior,* **73**(4):457-469.
7. Gelez, H., Archer, E., Chesneau, D., Campan, R. & Fabre-Nys, C. (2004). Importance of Learning in the Response of Ewes to Male Odor. *Chemical Senses,* **29**(7):555-563.
8. De Oliveira, D. & Keeling, L. (2018). Routine activities and emotion in the life of dairy cows: Integrating body language into an affective state framework. *PLoS ONE,* **13**(5):e0195674.
9. Lansade, L., Lemarchand, J., Reigner, F., Arnould, C. & Bertin, A. (2022). Automatic brushes induce positive emotions and foster positive social interactions in group-housed horses. *Applied Animal Behaviour Science,* **246**:105538.
10. Lecorps, B., Welk, A., Weary, D.M. & von Keyserlingk, M.A.G. (2021). Postpartum Stressors Cause a Reduction in Mechanical Brush Use in Dairy Cows. *Animals,* **11**(11):3031.
11. Mandel, R., Whay, H.R., Nicol, C.J., Klement, E. (2013). The effect of food location, heat load, and intrusive medical procedures on brushing activity in dairy cows. *Journal of Dairy Science,* **96**(10):6506–6513.
12. Lansade, L. & Simon, F. (2010). Horses' learning performances are under the influence of several temperamental dimensions. *Applied Animal Behaviour Science,* **125**(1-2):30-37.
13. Lansade, L., Pichard, G. & Leconte, M. (2008). Sensory sensitivities: Components of a horse's temperament dimension. *Applied Animal Behaviour Science,* **114**(3-4):534-553.
14. Rørvang M.V., Nicova, K. & Yngvesson, J. (2022). Horse odor exploration behavior is influenced by pregnancy and age. *Frontiers in Behavioral Neuroscience*, **16**:941517.

15. Stenfelt, J., Sassner, H., Bombail, V., Pálsdóttir, A.M. & Rørvang, M.V. (2023). *Olfactory conditioning for mitigating stress in horses* [Poster presentation]. ASAB Winter Meeting 2023 Animal Cognition: Pure to Applied, 13-14th December, Edinburgh, UK.

An investigation of several methods to monitor behaviour of housed dairy cows

W. Ouweltjes[1] and B. Loke[2]

1Wageningen Livestock Research, Wageningen, the Netherlands, 2 Noldus Information Technology BV, Wageningen, the Netherlands

Abstract

We monitored cows with multiple sensors: cameras, UWB tracking sensors, detection gates, accelerometers and pre-programmed pedometers. To retrieve information from cameras we have developed a software tool that can detect cows and 4 key points for each cow in snapshots of video footage. We investigated the insights in behaviour that can be obtained from data obtained from the different sensors and assess the practicalities for each of them and explore potential farm management applications.

Short paper text

Traditionally farmers visually assess the state of the animals for which they are responsible, and take action when animals need specific care. However, this is time consuming (particularly with large herd size), subjective and also requires expertise. Moreover, it is not realistic to do this continuously. Therefore, tools have been developed that enable automated measurement of aspects of behaviour and detect abnormalities in an automated way. For management purposes such detections only make sense when they can be interpreted (either stand alone or in combination with other measured parameters such as milk yield) and there is clarity about perspectives for action. The majority of currently available tools are based on accelerometry and focus on time budgets for some main activities (eating, resting, ruminating, movement/activity). For several reasons the set of behaviours monitored is limited, and the data recorded is usually a summary of the underlying raw data that was measured by the sensor. Moreover, the approach largely ignores the fact that farm animals are usually kept in groups, and interact with their herd mates. This hampers further development of automated measurement of behaviour and retrieval of interpretable features, both for management and research purposes.

Due to developments in sensor technology it is nowadays possible to accurately track animals indoors with the use of UWB technology [1], e.g. using Sewio Leonardo tags. Noldus Information Technology BV has developed an adaptor with which these tags can be attached to collars that dairy cows usually wear, and developed the TrackLab software to further process and interpret the data. However, this tool is developed for research purposes and not for farm management. Although position is linked to behaviour in cattle barns (with separate functional areas for feeding, drinking and resting), cows can both lie down or stand in a cubicle, be close to a drinker without drinking and be close to the feeding rack without eating. Moreover, when lying down they can take different poses and experience different degrees of comfort. When two or more animals are in close proximity they can interact in multiple ways (or not at all). Thus, knowledge of where animals are at specific time points only to a limited extent provides information about what they are doing and does not take into account qualitative aspects of their behaviours. An advantage of using position-information for animals in groups is that interactions between cows can be studied, which is impossible when using e.g. accelerometers only. Moreover, it can provide insights in the usage of e.g. individual cubicles, feeding places etc. and thus provide valuable information about the environment in which e.g. cows are kept. For monitoring animal positions throughout time from video footage automated object detection can be applied. When combined with detection of key points (definable anatomical elements of the body, see Cao et al. [2] for examples) for each individual pose estimation becomes possible. Both tools for automated object detection and pose estimation in video footage have shown a huge progress recently. This has enabled development of tools to automatically detect animals and their poses with sufficient accuracy. In principle, automated image processing can provide all information that can be retrieved from UWB tracking systems but also provide qualitative information about the behaviours performed and information about the pose/orientation of the cows. Correct identification can be challenging, but is a necessity when the information is to be used for management purposes. Identification errors also cause errors in tracking. All sensor applications benefit from the steady increase of computing power and data storage capacity.

At Dairy Campus, a research facility of Wageningen Livestock Research, an infrastructure was created in 7 barn units where dairy cows (16 in each unit) are tracked in their home pens with Sewio Leonardo tags with a resolution of 1 Hz. Because the study is purely observational approval from the Wageningen Research ethical committee was deemed unnecessary. The raw location data was processed with Noldus TrackLab software and with tailor-made Python scripts. Moreover, the tags also provided accelerometer data that so far is not exploited. One of the units is also equipped with 8 cameras (4 pairs) from which video footage was recorded with 25 fps. All of these cameras are calibrated, and their positions and orientations in the barn are known. We have developed a software tool to automatically detect cows and 4 key points for each cow (base of the left and right ear, withers and base of the tail) in video frames. Moreover, because all areas of the barn are visible from at least 2 viewpoints (cameras) we are able to determine 3D-coordinates for these key points. The aim of the setup is to develop tools for automated monitoring of behaviour, first of all for research purposes but ultimately also for farm management purposes. During a targeted measurement week (from 27-2-2023 until 6-3-2023) in one barn unit the Sewio-tags were set to communicate with 20 Hz and the cows were equipped with IceQube® pedometers (Peacock Technology Ltd, Stirling, Scotland UK). These provide information about duration of lying and standing bouts, times of transitions between lying and standing, number of steps and a motion index per 15 minutes [3]. Unfortunately, we only have video footage from 05:00 h until 19:00 h during the measurement week, from which 1 frame per second is analyzed with our image processing tool. To answer specific targeted research questions additional footage will be analyzed with higher frame rate. The animals leave the barn twice daily for milking, between 05 and 06 h and between 15 and 16 h. On return to the barn they are individually recognized at a selection gate, data of these detections is also available.

We investigate the insights in behaviour that can be obtained from the different sensors, and assess the practicalities for each of them. For image processing we focus on required storage capacity, streaming quality, required computing power, synchronization, animal identification and measurement accuracy. For the UWB sensors we focus on realized recording frequency, accuracy of the measured positions and impact of TrackLabs post-processing and battery capacity. Moreover, we will investigate the potential added value of the accelerometer data that is collected from the Sewio UWB sensors. Behavioural parameters obtained from both sensors is compared with information obtained from the IceQube pedometers.

References

1. Adriaens, I., Ouweltjes, W., Pastell, M., Ellen, E., & Kamphuis, C. (2022). Detecting dairy cows' lying behaviour using noisy 3D ultra-wide band positioning data. Peer Community Journal, 2: e55.
2. J. Cao, J., H. Tang, H. S. Fang, X. Shen, C. Lu and Y. W. Tai, 2019. Cross-Domain Adaptation for Animal Pose Estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9498-9507).
3. G. Charlton, Gauld, F. Veronesi, S.M. Rutter and E. Bleach,2022. Assessing the Accuracy of Leg Mounted Sensors for Recording Dairy Cow Behavioural Activity at Pasture, in Cubicle Housing and a Straw Yard. Animals (Basel). 2022 Mar 3;12(5):638. doi: 10.3390/ani12050638.

Measuring Dairy Cow Behavior Using a Barometric Sensor

Christiane Engels[1]

[1]Institute for Agricultural Engineering, University of Bonn, Bonn, Germany. christiane.engels@uni-bonn.de

Abstract

Barometric sensors are widespread in smartphones and wearable devices and have many applications, mainly in human or vehicle tracking. In this study, we collected barometric pressure and acceleration data of dairy cow behavior with a collar mounted sensor in a free-stall barn. The goal is to distinguish between standing and lying behavior based on the sensor height inferred from barometric pressure.

Introduction

In farm animal husbandry, the use of digital technologies becomes increasingly important for monitoring and managing the animals on the farm. Different sensor systems are employed in precision livestock farming to gather data about the animals' activity, feed and water intake, body temperature, weight, and barn climate. For dairy cows, the most commonly used sensors are pedometers, accelerometers, and locating systems recording the cows' locomotion activity. Especially lying behavior is an important and suitable parameter for assessing animal health and welfare [1, 2]. For example, lame cows spend longer time lying [3], heat stress reduces lying time [4] as well as uncomfortable bedding conditions [5].

The state-of-the-art technology for recording dairy cow lying time are neck- and leg-mounted accelerometers [6]. Accelerometers have been used in numerous studies on cow behavior monitoring [2], estrus detection [7], and lameness detection [3, 8].

In this study, we apply a sensor consisting of a 3D acceleration sensor and a barometric sensor. From relative changes in air pressure, the sensor height can be deduced by the barometric formula [9]. This enables the differentiation between standing and lying with a collar mounted sensor. Close to the Earth's surface (under 10 km altitude above sea-level), the change of pressure is almost linear with a decrease of 11.5 Pa per 1 m ascent [10].

There are various applications for barometric sensors which are widespread in smartphones and wearable devices. Most applications are in (indoor) positioning, e.g. fall detectors for humans [11], navigation in multi-floor buildings [12], and measuring vertical velocity (of elevators or drones) [13]. For animals, barometric sensors have so far been used in birds for flight dynamics analysis [14] and the position estimation of migratory movements [15]. Nabenishi et al. (2019) investigated a combined barometer and accelerometer for automated estrus behavior detection in tie stalls [16].

The absolute height accuracy of microelectromechanical system (MEMS) based barometers used in smartphones and in this study is low, ranging from 10 – 100 Pa (≙ 0.9 – 8.7 m), but their relative pressure accuracy on which most applications rely is high (0.2 – 2.0 Pa, ≙ 0.02 – 0.17 m) [10]. In particular, the relative pressure accuracy is significantly lower than the targeted height difference between standing and lying cows of approx. 0.7 – 0.8 m.

Besides the altitude, barometric pressure is influenced by environmental conditions like wind and temperature [9]. These effects have a much smaller amplitude than the altitude changes corresponding to behavioral activities. In addition, vertical displacements result in faster pressure responses and hence a larger gradient. However, it is important to analyze relative pressure variations rather than absolute pressure levels.

Whereas accelerometers need a high sampling frequency of > 50 Hz to achieve appropriate data, a sampling frequency of 1 Hz may be sufficient for barometric pressure data to capture vertical displacements [10]. This reduces the amount of data recorded, stored, (pre)processed, and transmitted by the sensor and consequently lowers the required power consumption.

The aim of the study was to collect barometric training data of different cow behaviors and to evaluate whether it is possible to distinguish between the categories *standing* and *lying* and potentially others under practical conditions. The intended application of the barometric sensor is as part of an indoor positioning system for animal tracking which – usually employing a 2D location – can only determine the position and not the behavior of an animal inside the barn. Thus, the barometric data enriches animal observation as a third axis and enables for example the recording of lying and standing behavior in the cubicles.

Material and Method

In the study, data was collected with two custom-built devices containing a digital barometric pressure sensor (DPS310, Infineon Technologies AG, Neubiberg, Germany) and a 3D accelerometer (ADXL435, Analog Devices Inc., Wilmington, U.S.) each. The barometric sensor has a relative pressure accuracy of ±0.5 Pa (≙ ±0.04 m) [17]. Data was collected at 1 Hz for the barometric sensor and 60 Hz for the acceleration data. The devices were embedded in industrial housings and attached to collars for the practical use in the barn. The sensors could be paired with a smartphone via Bluetooth. With a corresponding self-developed app, the cows' behavior was logged by means of an ethogram. The applied ethogram has two layers with eight main categories (lying, lying down/ standing up, standing, walking, feeding, ruminating, mounting (estrus), and other) and 25 subcategories. It was designed to capture especially those activities and poses affecting the height of the collar mounted sensor. The logged events were directly stored together with the recorded sensor data on the device such that no time synchronization was necessary.

In preliminary tests, the sensor was carried by a human up and down the staircase of a building at different speeds. Including the basement, the building had four floors with landings on the upper two staircases. The step height was approx. 168 mm.

The main study was performed in the free-stall barn of the Saxon State Office for Environment, Agriculture and Geology (Köllitsch, Germany). Ten sessions of 1:30 – 4:00 h duration were recorded with eight individual Holstein-Friesian dairy cows in September 2023. At the beginning of each session, a (healthy) cow was chosen randomly and equipped with the collar mounted sensor. The sensor was paired with a smartphone and the data recording started. The activities of each cow were logged without interfering with the cow's behavior. At the end of the session, the sensor was removed from the cow. In total 29:00 h of labelled data were collected. In total, 985 events were recorded, including 30 lying phases (comprising 14:28 h), 13 lying down and 15 standing up events. The most interesting events for distinguishing standing and lying behavior are the transitions *standing up* and *lying down*. These were further analyzed regarding the recorded pressure values.

In the recorded sessions, there was a huge variation in pressure ranging from 998.5 – 1022.5 hPa. This was due to different weather conditions on the trial days with temperatures of 15.0 – 29.4 °C measured with a weather station outside the barn. As most sessions were recorded in the morning, the barometric pressure slowly declines with increasing temperature throughout the session.

Results and Discussion

Preliminary Tests

The measured pressure data of an example stair climb of the preliminary tests is depicted in Figure 1. The walking phases on floors and landings with constant altitude are clearly recognizable (SD ±0.4 Pa ≙ ±0.03 m) and almost match the previous pressure levels (±1.9 Pa ≙ ±0.17 m). The abrupt change of altitude when positioning the sensor on the basement floor and subsequently picking it up (approx. 1.15 m) is detected within 2 s with a pressure difference of 13.4 Pa.

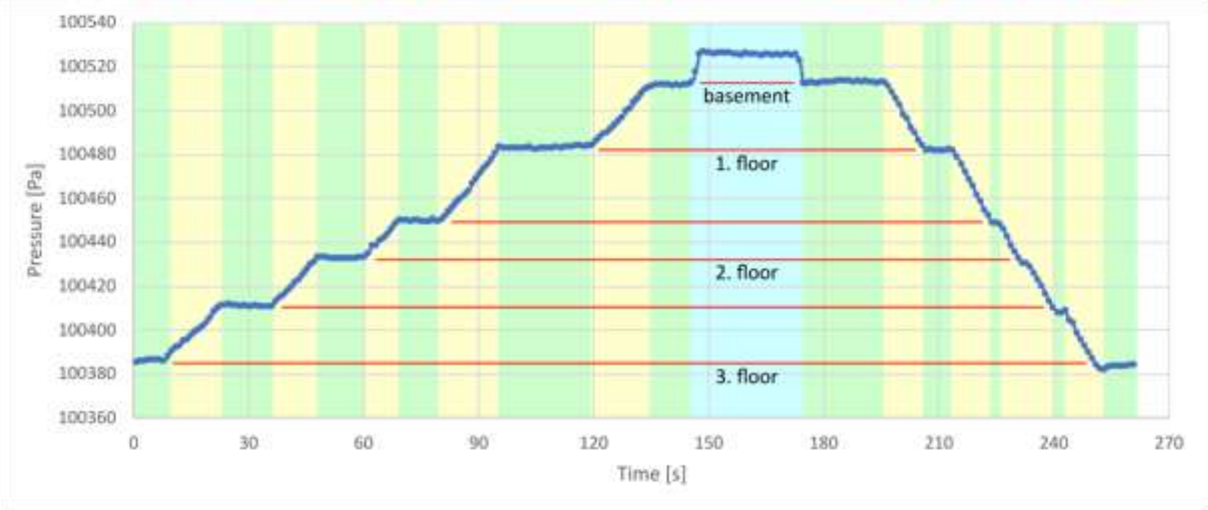


Figure 1. Measured pressure data of the preliminary test. Sensor carried by human first down slowly, then up faster the staircase of a building. Climbing is indicated in yellow, walking on landings in green. In the basement, the sensor was positioned on the floor for approx. 30 s (cyan).

Main Trial

Figure 2 shows two example recordings of a *standing up* events.

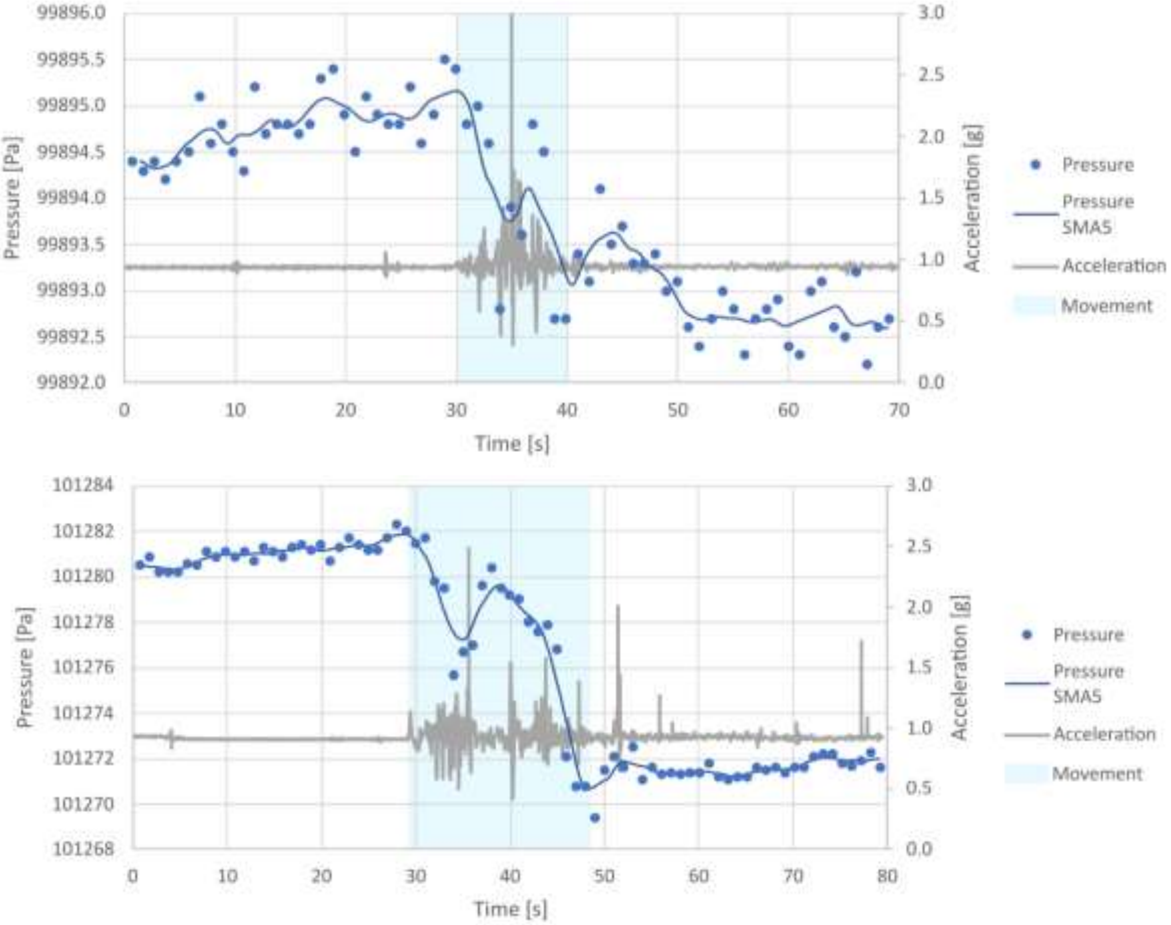


Figure 2. Recorded pressure and acceleration data of two *standing up* events (30 s prior to 30 s after the movement). Acceleration is plotted as combined magnitude of the three axes. Pressure values measured at 1 Hz are additionally depicted as simple moving average (SMA) over 5 s.

In both examples, there is a pressure difference prior and after the movement phase (101281.0 Pa vs. 101271. 6 Pa; 99894.8 Pa vs. 99892.9 Pa). In the first example, this difference is 9.4 Pa which corresponds to a rise of 0.82 m and matches the height difference of lying and standing cows. Nabenishi et al. (2019) reported pressure decreases of approx. 12 Pa and increases of approx. 13 Pa for cows standing up and lying down, respectively [16]. In the second example however, this difference is only 1.9 Pa which corresponds to a rise of 0.17 m. The curvature of the pressure course present in both examples during the movement phase may indicate the characteristic head swing of cows when standing up [18]. For *lying down* events, there are similar findings. Figure 3 shows an example. There are pressure differences prior and after the movement phase with varying magnitude (7.8 Pa ≙ 0.68 m in Figure 3). Here, the rise in pressure already starts before the actual lying down phase. This may again correspond to the lying down process of cows starting with lowering their head and bending the forelimbs to stand on the carpal joints [18].

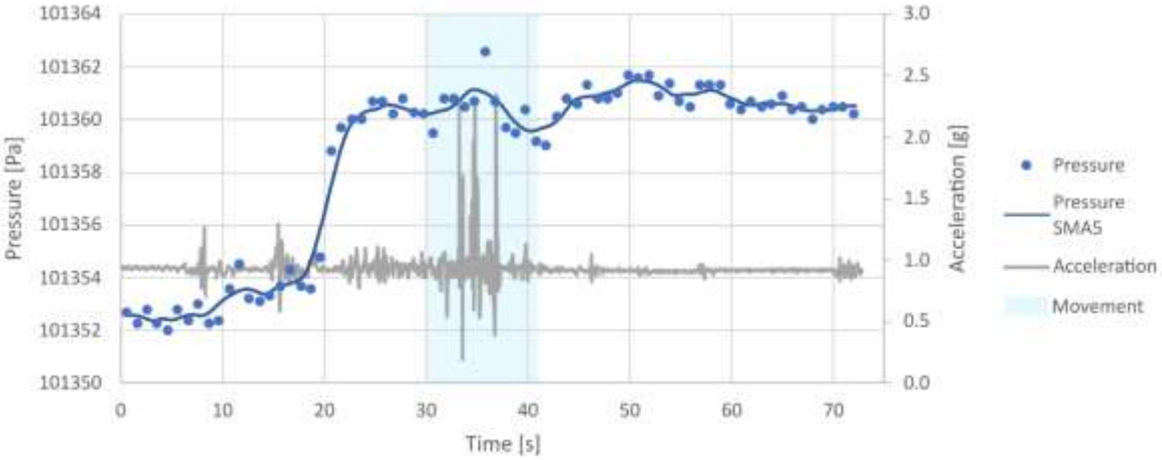


Figure 3. Recorded pressure and acceleration data of a *lying down* event (30 s prior to 30 s after the movement). Acceleration is plotted as combined magnitude of the three axes. Pressure values measured at 1 Hz are additionally depicted as simple moving average (SMA) over 5 s.

During the lying phases, the pressure varies due to the above-mentioned environmental influences. Tough the amplitude gets up to 50 Pa per hour, this variation is slow and small (⌀ 0.3 Pa per second) combined to the pressure difference of the transition movements.

Conclusion

Barometric sensors have the potential to distinguish between standing and lying behavior of dairy cows. In the recorded data, the pressure level difference before and after the transition between standing and lying was noticeable. However, a more detailed categorization of the remaining behaviors was not possible from the barometric data alone. The combination of barometric and acceleration data – or barometric and position data as intended with the sensor used in this study – are most promising. To this end, further research, the integration of machine learning techniques, and the recording of more training data is necessary.

Ethical statement

In accordance with the faculty’s animal welfare officer, an animal experimental application was not considered necessary. After equipping the cows with the collars, they were free to express their normal behavior while being monitored by a human from an appropriate distance.

Acknowledgements

The author thanks the team of the Saxon State Office for Environment, Agriculture and Geology in Köllitsch for recording the sessions of the main study.

This work was financially funded by the German Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany, granted by the Federal Office for Agriculture and Food (BLE; grant number: 28DE108A18).

References:

1. Fregonesi, J.A., Leaver, J.D. (2001). Behaviour, performance and health indicators of welfare for dairy cows housed in strawyard or cubicle systems. *Livestock Production Science* **68**, 205–216.

2. Mattachini, G., Antler, A., Riva, E., Arbel, A., Provolo, G. (2013). Automated measurement of lying behavior for monitoring the comfort and welfare of lactating dairy cows. *Livestock Science* **158**, 145–150.

3. Chapinal, N., de Passillé, A.M., Weary, D.M., von Keyserlingk, M.A.G., Rushen, J. (2009). Using gait score, walking speed, and lying behavior to detect hoof lesions in dairy cows. *Journal of Dairy Science* **92**, 4365–4374.

4. Tullo, E., Mattachini, G., Riva, E., Finzi, A., Provolo, G., Guarino, (2019). Effects of climatic conditions on the lying behavior of a group of primiparous dairy cows. *Animals*, **9**, 869.

5. Tucker, C.B., Jensen, M.B., de Passillé, A.M., Hänninen, L., Rushen, J. (2021). Invited review: Lying time and the welfare of dairy cows. *Journal of Dairy Science* **104**, 20–46.

6. Benaissa, S., Tuyttens, F.A., Plets, D., De Pessemier, T., Trogh, J., Tanghe, E., Martens, L., Vandaele, L., Van Nuffel, A., Joseph, W., Sonck, B. (2019). On the use of on-cow accelerometers for the classification of behaviours in dairy barns. *Research in Veterinary Science* **125**, 425–433.

7. Arcidiacono, C., Mancino, M., Porto, S.M.C. (2020). Moving mean-based algorithm for dairy cow's oestrus detection from uniaxial-accelerometer data acquired in a free-stall barn. *Computers and Electronics in Agriculture* **175**, 105498.

8. O'Leary, N.W., Byrne, D.T., O'Connor, A.H., Shalloo, L. (2020). Invited review: Cattle lameness detection with accelerometers. *Journal of Dairy Science* **103**, 3895–3911.

9. Spiridonov, V., Ćurić, M., Spiridonov, V., Ćurić, M. (2021). Atmospheric pressure and wind. *Fundamentals of Meteorology*, 87–114.

10. Manivannan, A., Chin, W. C. B., Barrat, A., Bouffanais, R. (2020). On the challenges and potential of using barometric sensors to track human activity. *Sensors* **20**, 6786.

11. Bianchi, F., Redmond, S.J., Narayanan, M.R., Cerutti, S., Lovell, N.H. (2010). Barometric pressure and triaxial accelerometry-based falls event detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **18**, 619–627.

12. Xia, H., Wang, X., Qiao, Y., Jian, J., Chang, Y. (2015). Using multiple barometers to detect the floor location of smart phones with built-in barometric sensors for indoor positioning. *Sensors* **15**, 7857–7877.

13. Monteiro, M., Martí, A.C. (2016). Using smartphone pressure sensors to measure vertical velocities of elevators, stairways, and drones. *Physics Education* **52**.

14. Shipley, J.R., Kapoor, J., Dreelin, R.A., Winkler, D.W. (2018). An open-source sensor-logger for recording vertical movement in free-living organisms. *Methods in Ecology and Evolution* **9**, 465–471.

15. Nussbaumer, R., Gravey, M., Briedis, M., Liechti, F. (2023). Global positioning with animal-borne pressure sensors. *Methods in Ecology and Evolution* **14**, 1104–1117.

16. Nabenishi, H., Kawakami, S., Shimo, S., Takeshita, K., Yamazaki, A., & Suzuki, K. (2019). Automated detection of estrous behavior in tie-stall housing using a barometer and accelerometer. *Journal of Reproduction and Development* **65**, 91–95.

17. Infineon (2016). Data Sheet: DPS310 - Digital Pressure Sensor, IFX-sch1406115644540, https://www.infineon.com/cms/de/product/sensor/pressure-sensors/pressure-sensors-for-iot/dps310/#!?fileId=5546d462576f34750157750826c42242

18. Schnitzer, U. (1971). Abliegen, Liegestellungen und Aufstehen beim Rind. *KTBL-Bauschrift* **10**.

Posters

A Semi-automated Gait Assessment Tool for Mouse Locomotion Recovery after Neurological Injury

A. Bernstein[1], A. Vivinetto[1], J. Kaiser[1], M. Soliman[1], A. Lammers[1], V. Sahni[1],[2,3], and E. Hollis[1,2]

1. Burke Neurological Institute, White Plains, USA. amb3005@med.cornell.edu,

2. Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, USA. edh3001@med.cornell.edu,

3. Department of Stem Cell and Regenerative Biology, and Center for Brain Science, Harvard University, Cambridge, USA. vis2763@med.cornell.edu

Abstract

Quantitative gait assessment is a critical tool for studying neurological injury. To address limitations in existing tools we present a novel apparatus using markerless tracking for comprehensive analysis of whole-body locomotion in mice following spinal cord injury.

Introduction

Quantitative and unbiased gait assessment is a critical tool in pre-clinical studies of neurological injury and dysfunction, including spinal cord injury (SCI). Gait analysis has been instrumental in assessing deficits and outcomes in the field of SCI [1, 2]. Multiple tools have been developed to standardize and look at various aspects of gait, such as open field ordinal testing of locomotor coordination [3, 4], narrow beam walking, ladder walking, footprint analysis, and coordinated locomotion (Rotarod, Ugo Basile, Gemonio, Italy). These tests have foci limited to the specific sensorimotor outcomes they were developed for and do not give a comprehensive view of locomotion. As manual analysis of gait in rodents is time-consuming and reproducibility relies on investigator expertise, multiple tools have been developed for automated and objective analysis, incorporating measures of a greater number of gait parameters than manual methods [1]. These automated methods were rapidly adopted, with Noldus' CatWalk emerging as the most highly cited in PubMed indexed publications (Figure 1). As CatWalk relies on filming from below to characterize gait, complimentary methods are required to evaluate body posture from a lateral viewpoint [3, 4]. Furthermore, in severe injury models, animals are often unable to make plantar steps, so systems designed for assessing gait solely from ventral views of paw placement have limited utility at early recovery phases. Here we describe a new behavioral apparatus to analyze whole-body locomotion using 37 tracking points and the open-source software DeepLabCut (DLC) to provide a more comprehensive understanding of locomotor recovery after SCI used on its own or combined with CatWalk.

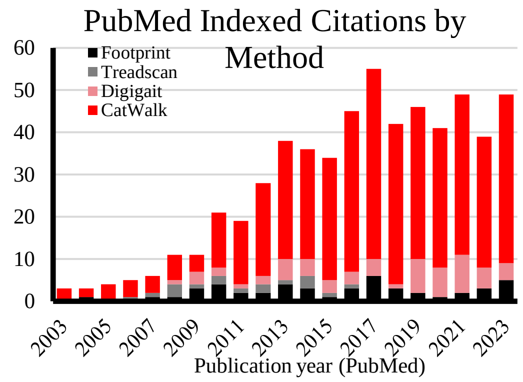


Figure 1. Adoption of automated gait assessment. Since first described in 2003, CatWalk became the most widely published automated gait assessment method, compared to Digigait and Treadscan, or manual footprint testing (source PubMed 26/12/2023)

Approach

We developed a semi-automated tool to analyze body posture, gait, and digit spread in mice with SCI. The walkway is made of clear acrylic (27cm x 6.4cm x 15.7cm) with a mirror positioned underneath at approximately 45-degree angle to acquire ventral views of the mouse. A Basler Ace acA1140-220 um camera was used to film the mice walking at various timepoints, 227 fps at 1.6MP resolution (Figure 2). The DLC networks were trained based on ResNet-50 by manually labelling 600 frames from randomly selected videos of different mice at various time points. Tracking points in the side view included, the tail tip, mid-tail, base of tail, knee, ankle, toe, shoulder, wrist, front paw, and nose. The mirror was used to track the tail (tip and base), front paws, hind paws, nose, and digits on all 4 paws. A custom Matlab script was developed to analyze a wide range of locomotor parameters including aspects of body posture, kinematics, and digit spread. This apparatus was used to detect differences in treatments as well as severity of SCI, which is presented below.

Results

Young adult C57BL/6J mice (*n* = 6, males, 3-4 months old) underwent either a 75 kDyn (severe) or 50 kDyn (moderate) contusion injury at spinal level T9 using an Infinite Horizons spinal cord impactor. Using our apparatus, we were able to analyze 22 different parameters of gait, most of which were focused on the lateral view of the mouse. The mice were tested prior to injury and at multiple time points after SCI to map the recovery process over 14 days. With the apparatus we were able to analyze both moderate and severe injury over time, included aspects of leg movements as early as four days after severe injury when the mice were dragging their hindpaws. We examined distinct components of stepping and found a graded functional impairment, with less impairment after moderate contusion SCI (Figure 2 C-D). Early after severe SCI, the mice tend to drag their tails more, as determined by the height of the tail base and the height of the tail tip, as compared to after moderate SCI (Figure 2-D). Also, digit spread and the angle of the ankle at heel-strike were smaller in the severe injury group early after SCI, indicating impaired stepping and body weight support (Figure 2-D). By analyzing relative positions, angles, and trajectories in the lateral view and bottom view of the mouse, this new apparatus overcame some of the challenges when using the CatWalk assessment in animals undergoing early recovery, such as decreased labeling due to weight loss after injury, or movement of the hindlimbs with no weight support. This analysis can be combined with the CatWalk to generate a comprehensive understanding of locomotor recovery in mouse models of contusive SCI.

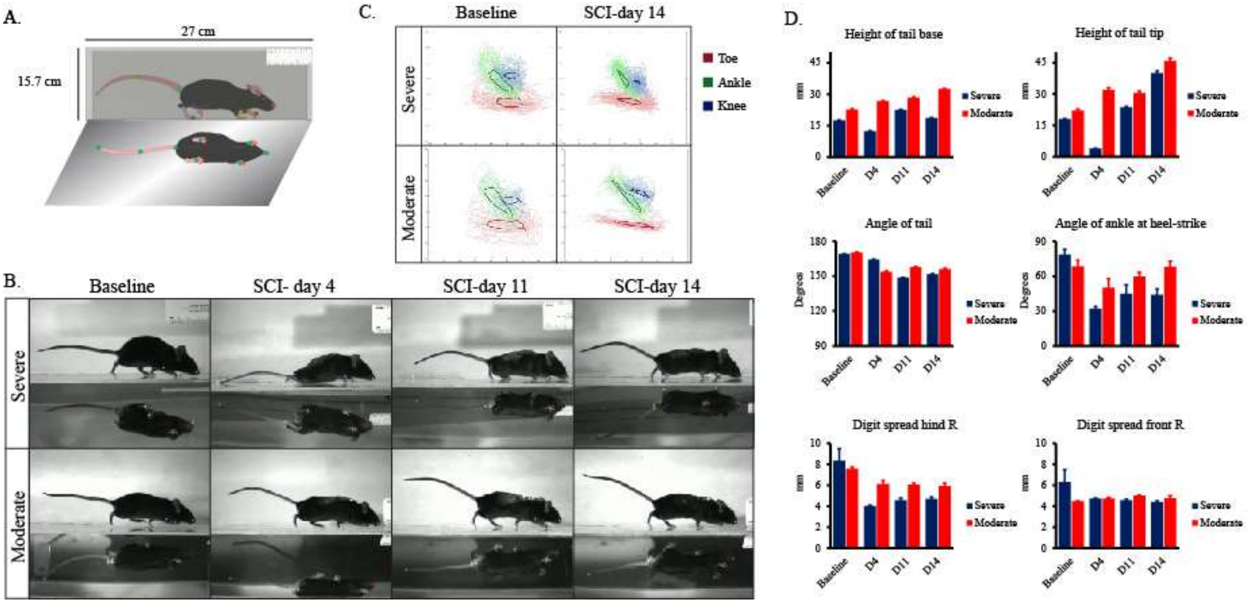


Figure 2: Evaluation of moderate vs. severe T9 contusive SCI. (A) Schematic of the locomotor analysis tool. (B) Examples of DeepLabCut tracking at various time points after SCI. (C) Average trajectory during steps of the toe, ankle, and knee at baseline and 14 days after SCI in the moderate and severe SCI. (D) Examples of some body posture and digit spread parameters that were analyzed. (*n* = 3 / group)

Future work

Building upon the initial success of this semi-automated gait analysis tool, our aim is to develop a modular version that can provide a comprehensive understanding of locomotor recovery after SCI as well as other injury and disease models. This new system is adapted from the work of Weber et al. [5], who developed an apparatus with ventral camera placement and 2 mirrors to simultaneously obtain both lateral views (Figure 3). Obtaining both lateral views simultaneously reduce the amount of time needed to run animals across the apparatus while also providing more information about coordination and bilateral effects of injury. The first two modules we tested, the ladder module (102.1cm x 5cm x 17cm) and the flat substrate (102.3 cm x 7.4cm x 15 cm) are made of 0.5 cm thick acrylic (Figure 3). These sit atop a frame made of 40 series square aluminum T-slot beams (101.92cm x 88.9cm x 83.5cm). A Basler ace acA1140-220 um camera with a wide-view Basler lens (C125-0418-5M-P f4mm) was placed 53 cm from the base of the walkway (Figure 3). Mirrors were placed at 30-degree angles from the side of the modules and will be adjusted to provide the optimal view in each module using custom 3D printed parts (Figure 3). DeepLabCut was used during the prototyping phase to determine optimal lighting conditions and viability of tracking points of interest. A custom script to analyze the videos, detect toe-off and heel-strike occurrences, and generate a 3D reconstruction of mouse gait will be developed for each modality including an irregular ladder, flat substrate locomotion, and narrow beam crossing.

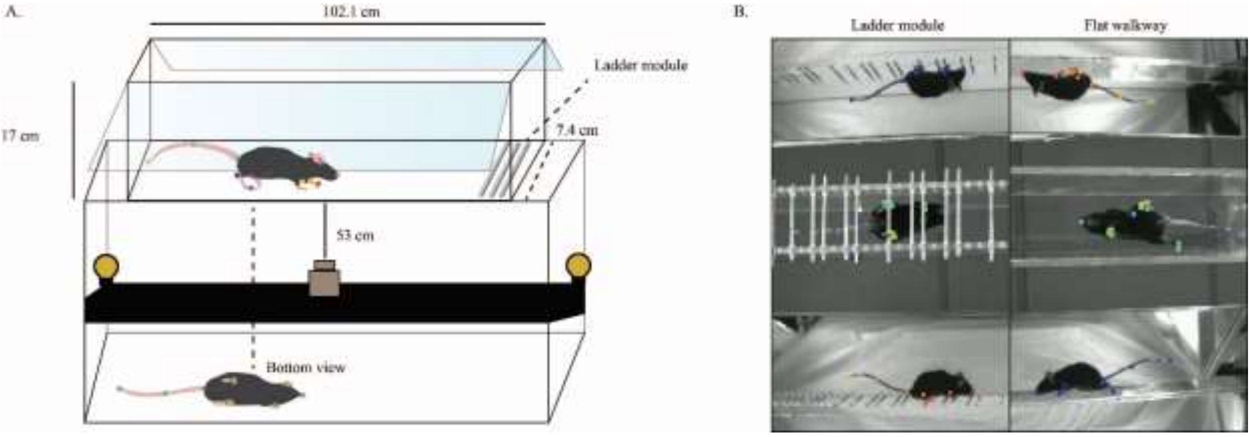


Figure 3: Schematic of new modular locomotor testing apparatus (A). Examples of initial proof-of-concept DeepLabCut testing for both the irregular ladder module and flat substrate

Acknowledgements

This work was supported by the New York State Department of Health C37717GG (EH) and C39069GG (VS), as well as the National Institutes of Health DP2NS106663 (EH), R01NS105725 (EH), R01NS131662 (VS), and R21NS127622-01A1 (VS).

Ethical Statement

All animal work was approved by the Weill Cornell Medicine Institutional Animal Care and Use Committee. Veterinary care and euthanasia were consistent with AVMA guideline.

References

1. Hamers, F.P., G.C. Koopmans, and E.A. Joosten, CatWalk-assisted gait analysis in the assessment of spinal cord injury. J Neurotrauma, 2006. 23(3-4): p. 537-48.

2. Preisig, D.F., et al., High-speed video gait analysis reveals early and characteristic locomotor phenotypes in mouse models of neurodegenerative movement disorders. Behav Brain Res, 2016. 311: p. 340-353.

3. Basso, D.M., M.S. Beattie, and J.C. Bresnahan, A sensitive and reliable locomotor rating scale for open field testing in rats. J Neurotrauma, 1995. 12(1): p. 1-21.

4. Basso, D.M., et al., Basso Mouse Scale for locomotion detects differences in recovery after spinal cord injury in five common mouse strains. J Neurotrauma, 2006. 23(5): p. 635-59.

5. Weber, R.Z., et al., Deep learning-based behavioral profiling of rodent stroke recovery. BMC Biology, 2022. 20(1): p. 232.

Using a Pop-up Tunnel to Record Vocalisations of Sea Turtle Hatchlings to Avoid Geophony

S.N.D. Melo[1,a], D. A. Melo[2,b], M.F.S.D. Silva[3,c] , V.C.S. Neves[3,d], B. Bezerra[1,e]

[1]Programa de Pós-Graduação em Biologia Animal, Departamento de Zoologia, Universidade Federal de Pernambuco, Recife, Brasil, [a]safira.melo@ufpe.br [e]bruna.bezerra@ufpe.br.

[2]Colégio Militar do Recife, [b]5705.deboramelo.cmr@gmail.com.

[3]NGO Ecoassociados, Ipojuca, Brasil, [c]matheus.diassilva@ufpe.br, [d]vivianecoassociados@gmail.com.

Abstract

Research on sea turtles' vocalisations is scarce, often focusing on the incubation period. Recording the hatchlings during their walk to the sea is difficult due to wind noise on the seafront. We propose a simple solution to standardise and optimise the acoustic recordings during this period. We placed a recorder inside a pop-up tunnel and let the hatchlings walk through it before reaching the sea. It allowed the recordings while minimising signal masking.

Keywords

Data collection, acoustic signal, noise, masking, *Eretmochelys imbricata*.

Introduction

Ecological information on the distribution of nesting sites, strandings, and feeding ecology is well-known for sea turtles [1; 2]. However, the general behaviour of these animals is still poorly understood due to the obvious difficulties of observing sea turtles in the wild. Research on vocal behaviour is also scarce, often focusing on recordings over the incubation period due to greater accessibility to the nests [3; 4; 5]. Nevertheless, the recording methods of these studies differ in sampling method and size, and sound characterisation [e.g., 3; 6; 7]. Recording sea turtle hatchlings during their walk to the sea is not simple due to strong wind noises on the seafront masking the acoustic signals. Here, we propose a simple solution to optimise the recordings of vocal behaviours in this developmental stage of sea turtles while avoiding masking from geophony. We tested our method on hatchlings of the species *Eretmochelys imbricata*.

Methods

We recorded *Eretmochelys imbricata* individuals from seven nests in the Ipojuca municipality coast in Pernambuco, Northeast Brazil (see Figure 1).

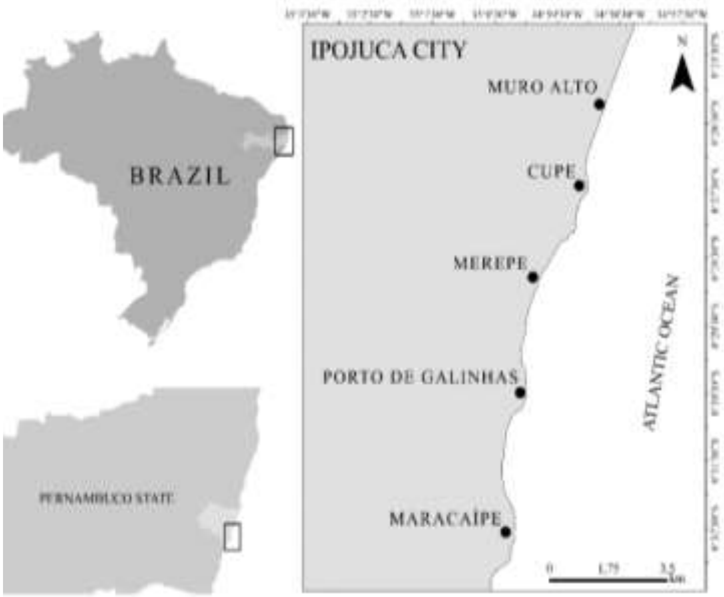


Figure 1. Map of the recording site, comprising five beaches over the coast of Ipojuca municipality in Pernambuco State, Northeast Brazil.

We conducted recordings in breeding seasons between April/2021 and March/2023, using an Audiomoth recorder (Open Acoustic Devices) under 48kHz and 16-bit sampling rate. We placed the recorder inside a 1,75m long pop-up tunnel with sand and simply let the hatchlings walk through the tunnel before reaching the sea (see Figure 2). We then analysed the recordings made over this period through the Raven Pro 1.5 software (Cornell Lab of Ornithology, Ithaca, NY). We manually registered the following acoustic parameters from the signals recorded: delta time, low frequency, high frequency, maximum frequency, and number of harmonics.

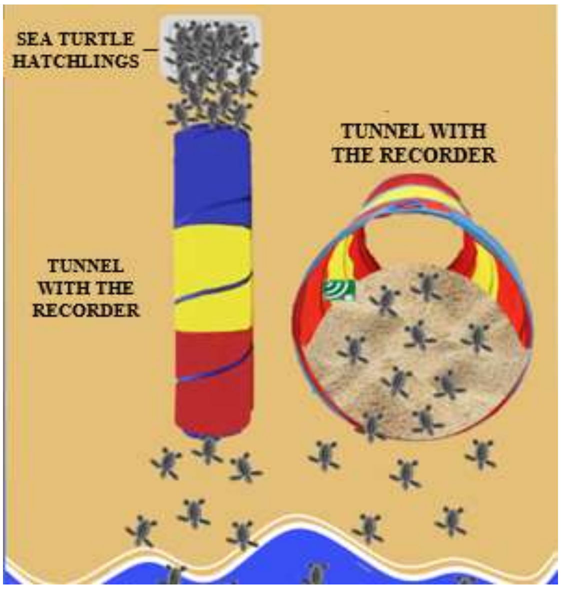


Figure 2. A - Illustration of the hatchlings going through the pop-up tunnel.

Results

The hatchlings took, on average, 7,8 minutes (Range 4 – 12 min) to pass through the pop-up tunnel. We successfully obtained recordings from hatchlings of six out of seven nests. We recorded four types of acoustic signals (see Figure 3).

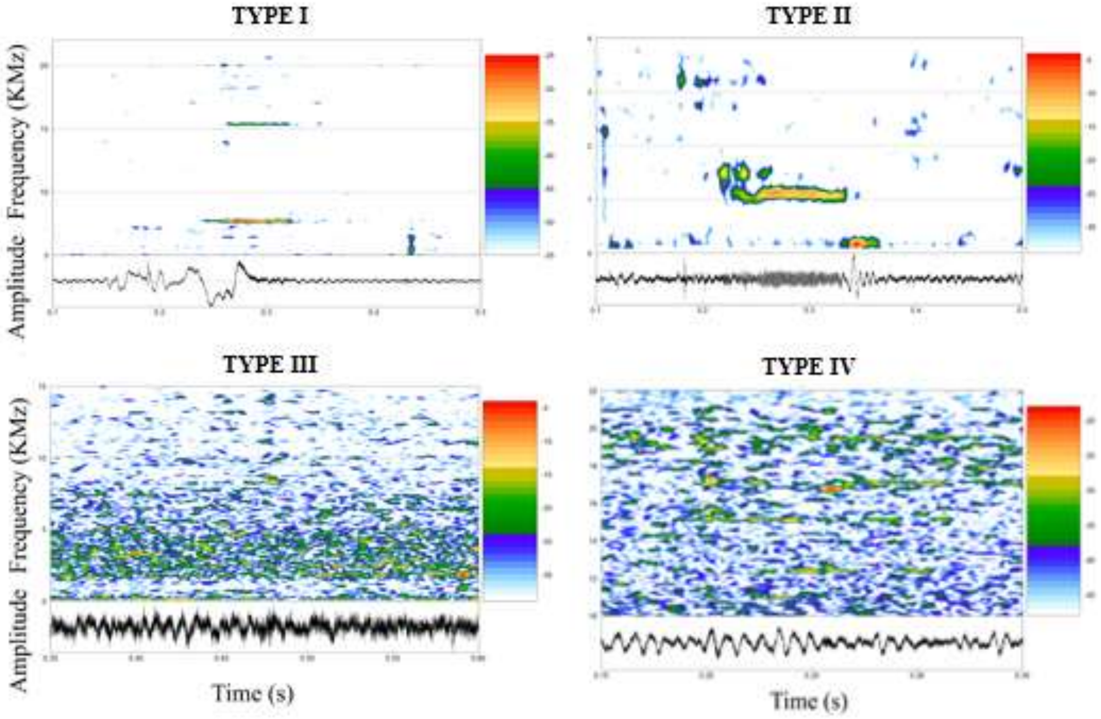


Figure 3. Spectrograms of the acoustic signals recorded from *Eretmochelys imbricata* hatchlings during their walk to the sea.

Discussion

The pop-up tunnel allowed the recordings of the hatchlings while minimising masking caused by strong wind noise on the beachfront. We obtained four acoustic signals from *Eretmochelys imbricata* in a developmental stage different from incubation (i.e., the walk to the sea), which, for instance, may help to respond to new questions related to synchronising behaviours when the hatchlings walk to the sea to help the survival of these animals. We currently have two studies describing acoustic signals for *Eretmochelys imbricata* - one study describing four signals [5] and another study describing five signals [7]. Our next step is to test different recording solutions to test the efficiency of the pop-up tunnel and add other sea turtle species to our tests. We hope to present these additional data during the Measuring Behaviour conference as we are conducting experiments for sea turtles in Ipojuca municipality in Brazil during the current breeding season. Our non-invasive simple solution may help standardise the collection of vocal behaviours at different sites.

Acknowledgements

We thank CAPES, CNPq, and Facepe for their financial support. The study agrees with Brazilian law, and data were collected under the SISBIO licence 75658-4.

References

1. Hays G. C. (2008). Sea turtles: A review of some key recent discoveries and remaining questions. *Journal of Experimental Marine Biology and Ecology* 356 (1-3):1-7. doi: 10.3354/esr00865

2. Duncan, E. M., Botterell, Z. R. L, Broderick, A. C., Galloway, T. S., Lindeque, P. K., Nuno, A., Godley, B. J. (2017). A global review of marine turtle entanglement in anthropogenic debris: a baseline for further action. Endangered Species Research, 34: 431–448. doi: https://doi.org/10.1016/j.jembe.2007.12.016

3. Muñoz, R.S. (2010). Estudio de los sonidos emitidos por las crías de tortuga boba, *Caretta caretta*, en el momento de la eclosión. *Anales Universitarios de Etología* 4(2):63–70.

4. Ferrara, C.R., Vogt, R.C., Harfush, M.R., Sousa-Lima, R.S., Albavera, E., Tavera, A. (2014). First evidence of leatherback turtle (*Dermochelys coriacea*) embryos and hatchlings emitting sounds. *Chelonian Conservation Biology* 13(1):110–114. doi: 10.2744/ccb-1045.1.

5. Monteiro, C.C., Carmo, H.M.A., Santos, A.J.B., Corso, G., Sousa-Lima, R.S. (2019). First record of bioacoustic emission in embryos and hatchlings of Hawksbill Sea turtles (*Eretmochelys imbricata*). *Chelonian Conservation Biology* 18(2):273–278. doi: 10.2744/CCB-1382.1.

6. Mckenna, L.N., Paladino, F.V., Tomillo, P.S., Robinson, N.J. (2019). Do sea turtles vocalise to synchronise hatching or nest emergence? *Copeia* 107(1):120–123. doi: 10.1643/ce-18-069.

7. Melo, S.N.D., Silva, M.F.S.D., Santos, P.J.P., Neves, V.C.S., Bezerra, B.M. (2023). Sound production in sea turtle nests and hatchlings (*Eretmochelys imbricata* and *Caretta caretta*) in Northeast Brazil. *Bioacoustics* 32(6): 1-15. doi: 10.1080/09524622.2023.2251936.

The OpenBehavior Project: A Database and Dissemination Platform for Open-source Tools for Behavioral Neuroscience Research

S. Bradley[1], L. Amarante[2], K. Lopez[3], J. Frie[4], J. Palmer[3], S. White[2], J. Khokhar[5], A. Kravitz[6], and M. Laubach[3]

1. **Rodent Behavioral Core & Section on Behavioral Neuroscience, National Institute of Mental Health, Bethesda, Maryland, USA. sean.bradley@nih.gov**

2. **Section on Neurobiology of Compulsive Behaviors, National Institute of Mental Health, Bethesda, Maryland, USA. linda.amarante@nih.gov, samantha.white@nih.gov**

3. **Department of Neuroscience, American University, Washington DC, USA. mark.laubach@american.edu**

4. **Department of Neuroscience, University of Guelph, Guelph, Ontario, Canada. jfrie@uoguelph.ca**

5. **Department of Anatomy and Cell Biology, University of Western Ontario, London, Ontario, Canada. jkhokha@uwo.ca**

6. **Department of Psychiatry, Washington University of St. Louis, St. Louis, Missouri, USA. alexxai@wustl.edu**

The OpenBehavior Project promotes the use of open-source tools for behavioral neuroscience research. Since 2016, the project has disseminated information on more than 250 research tools on a weekly basis through blog posts to our website (OpenBehavior.com) and through social media. Over the past 3 years, we have (i) created a database of all tools featured on openbehavior.com and issued Research Resource IDentifiers (RRIDs) that facilitate the citation and tracking of the tools in research publications; (ii) created a repository of raw video recordings of animals performing behavioral tasks that are commonly used in neuroscience research, organized a series of community conversations on video analysis tools, and written a commentary on setting video methods in a lab and best practices for the use of video methods; (iii) developed in-person and virtual training workshops on Arduino-based microcontrollers and 3D printing methods; (iv) created a repository of validated open-source designs for 3D printed objects used in neuroscience research.

We believe that open-source tools play an important role in addressing the issues of reproducibility and rigor that are endemic within the field of behavioral neuroscience. Open-source tools are inherently transparent, reproducible, and sharable. Adopting open-source tools allows the methodology of each tool or program to be rigorously evaluated and freely shared between laboratories. This allows researchers across laboratories and institutions to collect or analyze their data in the same way with greatly reduced financial barriers to entry. Further, these freely-available tools can meet or exceed the performance of commercial counterparts, offering value beyond being, 'the cheap option.' By contrast, commercial hardware and software is generally more costly. Especially at a time when the price of everything seems to be increasing, these costs can render such methods impractical for researchers facing tight budgets. Commercial software may also include proprietary algorithms that are opaque to the researcher and may vary between versions, especially for software-as-a-service models. Although this feature is not unique to proprietary software (and open-source packages may have their own issues with version control), open-source software allows researchers direct access to the methods used within or between laboratories to determine whether approaches are directly comparible.

The largest obstacle to the adoption of open-source tools is the expertise required to setup, implement, and analyze the output of these systems. Despite the advantages of open-source methods in transparency and reproducibility, 'off-the-shelf' solutions are a compelling option for laboratories without staff who may be experts in fabrication, electronics, or computer science. As such, the OpenBehavior Project attempts to lower this barrier of entry by disseminating information and best practices in the form of written journal articles and, more recently, by forming a community to encourage direct collaboration between developers, experts, and new users. This has taken the form of a community workshop as a Society for Neuroscience satellite in 2023 and an online community presence for user with questions to find experienced researchers who may provide answers.

This year, we will launch two new curated collections called "Setups and Protocols" and "Data and Analysis". These will include items needed to integrate open-source devices and programs into existing laboratory setups, protocols for using the methods, and example data sets and analysis code. Cognizant that there are many novel reseach methods presented outside of ‘methods papers,’ special attention will be given to including open-source methods included in research papers but not published as stand-alone publications.

Our poster presentation will be informational and is intended to both inform our audience of the suite of curated tools provided by the Project and improve the Project by fostering relationships with researchers who develop new methods and tools.

Capturing Differences in Mouse Behaviour Induced by Gene Mutations and Pharmacological Intervention using Motion Sequencing

Jack Bray[1] & Gernot Riedel[1]

[1]Institute of Medical Science, University of Aberdeen. Aberdeen, jack.bray1@abdn.ac.uk

Introduction

For many years, measuring behaviour in rodents typically employed point tracking over standard 2-D video to summarise 'typical' parameters such as distance, speed, and position. More recently, researchers have utilised depth cameras to track behaviour in 3-D and gain a more complete understanding of animal behaviour and a more comprehensive analysis of behavioural anomalies. Much like with 2-D videos, depth recordings also require a human to identify and label specific movements which can be time consuming and require the behaviour to be recognised in the first place, leading to results which encompass only those behaviours which can be observed by a human.

To address these limitations, there has been growing interest in devising unsupervised, data-driven methods that can identify the inherent patterns of behaviour and delineate how experimental interventions, like gene mutations or drug treatments, bring about alterations in the behavioural patterns [1].

Therefore, this study aimed to assess the feasibility of using motion sequencing algorithms applied to depth videos to establish any differences in behavioural patterns associated with pharmacological interventions and gene mutations in mice.

Materials and Methods

Mice & pharmacological intervention

A total of 20, 5-month-old, female mice were used for this experiment. 10 NMRI mice were used as control animals. 10 Line66 mice were used to assess changes in behaviour associated with gene mutations. The Line66 mouse model has an NMRI background strain and expresses the full-length human tau protein, carrying a double mutation (P301L & G335D). These mice have abundant tau pathology widely distributed throughout the brain and present with specific sensorimotor impairments [2].

In order to assess changes in behaviour associated with pharmacological intervention, mice were also injected with either intraperitoneal Saline or MK801 (0.2mg/kg). The non-competitive NMDA receptor antagonist, MK801, is known to increase locomotor activity and induce stereotypic behaviours [3].

Open field test

All mice underwent a 30-minute Open Field Test. Mice were transported from their housing room to the experimental room and allowed to habituate for 10 minutes. After which, mice received the injection and housed for a further 30-minutes before the start of the test.

The Open Field is a circular Perspex arena measuring 50cm in diameter. Matte, black, sticky backed vinyl was applied to the base and arena wall. This vinyl was sanded down with 120-grit sandpaper to stop the refection of infrared emitted from the time-of-flight sensor.

The mice were recorded with an overhead Kinect v2 sensor [4] which allowed for simultaneous recording of 2-D RGB, and depth videos. RGB videos were recored at 30 frames-per-second (FPS) and depth videos were recorded at 15 FPS, they were stored on a PC in .mp4 and .dat format respectivly.

Motion sequencing

The depth videos were then fed into a motion sequencing pipeline [5]. In short, this pipeline performs principal component analysis on depth images to identify the reused set of sub-second duration movements which can be sequenced together to form more complex behaviours. This set of behavioural motifs is sorted by using an auto-regressive, hidden Markov model which automatically recognises specific movements based on the latent structure present in the behavioural data as well as the associated transition statistics.

Ethical statement

All procedures were approved by local ethical review, carried out under a UK Home Office project licence and complied with the EU directive 63/2010E and the UK Animal (Scientific Procedures) Act 1986.

Results

The unsupervised motion sequencing pipeline was able to identify around 30, short duration, behavioural motifs which significantly differed between animal which received MK801 and those which received saline. After a human review, the animal behaviours which increased in usage with administration of MK801 closely matched the known pharmacological effect of MK801 at this dose, namely uncoordinated locomotor activity, bursts of rapid turning behaviour, head weaving and body rolling.

Administation of MK801 also significant reduced the usage of certain behaviours compared to saline treated animals, these mainly consisted of vertical movements such as rearing and sniffing but also incidence of grooming was reduced in MK801 animals.

Whilst the usage of most behaviours was similar for both Line66 and NMRI animals, the motion sequecing pipeline was able to identify a number of differences between the two genotypes. After reviewing by a human, these behaviours were found to mostly consist of exploratory rearing, and whilst Line66 animals did perform these actions, their usage was lower than NMRI animals.

Conclusions

Motion sequencing, performed on 3-D, depth videos was able to capture and quantify the nuanced changes in mouse behaviour induced by MK801, as well as, gene mutiations. Whist this experiment could benefit from larger sample sizes, the unsupervised machine learning pipeline was shown to be robust and sensitive to small changes which may not have been identified using ‘typical’ 2-D paramters and human scoring. This experiment therefore contributes to the broader validation of such methodologies in behavioural pharmacology research.

References

1. Egnor, S. E. R. & Branson, K. (2016). Computational analysis of behavior. Annu. Rev. Neurosci. 39, 217-236. https://doi.org/10.1146/annurev-neuro-070815-013845
2. Melis, V. et al (2015). Different pathways of molecular pathophysiology underlie cognitive and motor tauopathy phenotypes in transgenic models for Alzheimer's disease and frontotemporal lobar degeneration. Cell. Mol. Life. Sci. 72(11), 2199-2222. doi: 10.1007/s00018-014-1804-z
3. Liljequist S, Ossowska K, Grabowska-Andén M, Andén NE. (1991). Effect of the NMDA receptor antagonist, MK-801, on locomotor activity and on the metabolism of dopamine in various brain areas of mice. Eur J Pharmacol.195(1),55-61. doi: 10.1016/0014-2999(91)90381-y
4. Microsoft, https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows
5. Wiltschko, A.B., Tsukahara, T., Zeine, A. et al. (2020) Revealing the structure of pharmacobehavioral space through motion sequencing. Nat Neurosci 23, 1433–1443. https://doi.org/10.1038/s41593-020-00706-3

Micro-behavioral coding system for dynamic systems theory analysis of parent-child interactions in developmental disabilities

B. J. Byiers[1], J. Gunderson[2], C. Roberts[1], A. Dimian[3], & F. J. Symons[1]

1 Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA. Byier001@umn.edu; robe2020@umn.edu; symon007@umn.edu

2 Mayo Clinic, Department of Pediatric and Adolescent Medicine, Rochester, MN USA. Gunderson.Jaclyn@Mayo.edu

3 Institute on Community Integration, University of Minnesota, Minneapolis, MN, USA. Dimia006@umn.edu

Children with developmental disabilities (DD) are more likely to exhibit behavior problems including aggression and self-injurious behavior compared to typically-developing peers [1]. Although early parent-child interactions have been shown to be highly predictive of behavioral outcomes among typically developing children, it is unclear whether the same patterns apply to families of children with DD. Dynamic systems theory (DST) provides a conceptual framework for the study of change and posits that complex dynamic processes, including dyadic interactions, require a balance between stability and flexibility for optimal growth [2]. Within this framework, flexibility in parent-child interactions is adaptive, as it indicates that the dyad can engage in positive reciprocal moderation of the interaction. Although flexibility is necessary, predictability of parental behavior is also critical for optimal development [3,4], suggesting that too much variability may be detrimental, particularly within negative or hostile interactions. To test whether this framework is associated with behavioral outcomes among families of young children with DD, we have developed a micro-behavioral coding system and associated data collection protocol, as none of the existing protocols met the following criteria: 1) can be applied to interactions among children with a wide range of levels of ability, from no verbal language and limited motor/adaptive skills to children with mild developmental delays and age-appropriate speech; 2) can be applied to videos collected remotely via video conferencing software in families' homes using available toys and materials; 3) would allow us to quantify aspects of the content of the interaction (e.g., positive dyadic engagement, attractor states), and the structure of the interaction (e.g., dyadic behavioral flexibility or entropy); and 4) could be reliably coded by independent observers; and 5) is acceptable to families as indicated by continued participation in a longitudinal project.

All study procedures were approved by the local institutional review board and all participating families provided informed consent. After several iterations, we have developed a data collection protocol that involves collection of brief (5-min) semi-structured parent-child play sessions that includes three distinct tasks: an unstructured free play condition (conducted twice during each remote visit), a book sharing condition, and a demand condition during which the parent is instructed work on a challenging task with the child. The initial coding scheme was based primarily on behavioral definitions from the Dyadic Interaction Coding System [5] (Lunkenheimer, 2009) and the maternal responsivity/language development literature (e.g., Haebig et al., 2013) [6]. Behavioral codes were grouped into six categories for parent behavior (positive directing, positive following, neutral active, neutral passive, negative active, negative passive), and five categories for child behavior (positive social interaction, positive interaction with objects, neutral unengaged, neutral refusal/protest, and negative noncompliance and dysregulation). From these categories, we have created measures of positive dyadic engagement as a measure of the affective content of the interaction, and measures of flexibility (i.e., the number of transitions between dyadic behavioral states), and entropy (the unpredictability of the sequence of events).

Although data collection and analysis are ongoing, preliminary results suggest that the coding system produces relatively reliable patterns of results across sessions within dyads, that it can be coded reliably by trained independent observers, and that both the content and structural measures are associated with concurrent behavior problems as reported by the caregivers, and would be hypothesized according to the DST framework. The project has maintained relatively high rates of retention and relatively low data loss to date, supporting the overall feasibility of the approach.

References

1. Davies, L., & Oliver, C. (2013). The age-related prevalence of aggression and self-injury in persons with an intellectual disability: A review. Research in Developmental Disabilities, 34(2), 764–775. https://doi.org/10.1016/j.ridd.2012.10.0042. Spink, A.J., Tegelenbosch, R.A.J., Buma, M.O.S., Noldus, L.P.J.J. (2000). The EthoVision video tracking system: a tool for behavioral phenotyping of transgenic mice. *Physiology & Behavior* **73**, 731-744.

2. Thelen, E. (2005). Dynamic Systems Theory and the Complexity of Change. Psychoanalytic Dialogues, 15(2), 255–283. https://doi.org/10.1080/10481881509348831

3. Glynn, L. M., & Baram, T. Z. (2019). The influence of unpredictable, fragmented parental signals on the developing brain. Frontiers in Neuroendocrinology, 53, 100736. https://doi.org/10.1016/j.yfrne.2019.01.002

4. Wahler, R. G., & Dumas, J. E. (1986). Maintenance Factors in Coercive Mother-Child Interactions: The Compliance and Predictability Hypotheses. Journal of Applied Behavior Analysis, 19(1), 13–22. https://doi.org/10.1901/jaba.1986.19-13

5. Lunkenheimer, E. (2009). Dyadic Interaction Coding Manual. Colorado State University.

6. Haebig, E., McDuffie, A., & Ellis, W. S. (2013). The Contribution of Two Categories of Parent Verbal Responsiveness to Later Language for Toddlers and Preschoolers on the Autism Spectrum. American Journal of Speech-Language Pathology, 22(1), 57–70. https://doi.org/10.1044/1058-0360(2012/11-0004)

Deep Learning Behavioral Phenotyping System in the Diagnosis of Parkinson's Disease with *Drosophila melanogaster*

Keyi Dong[1], April Burch[2] and Kang Huang[3,4,*]

1. University of Southern California, 3551 Trousdale Parkway, Los Angeles, CA 90089. fionadon@usc.edu

2. Berkshire School, 245 North Undermountain Rd, Sheffield, MA. aburch@berkshireschool.org

3. Shenzhen Institutes of Advanced Technology, Shenzhen, China. huangkang314@gmail.com

4. Shenzhen Bayone BioTech Co.Ltd., Shenzhen, China 518100.

*** Corresponding Author: Dr. Kang Huang. huangkang314@gmail.com**

Abstract

Drosophila melanogaster is widely used as animal models for Parkinson's disease (PD) research. Because of the complexity of MoCap and quantitative assessment among *Drosophila* melanogaster, however, there is a technical issue that identify PD symptoms within drosophila based on objective spontaneous behavioral characteristics. Here, we developed a deep learning framework generated from kinematic features of body posture and motion between wildtype and SNCA^{E46K} mutant drosophila genetically modeled ɑ-Syn, supporting clustering and classification of PD individuals. We record locomotor activity in a 3D-printed trap, and utilize the pre-analysis pose estimation software DeepLabCut (DLC) to calculate and generate numerical data representing the motion speed, tremor frequency, and limb motion of *Drosophila melanogaster*. By plugging these data as the input, the diagnosis result (1/0) representing PD or WT as the output. Our result provides a toolbox which would be valuable in the investigation of PD progressing and pharmacotherapeutic drug development.

Keywords

Deep Learning, Behavioral Analysis, Drosophila melanogaster, Parkinson's Disease, DeepLabCut

Introduction

Parkinson's disease (PD) is a recognizable neurodegeneration with complex etiology and rapid progress. Compared with the 1990s, the age standardized prevalence rate of PD increased by 21.7% over the same period[1]. Considerable evidence in recent studies showed that the progressive degeneration of dopaminergic (DA) neurons in the substantia nigra (SN) and the parenchyma (SNpc) was the main cause of the motor symptoms [2], and the pathological hallmarks of PD are the deposition of Lewy bodies (LBs), which are mainly composed of misfolding α-synuclein (α-Syn)[3]. For this reason, several trans-genetic *Drosophila melanogaster*, such as E46K mutation of SNCA gene, have been utilized as animal models to explore the pathological mechanism of PD induced by single gene mutation [4-6].

Compared with mammalian models such as rodents and NHPs (non-human primates), PD related trans-genetic *Drosophila melanogaster* has obvious advantages in distinguished physical signs, abnormal phenotypes in motor and non-motor behaviors, and short lifespan[7]. Whereas, due to the immatureness of the hierarchical and classified observation technology of entomological behaviors with *Drosophila melanogaster*, researchers still need a quantitative and time-scale dynamic phenotypic classification of *Drosophila melanogaster* during exploring the mechanism of PD affected by multiple factors, especially for non-transgenetic models[8]. Recently, scientists have developed and optimized a variety of software toolbox to track and quantify the behavioral characteristics of small but fast-moving animals including *Drosophila melanogaster*, such as DeepGraphPose, DeepPoseKit, DeepLabCut, LEAP etc [9-12].

In current study, we compared wildtype and PD trans-genetic *Drosophila melanogaster* with E46K mutation of SNCA gene for spontaneous behavioral classification and mapping. By using the pre-analysis software

DeepLabCut (DLC), which provides marker-less pose estimation of user-defined body parts[13,14], we monitored the posture information of *Drosophila melanogaster* during walking in a 3D-printed trap. Finally, we constructed a binary diagnosis system method based on extracting and clustering the multi-dimensioned behavioral features of these two types of *Drosophila melanogaster*. We hope this project provide an objective and quantitative behavioral assessment tool for the research on neural mechanism of PD with *Drosophila melanogaster*.

Materials and Methods

Animals

Two strains of *Drosophila melanogaster* are included in this study: Oregon R strain wild type drosophila ordered from Carolina as the control (WT) group and the genetically modified α-Synuclein PD strain (SNCA^{E46K} mutant) ordered from Bloomington Drosophila Stock Center as the PD group. Both strains of drosophila are housed in test tubes, and maintained under constant conditions (12-h/12-h light/dark cycle, 23℃, and a standard yeast-sugar-based food medium). Male and female are studied separately within each group because the behavioral changes of PD are usually more obvious for male.

Behavioral apparatus and data collection

Drosophila melanogaster's movements are restricted into a specific area under the microscope with the 3D-printed trap (7mm*9mm*3mm, size decided upon the field of microscope). Videos of the WT and PD groups are recorded with a microscope and Motic software. The recording resolution was 2048×1536 at 25 frame-per-second (fps). We collected 40 animal videos in total ($N_{PD\ female}$ = 11, $N_{PD\ male}$ = 7, $N_{PD\ female}$ = 14, $N_{PD\ male}$ = 8).

Drosophila melanogaster pose estimation with DeepLabCut

Before generating the pose estimation model, we resize the resolution (resized resolution: 512×384) of these collected videos to improve the computational efficiency. Thirty *Drosophila melanogaster* videos were selected to generate the training set. We used the K-means clustering method to extract 20 frames from each video, and finally obtained a 600-frame to be labeled images. Then, we created a DeepLabCut (DLC, version: 2.2.0.6) project and configured nine pupil feature points to be detected: head, body center, tail, left front leg, left middle leg, left hind leg, right front leg, right middle leg, right hind leg (see Fig1). To train the DLC model, we used graphical user interfaces (GUIs) to manually label 600 extracted images. After completing the manual labeling step, the DLC model was trained in Python3 environment. These training steps were executed on the NVIDIA GeForce RTX 2080ti graphic processing unit. We set the maximum number of training iteration as 1030000, and it took more than 10 hours to complete these processes.

We used the trained model to track the 40 *Drosophila melanogaster* videos and obtain 9 key body parts. Since each body part has 2 dimensions, each animal contains 18-dimension behavioral trajectory data.

Behavioral data pre-processing

When collecting *Drosophila Melanogaster* videos, the distance between behavioral assay and the microscope lens was not always consistent. Therefore, for each video, we manually determined the trap boundary as ROI to ensure the consistency of the trap in the cropped image. Then we normalized the trajectory data, resizing values in the horizontal and vertical directions to a range of 0-100.

Kinematic features extraction

In order to better observe the locomotion throughout the recording process, we plotted a velocity trajectory heatmap for each animal. We used the normalized trajectory of the body center point to represent the motion of whole body. First, we calculated the frame-by-frame velocity of each animal. Then to reduce noise and improve the visualization, we also smoothed the velocity with half a second time window. Finally, we used the blue-to-red colors in the heatmap to represent the low-to-high speed (Figure. 2A).

In addition, we calculated 20 kinematic features based on the normalized 18-dimensional body trajectories, including the velocity of 9 body parts, the acceleration of 9 parts, the per frame movement distance, and the ration of the animals entering the central region. We calculated these features as the following rules:

1) *Velocity*

We set $T_{body} = \{x_t, y_t\}, t = 1{:}\,n$ to be the coordinates of one of the body points, where n is the number of frames corresponding collected video. Therefore, the velocity of this body part can be calculated as follows:

$$v_{body} = \frac{\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}}{\Delta t}$$

where Δt is the frame rate of the recorded video.

2) *Acceleration*

Similarly, the acceleration of a body part can be calculated as follows:

$$a_{body} = \frac{\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}}{(\Delta t)^2}$$

3) *Per frame distance*

We used the body center point $T_{bodycenter}$ to calculate the per frame distance, the per frame distance of a body part can be calculated as follows:

$$v_{body} = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$$

4) *Ratio in centre region*

Since we have normalized the trajectory data to the range of 0-100, for each frame, both the x and y coordinates of the body center in the range of 25-75 were considered to be in the centre. Then, the ratio in centre can be calculated as the ratio between the total number of frames in the centre and the total number of frames recorded.

Unsupervised movement clustering

Since the symptoms of some mental diseases usually manifest themselves in specific stereotyped movements, we need to convert the 18-dimension trajectory data into movement sequences. We, therefore, used the Behavior Atlas (https://behavioratlas.tech/) to perform the trajectory decomposition and unsupervised movement clustering. Because the tracked 2D *Drosophila melanogaster* trajectories were different from the 3D skeleton used in Behavior Atlas, it was difficult to perform the body alignment. Thus, we adopted part of the original method of Behavior Atlas. We calculated the dynamic time alignment kernel (DTAK) matrix of the 18 kinematic features (9 body parts velocity and 9 body parts acceleration) for each animal. After decomposing the trajectories of all animals, we calculated the 2D movement features space. Then we used hierarchical clustering to identify 10 clusters of these movement sequences.

Dimensionality reduction and classification of *Drosophila melanogaster* based on behavioral features

Twenty kinematic features and 10 movement subtypes were involved in the diagnosis of PD *Drosophila melanogaster*. To intuitively evaluate the distribution between different types of animals according to behavioral features, we performed dimensionality reduction analysis with the t-SNE (t-distributed stochastic neighbor embedding) algorithm on the average value of these features. We used dimensionality reduction two-dimensional scatters and different colors to represent two types of animals (Figure 2A, B).

In the *Drosophila Mp]mmelanogaster* classification section, we combined the kinematic features and the movement sequence fraction features into a 30×40 feature matrix. Correspondingly, the 40 PD and WT animals formed 1×40 labels (22 WT and 18 PD). We used the MATLAB (Version: 2020b, MathWorks, Massachusetts, USA) classification app and selected 6 models for training: Fine Tree, Weighted KNN (k-Nearest Neighbor), Fine KNN, Logistic Regression, Linear SVM (Support Vector Machine), Cubic SVM. Set validation to 10-fold cross-validation.

Statistics

To compare kinematics and movement sequence fractions between the two groups, we imported these values for each animal using MATLAB and used Prism 8.0 (GraphPad Software, Inc.) for statistical analysis. Before hypothesis testing, the data were first tested for normality using the Shapiro-Wilk normality test and for homoscedasticity using the F test. One-way analysis of variance with the Kruskal–Wallis's test was performed to determine which feature has a significant difference between the two groups.

Results

A Deep Learning-based System for Parkinson's Disease *Drosophila melanogaster* Diagnosis

Drosophila melanogaster PD diagnostic system based on deep learning includes three parts: 1) data collection (see in Figure 1). For WT and PD *Drosophila melanogaster*, we used a microscope to record the track spontaneous locomotion in a 3D-printed trap, and captured video of each animal. 2) *Drosophila melanogaster* pose estimation based on deep learning: based on the DeepLabCut pose estimation toolkit, 600 photos are manually marked, each of which is marked with 9 body points as a training set. Then we used the trained model to track the body trajectory of *Drosophila melanogaster*. 3) PD diagnosis of *Drosophila melanogaster* based on behavioral features: kinematics parameters of body motion and sequential activity characteristics are extracted respectively to form a matrix. Then, we reduce the dimension of visualization and use different machine learning models to achieve PD diagnosis with *Drosophila melanogaster*.

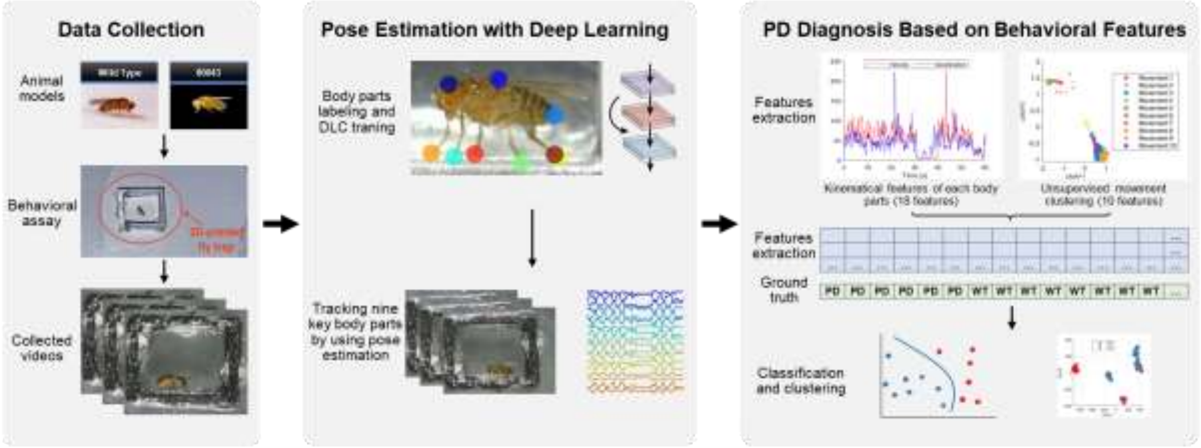

Data Collection
Animal models
Wild Type
Behavioral assay
3D-printed fly trap
Collected videos
Pose Estimation with Deep Learning
Body parts labeling and DLC traning
Tracking nine key body parts by using pose estimation
PD Diagnosis Based on Behavioral Features
Features extraction
Kinematical features of each body parts (18 features)
Unsupervised movement clustering (10 features)
Features extraction
Ground truth
PD
PD
PD
PD
PD
PD
WT
WT
WT
WT
WT
WT
WT
WT
...
Classification and clustering

Figure 1. Diagram of deep-learning-based system for Parkinson's Disease Drosophila melanogaster diagnosis.

Kinematic Features are Insufficient for the Diagnosis of PD *Drosophila melanogaster*

The variation of locomotor activity is a typical characteristic of PD in *Drosophila melanogaster*. We measure the velocity trajectory heatmap for each animal, and analyzed normalized trajectory of body center frame-by-frame to portrayal the action speed and spatial-temporal distribution of different genotype *Drosophila melanogaster* (see in Figure 2A). The color from blue to red represents speed from slow to fast. We found no difference in the velocity trajectory between WT and PD group, despite the female PD group presented higher probability of immobility. After that, we compared the kinematic features of *Drosophila melanogaster* with 20 parameters. We observed no difference in the velocity, acceleration, per frame distance, ratio in the centre individually (see in Figure 2B). These findings suggested that only kinematic features are insufficient for recognizing and distinguishing WT and PD *Drosophila melanogaster*.

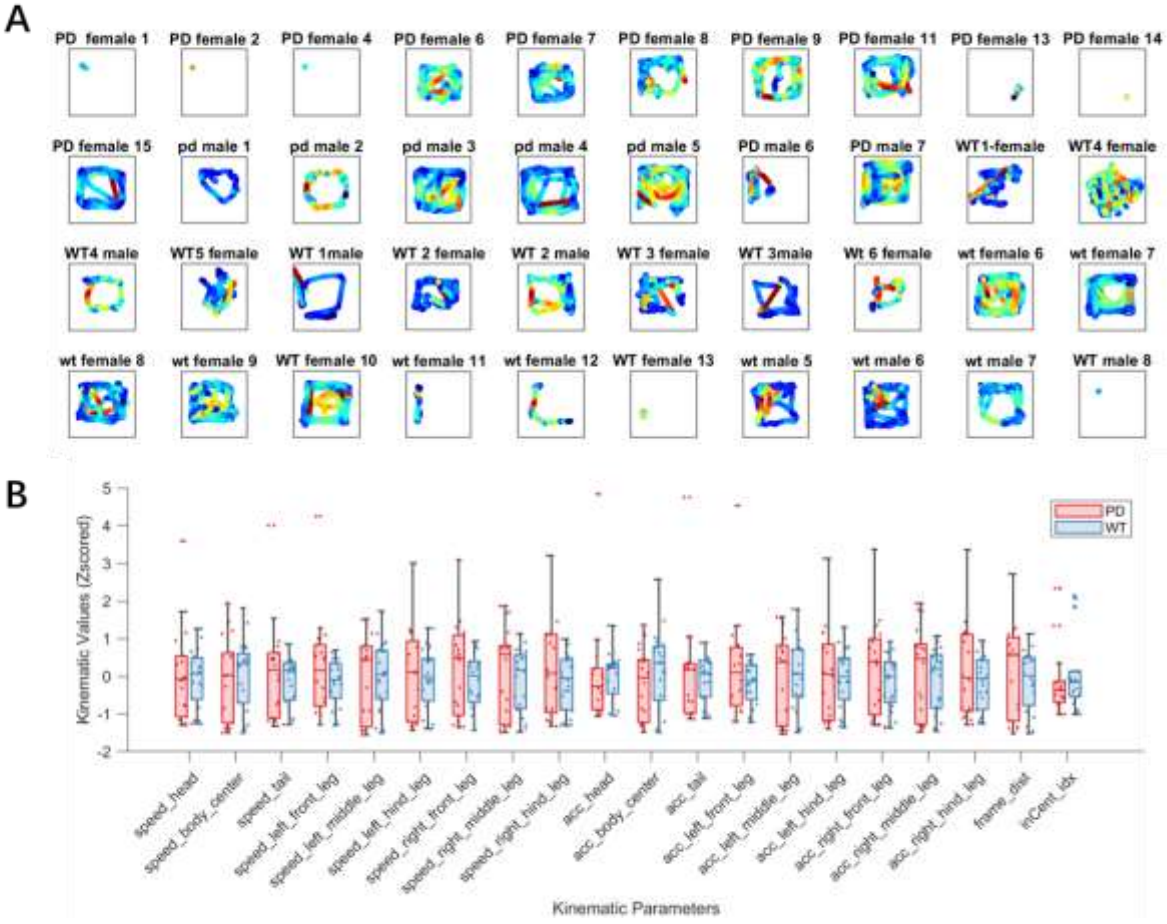


Figure 2. Comparison of kinematic parameters between WT and PD groups. A) Heatmap of velocity track of 40 *Drosophila melanogaster*. The activity track of the back point represents the overall movement. From blue to red in the heatmap, the speed is from low to high. B) Display the distribution of 20 kinematic parameters of PD and WT groups at the same scale, the values are normalized with z-score. The scatter points in the graph represents each sample. Boxplot is expressed in mean ± sem.

Significant Difference between WT and PD Groups in the Movement Sequence Level

According to previous study of PD in *Drosophila melanogaster*, stereotyped behaviors is the most obvious alteration in PD *Drosophila melanogaster* compared to WT. Therefore, we establish an 18-dimension trajectory for clustering the spontaneous behaviors into 10 movement subtypes. Since the 10 movement subtypes occur in different proportions in each animal, we calculated the movement fraction of each animal. The movement fractions are defined as the bouts number of each movement phenotype divide by the total number of movement bouts the animal occurred during the experiment. Finally, we exacted two distinctive movement subtypes from 10 clusters (see in Figure 3A, B), No.1 and No7, which represent forelimb rubbing without body center movement in a corner, and forelimb dominated fast unilateral movement respectively (see in Figure 3C). These results revealed a significant difference between WT and PD *Drosophila melanogaster* in the movement sequence.

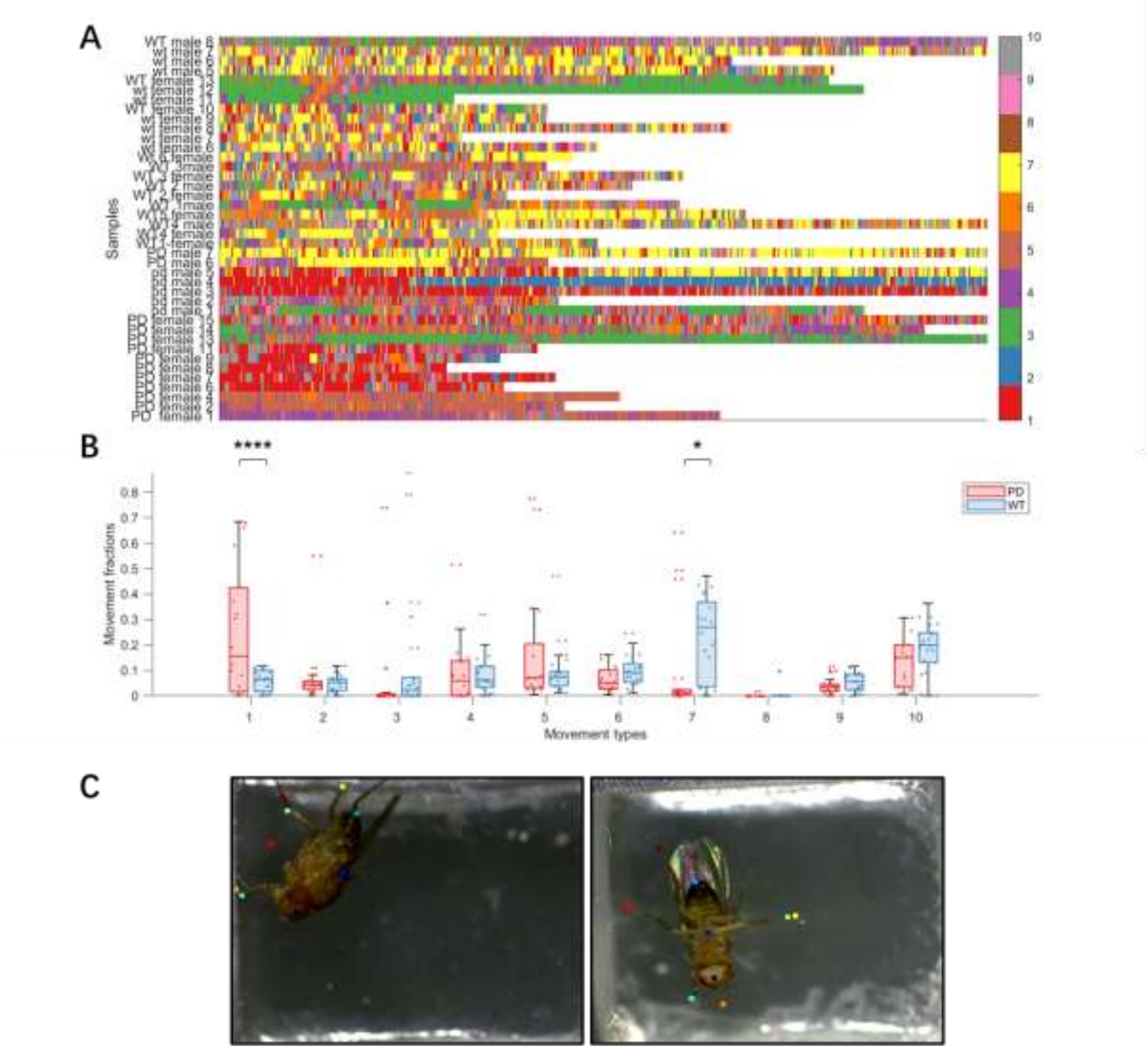


Figure 3. Comparison of movement fractions between WT and PD groups. A) Ethogram of 40 *Drosophila melanogaster*. Use 10 colors to show the types of spontaneous actions that *Drosophila melanogaster* perform during the corresponding period. Since the recording time of each animal is different, the end of the figure is not aligned. B) Comparison of 10 action types of PD and WT. Scatter points in the figure represent the fraction of each sample on this action, and boxplot is represented by mean ± sem. C) Typical behavioral performance of No.1(left) and No.7, which represent forelimb rubbing without body center movement in a corner, and forelimb dominated fast unilateral movement separately.

Classification of Drosophila melanogaster using combined features

To develop a deep learning binary PD diagnosis system based on extracting and clustering the multi-dimensioned behavioral features of these two types of *Drosophila melanogaster*, we further performed dimensionality reduction analysis with the t-SNE algorithm and testify the performance of different model for *Drosophila melanogaster* classification. The results showed the behavior characteristics of WT and PD groups could be clearly separated into two clusters after dimensionality reduction (see in Figure 4A, B). In addition, compared to other methods for classification, the Fine Tree model has the highest accuracy (85%) and true positive rate (TPR, 88.9% in PD and 81.8% in WT), and the lowest false negative rate (FNR, 11.1% in PD and 18.2% in WT) (see in Figure 4C). The other models for classification, such as Weighted KNN, Fine KNN, Logistic Regression, Linear SVM, Cubic SVM, cannot achieve the same accuracy of PD diagnosis (see in Table 1). This result verified that our deep learning binary PD diagnosis system possessed the ability to classify PD *Drosophila Melanogaster* via behavioral mapping.

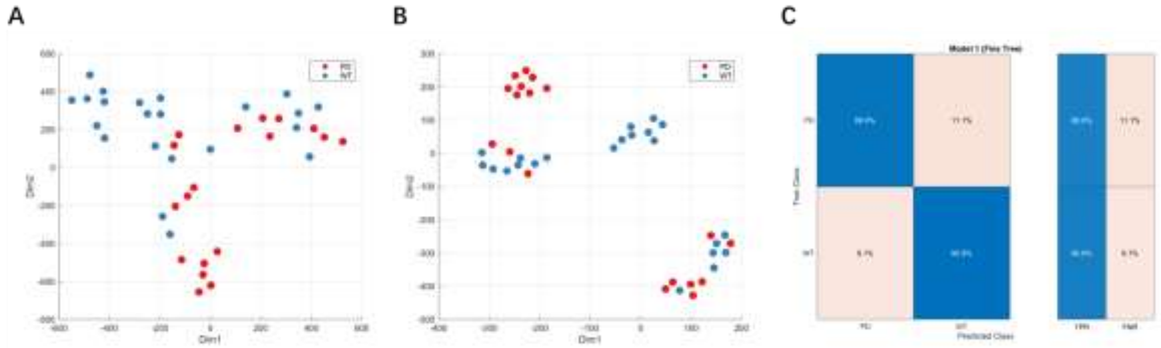


Figure 4. Clustering and classification of *Drosophila melanogaster* based on behavioral features. A & B) t-SNE algorithm is used to reduce the dimensions of kinematics and action sequences. The two colors represent clusters from different genotypes. C) The confusion matrix of Fine tree model with the best performance among the six machine learning models.

Table 1 Performance of different model for Drosophila Melanogaster classification

Performance ***Methods***	***Accuracy***	***TPR (PD, WT)***	***FNR (PD, WT)***	***AUC***
Fine Tree	85.0%	88.9%, 81.8%	11.1%, 18.2%	83.0%
Weighted KNN	62.5%	50.0%, 72.7%	50.0%, 27.3%	70.0%
Fine KNN	62.5%	50.0%, 72.7%	50.0%, 27.3%	70.0%
Logistic Regression	57.5%	66.7%, 50.0%	33.3%, 50.0%	62.0%
Linear SVM	72.5%	61.1%, 81.8%	38.9%, 18.2%	75.0%
Cubic SVM	65.0%	61.1%, 68.2%	38.9%, 31.8%	68.0%

* TPR: True Positive Rates; FNR: False Negative Rates; AUC: Area Under Curve

Discussion

Previous studies were used to observe the survival and locomotion of *Drosophila melanogaster* larvae, and climbing ability, courtship, olfactory memory of adult to explore the effects of genetic, environmental and drug on PD phenotypes. However, the traditional climbing test (negative geotaxis) is tedious, labor-intensive, time-consuming, and deliver significant differences between different experiments[15]. *Alexander Mathis et al.* developed a DeepLabCut toolbox for markerless tracking of *Drosophila melanogaster*. DeepLabCut allows accurate extraction of low-dimensional pose information from videos of freely behaving Drosophila melanogaster[14]. Meanwhile, *Anqi Wu et al.* proposed a probabilistic graphical model built on top of deep neural networks, Deep Graph Pose (DGP), to leverage these useful spatial and temporal constraints and develop an efficient structured variational approach to perform inference[12]. For accelerating processing speed and improving robustness, *Jacob M Graving et al.* used an efficient multi-scale deep learning model, StackedDenseNet, for developing a novel friendly used toolkit DeepPoseKit, which increase processing speed by 2 times without loss of accuracy[11].

In current study, we used the pre-analysis animal pose estimation software DeepLabCut (DLC) to transform the video data into numerical data representing the motion speed, tremor frequency, and limb motion of *Drosophila melanogaster*. We initially observed the locomotion dynamics of *Drosophila melanogaster* in 3D-printed trap, and compared the motion trajectory with 20 kinematic parameters of limbs speed and location. According to an unsupervised technology to discover and track the stereotypical behavior of drosophila recently developed by *Gordon J. Berman et al.*, we compared the behavioral sequence of WT and PD *Drosophila melanogaster* with their alternative internal states[16]. We identified and clustered 10 typical behavioral features and found that PD *Drosophila Melanogaster's* often perform forelimb rubbing and rapid shrinking abnormally, while forelimb-dominated rapid lateral movement significantly decreased. These phenotypes are similar to the bradykinesia and tremor in PD patients. The dichotomous PD diagnosis system developed in this project takes the massive data collected by DLC as input and outputs the diagnosis results. Moreover, we identified the best algorithm of clustering and classification is the Fine tree, of which the accuracy of prediction is the highest, up to 90%.

Parkinson's disease is the second most prevalent neurodegeneration in the world. Not only PD cause a great burden of patients, but also lead to a significant increase in the risk of suicide[17]. Actually, the pandemic of Covid-19 also has a severe impact on the central nervous system, such as disrupting the blood brain barrier[18]. The spread of Covid-19 also challenges the prevention and the hospitalized management of PD patients[19,20]. The pathological progress of PD triggered by poisoning, infection and social stress is different from that with gene mutation. PD induced by environmental factors progresses slowly, with hidden or inaccurate symptoms. It will be hard to study the dose-effect of pharmacological therapy, the mechanism of chronic environmental stress in *Drosophila melanogaster* without objective and continuous dynamic observation. Our project has successfully constructed a data-driven diagnostic toolbox for PD research with *Drosophila melanogaster*. However, in future studies on this topic, larger arena should be used to measure the kinematic features as the small-sized arena used in this experiment may have restricted *Drosophila melanogaster* movement.

References

1. Collaborators, G. B. D. P. s. D. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* **17**, 939-953, doi:10.1016/S1474-4422(18)30295-3 (2018).
2. Kalia, L. V. & Lang, A. E. Parkinson's disease. *Lancet* **386**, 896-912, doi:10.1016/S0140-6736(14)61393-3 (2015).
3. Ma, M., Moulton, M. J., Lu, S. & Bellen, H. J. 'Fly-ing' from rare to common neurodegenerative disease mechanisms. *Trends Genet* **38**, 972-984, doi:10.1016/j.tig.2022.03.018 (2022).
4. Tang, Y., Tahmasebinia, F. & Wu, Z. Evaluation of Mitochondrial Function and Morphology in Drosophila. *Methods Mol Biol* **2322**, 195-206, doi:10.1007/978-1-0716-1495-2_19 (2021).
5. O'Hanlon, M. E. *et al.* Mitochondrial electron transport chain defects modify Parkinson's disease phenotypes in a Drosophila model. *Neurobiol Dis* **171**, 105803, doi:10.1016/j.nbd.2022.105803 (2022).
6. Stoessl, A. J., Lehericy, S. & Strafella, A. P. Imaging insights into basal ganglia function, Parkinson's disease, and dystonia. *Lancet* **384**, 532-544, doi:10.1016/S0140-6736(14)60041-6 (2014).
7. Naz, F. & Siddique, Y. H. Drosophila melanogaster a Versatile Model of Parkinson's Disease. *CNS Neurol Disord Drug Targets* **20**, 487-530, doi:10.2174/1871527320666210208125912 (2021).
8. Bolus, H., Crocker, K., Boekhoff-Falk, G. & Chtarbanova, S. Modeling Neurodegenerative Disorders in Drosophila melanogaster. *Int J Mol Sci* **21**, doi:10.3390/ijms21093055 (2020).
9. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat Methods* **16**, 117-125, doi:10.1038/s41592-018-0234-5 (2019).
10. Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat Methods* **19**, 496-504, doi:10.1038/s41592-022-01443-0 (2022).
11. Graving, J. M. *et al.* DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **8**, doi:10.7554/eLife.47994 (2019).
12. Anqi Wu, E. K. B., Matthew R Whiteway, Michael Schartner, Guido Meijer, Jean-Paul Noel, Erica Rodriguez, Claire Everett, Amy Norovich, Evan Schaffer, Neeli Mishra, C. Daniel Salzman, Dora Angelaki, Andrés Bendesky. Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose tracking. doi:https://doi.org/10.1101/2020.08.20.259705 (2020).
13. Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152-2176, doi:10.1038/s41596-019-0176-0 (2019).
14. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281-1289, doi:10.1038/s41593-018-0209-y (2018).
15. Nichols, C. D., Becnel, J. & Pandey, U. B. Methods to assay Drosophila behavior. *J Vis Exp*, doi:10.3791/3795 (2012).
16. Berman, G. J., Bialek, W. & Shaevitz, J. W. Predictability and hierarchy in Drosophila behavior. *Proc Natl Acad Sci U S A* **113**, 11943-11948, doi:10.1073/pnas.1607601113 (2016).

17. Erlangsen, A. *et al.* Association Between Neurological Disorders and Death by Suicide in Denmark. *JAMA* **323**, 444-454, doi:10.1001/jama.2019.21834 (2020).
18. Denaro, C. A. *et al.* COVID-19 and neurodegeneration: The mitochondrial connection. *Aging Cell*, e13727, doi:10.1111/acel.13727 (2022).
19. Leta, V. *et al.* Covid-19 and Parkinson's disease: Acute clinical implications, long-COVID and post-COVID-19 parkinsonism. *Int Rev Neurobiol* **165**, 63-89, doi:10.1016/bs.irn.2022.04.004 (2022).
20. Scherbaum, R. *et al.* COVID-19 outcomes in hospitalized Parkinson's disease patients in two pandemic waves in 2020: a nationwide cross-sectional study from Germany. *Neurol Res Pract* **4**, 27, doi:10.1186/s42466-022-00192-x (2022).

Automated Recordings of Neural Activity in a Naturalistic Setting: the Drivemaze

C. Hartmann, H. Mansvelder and M. Karnani

Department of Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. c.s.e.hartmann@vu.nl

Motivated behaviors such as feeding, drinking, and social interaction are modulated by internal needs and external factors which fluctuate throughout the day. In particular, the presence of the experimenter can have a profound influence on the expression of these behaviors, as well as the neural activity underlying the expressed behaviors. In this study, we demonstrate an open-design based device, the Drivemaze, which allows standardized recording of motivated behaviors in mice. This is achieved through discretization of motivated behaviors in a naturalistic burrow-like environment and automated recordings, thus eliminating the need for the experimenter's constant presence.

The Drivemaze allows individual mice to choose between distinct chambers, containing various goal-objects, such as food, water, an exercise wheel, and social contact with other mice (see Figure 1). The modular construction of the Drivemaze allows the researcher to change object locations and availability, thus mimicking a natural habitat. Animals enter a central decision point from a home cage through a single-entry module, which automatically identifies and weighs the animal. Within the maze, the animal must return to the central decision point before switching to a new goal or repetitively exploiting the same goal. Such spatial discretization of natural behaviors allows us to separate appetitive and consummatory aspects with the aim of studying their neural control [1,2]. The Drivemaze also incorporates open-source components, such as the feeding experimentation device 3 (FED3)[3] and the UCLA V4 miniscope for neural recordings.

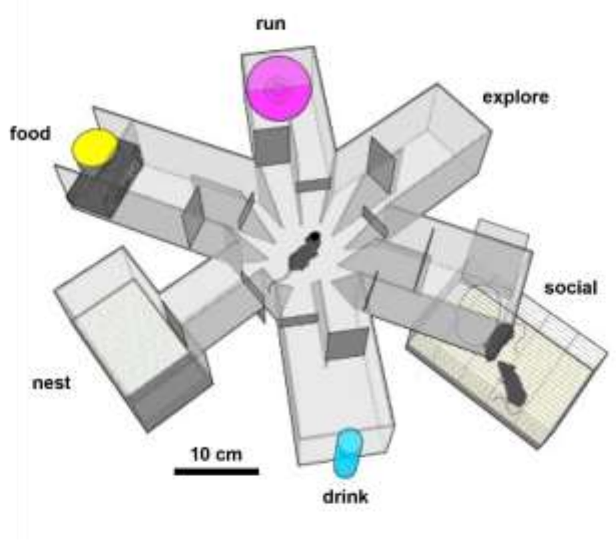


Figure 1. Overview of the Drivemaze.

We describe the behavioral patterns of mice living continuously within the Drivemaze (see Figure 2B), and go on to demonstrate the utility of this approach by recording, at single-cell resolution, the firing rates of prefrontal cortex (PFC) neurons which project to the lateral hypothalamus (LH) (see Figure 2A/C). We do so, by applying a dual viral strategy to selectively express a genetically encoded calcium indicator (GCamp8f) in PFC neurons projecting to the LH. This neuronal population is then recorded *in vivo* using miniscopes and gradient-index lenses implanted into the PFC.

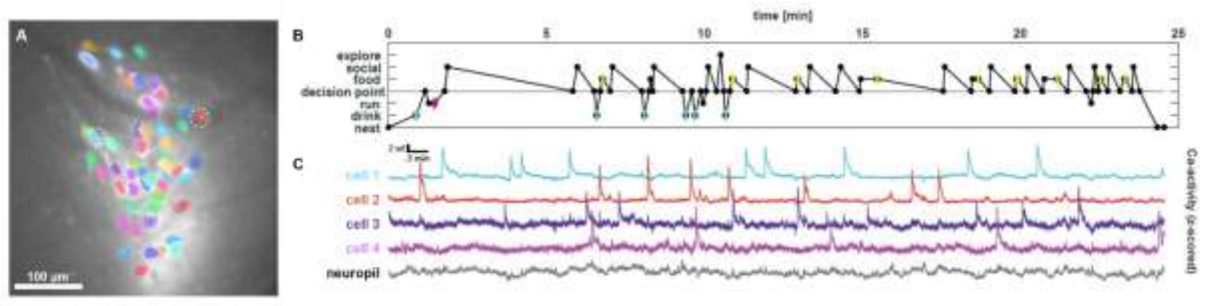


Figure 2. Ethogram and calcium imaging traces of one animal. A) Field of view of one animal. B) Ethogram of one animal showing a standard maze entry. C) Calcium traces of four example cells and their average neuropil signal. Color-coded to cells in panel A.

The LH has been shown to play a central role in homeostasis and motivated behaviors [4]. As there is minimal local synaptic communication in the LH [5], upstream glutamatergic neurons in the PFC are critical regulators of LH output [6]. We find that specific subsets of PFC→LH projection neurons are active during appetitive and consummatory behavioral epochs. Consummatory behavior, however, produced stronger, more robust responses and recruited more cells. Specifically during continuous feeding, a proportion of cells shows activity that is time-locked to food pellet retrievals. Taken together, these data suggest that the PFC encodes various aspects of goal-directed behavior and potentially regulates meal length and/or size through its connection to the LH.

In conclusion, this research provides insight into the neural mechanisms that govern motivated behavior. Leveraging *in vivo* calcium imaging in a naturalistic environment, our study sheds light on the coordination of LH activity via forebrain inputs, specifically from the PFC. Our findings demonstrate the feasibility and potential of ethologically relevant recording of neural populations. Understanding these circuits could unveil potential targets for addressing dysregulated motivated behavior, e.g. in obesity.

Ethical Statement

All experimental procedures were approved by the Netherlands Central Committee for Animal Experiments and the Animal Ethical Care Committee of the Vrije Universiteit Amsterdam (AVD11200202114477).

References

1 J. H. Jennings, R. L. Ung, S. L. Resendez, A. M. Stamatakis, J. G. Taylor, J. Huang, et al. (2015). Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell*. **160**, 516–527.

2 J. N. Siemian, M. A. Arenivar, S. Sarsfield, C. B. Borja, C. N. Russell, Y. (2021). Aponte Lateral hypothalamic LEPR neurons drive appetitive but not consummatory behaviors. *Cell Reports*. **36**, 109615.

3 B. A. Matikainen-Ankney, T. Earnest, M. Ali, E. Casey, J. G. Wang, A. K. Sutton, et al. (2021). An open-source device for measuring food intake and operant behavior in rodent home-cages. *Elife*. **10**, e66173.

4 M. A. Rossi Control of energy homeostasis by the lateral hypothalamic area. (2023) *Trends in Neurosciences*. **46**, 738–749.

5 D. Burdakov, M. M. Karnani Ultra-sparse Connectivity within the Lateral Hypothalamus. (2020) *Current Biology*. **30**, 4063-4070.e2.

6 J. D. Hahn, L. W. Swanson Distinct patterns of neuronal inputs and outputs of the juxtaparaventricular and suprafornical regions of the lateral hypothalamic area in the male rat. (2010) *Brain Research Reviews*. **64**, 14–103.

Analyzing the Relationship between Selected Flight Parameters in the Saker Falcon (*Falco cherrug*)

T. Hegerová[1], A. Juráška[2] and P. Juhás[1]

1 Institute of Animal Husbandry, Slovak University of Agriculture in Nitra, Nitra, Slovakia; 949 01. terezia.hegerova@uniag.sk / peter.juhas@uniag.sk

2 Astur Falconry, Bratislava, Slovakia; 851 07. astur@astur.sk

Abstract

The aim of the study was to evaluate the relationship between selected flight parameters of the Saker Falcon (*Falco cherrug*). A positive correlation was found between maximum speed and flight altitude (r = 0.217), from which the flight mode of falconiformes birds of prey can be inferred. A positive correlation was also found between maximum flight speed and distance flown (r = 0.241) and between maximum speed and duration of flight (r = 0.238).

Introduction

The Saker Falcon (*Falco cherrug*) is classified among the larger falcons, in size between the Gyr Falcon (*Falco rusticolus*) and the Peregrine Falcon (*Falco peregrinus*) [2, 9]. The Saker Falcon is one of the most commonly used birds of prey for falconry purposes [3, 11]. Their use can be observed at airports, where they are employed to scare off wild birds, thereby reducing the risk of bird strikes with aircraft [8, 10]. Falcons are also used in biological protection for agricultural areas or to eliminate unwanted unmanned aerial vehicles (UAVs). They are similarly used in the entertainment industry or for educational purposes [10]. Birds of prey are highly complex and intelligent species, prone to undesirable behavioral changes if kept in improper conditions or if incorrect training methods are used [7]. The aim of the study was to evaluate the relationship between selected flight parameters of the Saker Falcon (*Falco cherrug*), thereby providing an analysis for better training settings for the Saker Falcon used for falconry purposes.

Material and methods

We conducted our research at the Astur falconry centre at Red Rock Castle over a period of six months (June – November 2023). Flight parameters of the bird of prey were obtained using the Marshall Turbo GPS System. Statistical analysis was carried out using the SPSS software - Pearson correlation coefficient (r). The selected flight parameters include total flight distance, maximum speed achieved, maximum altitude, average speed, duration of flight, and duration of free flight. The statement from the ethics committee is not required for the specified research, as it involved the usual process of training a bird of prey based on the authorization for the practice of falconry - falconry exam.

Results and discussion

During the examined period, the average weight of the Saker Falcon was 759.36 g (with a standard deviation of 25.70 g), where the maximum recorded weight was 832 g and the minimum was 725 g. A similar weight was recorded in the study by [4], who used a uniform weight average for both males and females, resulting in a higher average weight of 915 g (with a standard deviation of 175 g), as our studied bird of prey was a male. The falcon flew an average distance of 4.484 km (with a standard deviation of 2.516 km) in a single flight, where the maximum recorded distance was 16.6 km and the minimum was 1.1 km. In our study, the falcon flew 4 times the distance compared to the study by [5]. The average value of the maximum speed of the falcon during flights was 108.25 km/h (with a standard deviation of 20.096 km/h), where the highest recorded maximum speed was

166 km/h and the lowest was 70 km/h. [6] reports a speed of 120 – 150 km/h for the Saker Falcon. The falcon reached an average altitude of 53.65 m (with a standard deviation of 29.408 m), where the highest altitude was 153 m and the lowest was 12 m. The average value of average speeds during the flight was 41.32 km/h (with a standard deviation of 3.949 km/h), where the highest average speed was 50 km/h and the lowest was 26 km/h. The average duration of a single flight of the falcon was 5 minutes and 55 seconds (with a standard deviation of 3 minutes and 16 seconds), where the longest recorded flight lasted 21 minutes and 26 seconds and the shortest only 5 seconds. The average duration of free flight of the falcon was 5 minutes and 20 seconds (with a standard deviation of 3 minutes and 35 seconds), where the longest recorded flight lasted 21 minutes and 26 seconds and the shortest lasted only 1 minute and 3 seconds. A similar flight duration for studying the training of birds of prey was chosen by [5], where their birds of prey flew for 5 minutes. We were able to identify several significant correlations between selected flight parameters at the level ($P < 0.05$). The strongest correlational relationships were recorded between parameters such as total distance flown, duration of flight, and duration of free flight, but since these parameters logically relate to each other, we will not directly address their relationship. Interestingly, a positive correlation was found between maximum speed and flight altitude ($r = 0.217$), from which the flight mode of falconiform birds of prey can be inferred, utilizing altitude to gain higher speed when attacking prey, confirming our hypothesis also supported by [1]. Positive correlations were also found between maximum flight speed and distance flown ($r = 0.241$) and between maximum speed and duration of flight ($r = 0.238$). These positive correlations may also be related to the current condition of the falcon, where it can exert enough energy to achieve higher speed and subsequently fly a greater distance with a longer flight duration.

Conclusion

Based on our study, we can conclude that some flight characteristics of birds of prey *Falconiformes* are closely interconnected. We have identified significant relationships between various flight parameters such as flight altitude and maximum speed. In the next phase of our study, we will focus on comparing the identified flight parameters with environmental characteristics to better understand how external factors influence the flight behavior of falconiformes. We will also include the current weight of the bird of prey in the study and determine the optimal feeding dose. The aim of the contribution is not the extrapolation of results to the species but the verification of the methodology for measuring flight parameters.

Acknowledgments

We are grateful to the management and staff of the Astur falconry centre at Red Rock Castle in Slovakia. This research was funded within the project of Grant Agency of Faculty of Agrobiology and Food Resources, SUA in Nitra (O-23-102/0100-02) and supported by VEGA 1/0572/22.

References

1. Dekker, D. (2009). Hunting tactics of Peregrines and other falcons. Wageningen University and Research. ISBN: 978-90-8585-328-2

2. Dixon, A. (2012). Conservation of the Saker Falcon Falco cherrug and the use of hybrids for falconry. Aquila, 119, 9-19. https://api.semanticscholar.org/CorpusID:54802418

3. Eastham, C. P., Nicholls, M. K., & Fox, N. C. (2002). Morphological variation of the saker (Falco cherrug) and the implications for conservation. Biodiversity & Conservation, 11, 305-325. https://doi.org/10.1023/A:1014566024582

4. Gaudio, E., Franceschinis, C., McKinney, P., & Azmanis, P. (2023). Anesthetic effects of dexmedetomedineketamine sedation followed by isoflurane induction and maintenance in the saker falcon (Falco cherrug). Journal of Exotic Pet Medicine, 47, 27-33. https://doi.org/10.1053/j.jepm.2023.07.004

5. Granati, G., Cichella, F., & Lucidi, P. (2021). High-Tech Training for Birds of Prey. Animals, 11(2), 530. https://doi.org/10.3390/ani11020530

6. Hekman, V. (2005). Falco cherrug,. Animal Diversity Web. https://animaldiversity.org/accounts/Falco_cherrug/

7. Jones, M. P., & Heidenreich, B. (2021). Behavior of birds of prey in managed care. Veterinary Clinics: Exotic Animal Practice, 24(1), 153-174. https://doi.org/10.1016/j.cvex.2020.09.007

8. Kitowski, I., Grzywaczewski, G., Cwiklak, J., Grzegorzewski, M., & Krop, S. (2011). Falconer activities as a bird dispersal tool at Deblin Airfield (E Poland). Transportation Research Part D: Transport and Environment, 16(1), 82-86. https://doi.org/10.1016/j.trd.2010.07.010

9. Nittinger, F., Gamauf, A., Pinsker, W., Wink, M., & Haring, E. (2007). Phylogeography and population structure of the saker falcon (Falco cherrug) and the influence of hybridization: mitochondrial and microsatellite data. Molecular Ecology, 16(7), 1497-1517. https://doi.org/10.1111/j.1365-294X.2007.03245.x

10. Panter, C. T., Jones, G. C., & White, R. L. (2023). Trends in the global trade of live CITES-listed raptors: Trade volumes, spatiotemporal dynamics and conservation implications. Biological Conservation, 284, 110216. https://doi.org/10.1016/j.biocon.2023.110216

11. Samour, J., Silvanose, C., Pendl, H., & Bailey, T. (2016). Clinical and laboratory diagnostic examination. Avian Medicine, Third Ed.(J. Samour, Editor), Elsevier Publishers, New York, NY, USA, 73-178. https://doi.org/10.1016/B978-0-7234-3832-8.00006-7

Measuring Variability of Behavioral Traits, Rectal Temperature, and Heart Rate After Delivery in Holstein Cattle.

P. Juhás[1], T. Hegerová[1]

1Institute of Animal Husbandry, Slovak University of Agriculture in Nitra, Nitra, Slovakia. Peter.Juhas@uniag.sk

Abstract

Presented work focused to evaluate the measures of variability of behavioral traits during the first two hours after delivery, and the variability and changes in rectal temperature (RT) and heartbeat rate (HR) during the first hour after delivery. The highest variability was recorded for mean duration of locomotion bout (unbiased coefficient of variation CV_U = 249.86%). The RT shoved the lowest variability (CV ranged from 1.14% to 1.3%).

Introduction

The estimation a calf's future performance potential immediately after birth is a important contribution to farm management. Dystocia and consequential complications are reason for significant economic losses each year [1, 2]. The prolongation of second stage labor duration phase [3, 4, 5] and gestation length [1, 6] are the main causes of reduced viability. The set of different physiological and behavior traits, e. g. heart rate and respiration rate, oxygenation, mucous membrane color, response to stimuli, suckling, standing [7, 8, 9] are used for determination postnatal vitality in a calf. The standing and lying are commonly published behavioral indicators connected to vitality [4, 10]. Calves with assisted at delivery have longer latencies to first attempt to stand, standing and locomotion [3]. A necessary condition for estimating differences between individuals is sufficient variability in the traits evaluated. Aim of this pilot study was analyze the variability of selected behaviors, rectal temperature and hearth rate in Holstein dairy calves in Slovakia.

Material and methods

The pilot study was performed at the dairy farm of the Slovak University of Agriculture in Nitra (Oponice, Slovakia). Twenty-two Holstein calves (12 heifers, 10 bulls) were used for observation. The difference between heifers and bulls was recorded in birth weight (median M = 7.45 kg resp. 43.9, $P < 0.01$) and gestation length (M = 276.5 days resp. 280.5 days). Calving ease was rated by 3-point scale, score 2 p. (n = 9) gains calves delivered without any support, score 1 p. (n = 3) gains calves delivered by injection of Sensiblex® but without assistance, score 0 p. (n = 10) gains calves delivered by injection of Sensiblex® and with assistance. Behavior of calves was recorded by two video cameras Planet ICA-HM316W connected to Planet NVR-401 network video recorder. Video records were analyzed, and behaviors were scored by BORIS 8.20.6 software [11]. The recorded and analyzed behavior were: lying (state) – lying fully on body side or upright on sternum or partially standing with front legs under body and lying down without reaching full standing; standing (state) – standing with all 4 legs touching ground; locomotion (state) – body in standing position and calf does at least one step; standing attempt (point event) – body movement with all four legs under body, ventral part do not touch ground; first standing (point event) – all four legs are at ground, fully extended for at least 5 seconds; head movement (point event) – movements of the head on the ground in the lying position before the first lift of the head; first head lifting (point event) – head lifted from the ground to an upright position for at least 5 seconds. Behavior was analyzed in two hours after delivery. The heartbeat rate (HR) was measured by a stethoscope on the left side of the chest (dual-tubed, Kruuse brand). The rectal temperature (RT) was taken using a Microlife brand waterproof digital thermometer, inserted 2.5 cm deep into the rectum. The HR was measured before RT because of minimize calf stress during the measurement. Th HR and RT were measured 4 times (M1 – M4) in 15 minutes intervals, first measure (M1) was taken 15 minutes after delivery. Statistical analyses were performed in the IBM SPSS 26. Variability of measured traits was evaluated by unbiased coefficient of variation $CV_U = (1+1/4n)*CV$, CV – coefficient of variation, n = sample size [12].

Observations of the behavior of calves were made only from video recordings obtained from a commercial dairy farm. Heart rate measurements and saliva sampling were performed under the supervision of the veterinarian as part of the calf's routine postpartum care. This article does not contain any studies with human or animal subjects and did not require IACUC/IRB/Ethic committee approval.

Results and discussion

There was no difference between sexes in any of the evaluated behavior trait nor rectal temperature (RT) and heart rate (HR). The longest recorded behavior was lying (median M = 6066.25 sec, min = 29598.34 sec, max = 7200 sec., n = 22). Eight of calves were lying whole observation period – 7200 sec. Latency of head lifting varied from 11 sec. to 16 min. 39 sec., M = 01:33 (mm:ss). All 22 calves performed attempt to stand up within first two hours after delivery, latency M = 0:09:13 (h:mm:ss), min = 0:02:11, max = 1:20:37. Standing was recorded in 14 calves (M = 1534,27 sec., min = 19,47 sec., max = 2841,40 sec., n = 14). Latency of standing varied from 0:17:27 (h:mm:ss) to 1:59:00, M = 0:38:42. Locomotion was recorded in 13 calves (M = 435,33 sec., min = 13,49 sec., max = 2562,34 sec., n = 13). Locomotion was divided into several bouts (M = 72, min = 13, max = 163). The highest variability was recorded in duration of one locomotion bout CV_U = 249.86%. The mean of one locomotion bout varied from 0.48 sec. to 197.1 sec. Second highest variability was recorded in proportion of locomotion duration (CV_U = 166.53%). High variability, over 100%, was recorded in total duration of locomotion (CV_U = 112.22%), latency of head lift (CV_U = 138.79%), latency of standing attempt (CV_U = 137.3%), number of head movements (CV_U = 107.63%) and proportion of standing duration (CV_U = 104.47%).

Rectal temperature showed decreasing tendency, M1 RT M = 39.3°C, M2 RT M = 39.0°C, M3 RT M = 38.7°C and M4 RT M = 38.55°C, $P < 0.001$, see Figure 1. Similar results published [13], temperature decreases depending on the time and is stabilized at optimal values, ranged between 37.8 – 39.2°C. Heartbeat rate was stable, in all measurements (M1 – M4) ranged between 136 – 214 bpm, M = 160, differences among M1 – M4 are not significant. The highest HR was recorded in M2 (M = 168 bpm), result should be connected to increased activity of calves (attempts to standing, standing). Similar results reported [14, 15]. Physiological traits showed very low variability. The lowest variability was recoded in RT 30 minutes after delivery (M2), CV_U = 1.14%, the highest variability of RT was recoded in M1 and M4, CV_U = 1,3%. Heartbeat rate had slightly higher variability, the highest value was recorded 30 minutes after delivery (M2), CV_U = 11.11%. The lowest HR CV_U = 8.98%, recorded in M1.

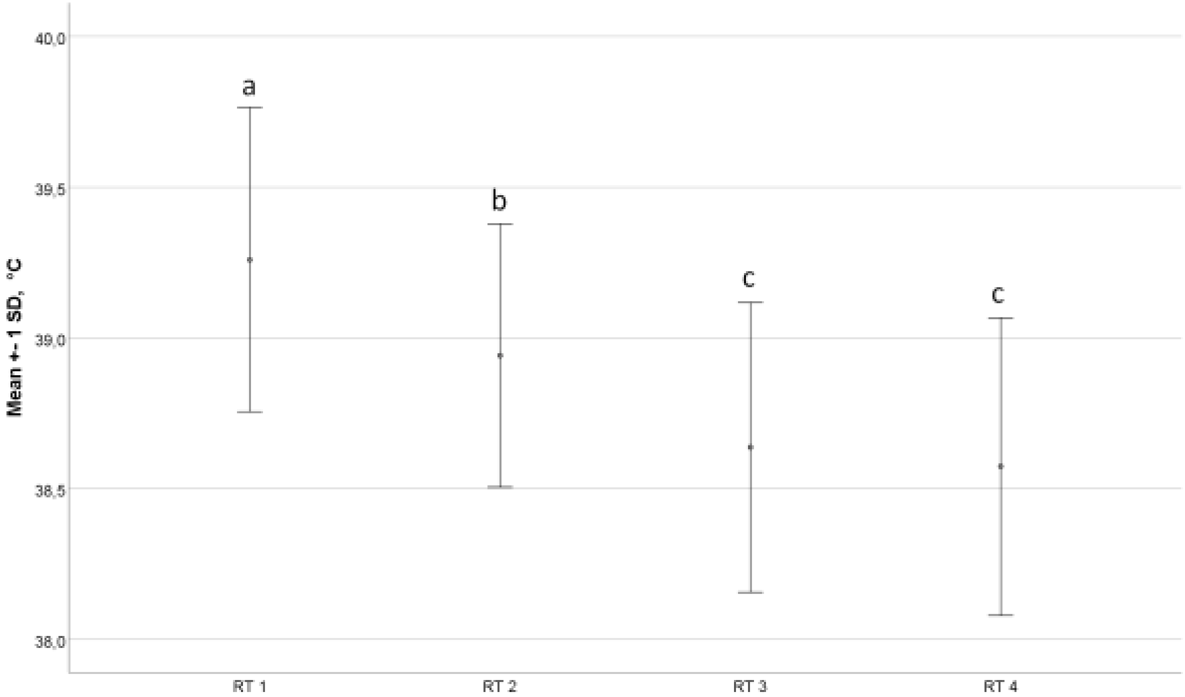


Figure 1. Decreasing tendency of rectal temperature in calves during first hour after delivery, n = 22. Different letters correspond to significant differences in the paired t-test at $P < 0.001$.

Behavior of calves in first two hours was influenced by delivery duration and calving easy. The most significant impact is on standing, calves with longer delivery had shorten total duration of standing ($r = -0.653$, $P < 0.05$). Calves after easier delivery has longer standing duration ($\rho = 0.603$, $P < 0.05$), shorten standing latency ($\rho = -0.599$, $P < 0.01$) and shorten interval between 1st standing attempt and standing ($\rho = -0.665$, $P < 0.01$). Similar impact to standing was reported by [3, 4].

Conclusion

Seven behavior traits associated with activity in calves reached appropriate high variability, suitable for identifying differences between individuals. Physiological traits showed low levels of variability and the feasibility of using it for the development of a viability index together with behavior traits must be validated by statistical methods in further research. Behavior traits can be analyzed and evaluated by advanced methods of live stream video analysis and used for online evaluation of viability in newborn calves.

The next research will be focused to relation among behavior of newborn calves and lactation performance, reproduction traits and health of heifers.

Acknowledgments

We are grateful to the management and staff of the Dairy farm Slovak University of Agriculture in Nitra for their help. The research was supported by VEGA 1/0572/22.

References

1. Meyer C. L., P. J. Berger, K. J. Koehler, J. R. Thompson and C. G. Sattler. (2001). Phenotypic Trends in Incidence of Stillbirth for Holsteins in the United States. J. Dairy Sci. 84:515–523. https://doi.org/10.3168/jds.s0022-0302(01)74502-x.

2. Umaña Sedó, S.G., Winder, C.B., Renaud, D.L. (2023). Graduate Student Literature Review: The problem of calf mortality on dairy farms. J. Dairy Sci., Vol. 106, Issue 10, p. 7164-7176. https://doi.org/10.3168/jds.2022-22795

3. Barrier, A.C., E. Ruelle, M.J. Haskell and C.M. Dwyer. (2012). Effect of a difficult calving on the vigour of the calf, the onset of maternal behaviour, and some behavioural indicators of pain in the dam. Prev. Vet. Med. 103:248– 256. https://doi.org/10.1016/j.prevetmed.2011.09.001.

4. Kovács, L., F. L. Kézér, S. Bodó, F. Ruff, R. Palme and O. Szenci. (2021). Salivary cortisol as a non-invasive approach to assess stress in dystocic dairy calves. Scientific Reports volume 11:62000. https://doi.org/10.1038/s41598-021-85666-9

5. Reis, M. E., M. Cantor, C. M. M. Bittar and J. H. C. Costa. (2022). Association of a green tea extract with serum immunoglobulin G status and neonatal vitality in newborn dairy calves. J. Dairy Sci. 105:9961–9970. https://doi.org/10.3168/jds.2022-22099.

6. Herfen, K., H. Bostedt. (1999). Correlation between clinical and laboratory diagnostic evaluation of the vitality of newborn calves under particular consideration of length and type of parturition. Wiener Tierärztliche Monatsschrift 86(8):255–261.

7. Murray, Ch. F. and K. E. Lesslie. (2013). Newborn calf vitality: Risk factors, characteristics, assessment, resulting outcomes and strategies for improvement. The Vet. J. 198:322–328. https://doi.org/10.1016/j.tvjl.2013.06.007.

8. Murray, Ch. F., D. M. Veira, A. L. Nadalin, D. M. Haines, M. L. Jackson, D. L. Pearl and K. E. Leslie. (2015). The effect of dystocia on physiological and behavioral characteristics related to vitality and passive transfer of immunoglobulins in newborn Holstein calves. Can. J. Vet. Res. 79(2):109–19. https://pubmed.ncbi.nlm.nih.gov/25852226/.

9. Homerosky, E. R., E. Timsit, E. A. Pajor, J. P. Kastelic and M. C. Windeyer. (2017). Predictors and impacts of colostrum consumption by 4 h after birth in newborn beef calves. Vet. J. 228:1–6. https://doi.org/10.1016/j.tvjl.2017.09.003.

10. Campler, M., L. Munksgaard , and M. B. Jensen. (2015). The effect of housing on calving behavior and calf vitality in Holstein and Jersey dairy cows. J. Dairy Sci. 98:1797–1804. http://dx.doi.org/10.3168/jds.2014-8726.

11. Friard, O., Gamba M. (2016). BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods in Ecology and Evolution. 7 (11), 1325-1330. DOI: https://doi.org/10.1111/2041-210x.12584

12. Sokal R.R., Rohlf F.J. Biometry (3rd Ed). New York: Freeman, (1995). p. 58. ISBN 0-7167-2411-1

13. Vermorel, M., Dardillat, C., Vernet, J., Renseigné, N., & Demigne, C. (1983). Energy metabolism and thermoregulation in the newborn calf. In Annales de Recherches Veterinaires (Vol. 14, No. 4, pp. 382-389). https://doi.org/10.4141/cjas89-013.

14. Vannucchi, C. I., Silva, L. C. G., Unruh, S. M., Lúcio, C. F., & Veiga, G. A. L. (2018). Calving duration and obstetric assistance influence pulmonary function of Holstein calves during immediate fetal-to-neonatal transition. PloS one, 13(9), e0204129. https://doi.org/10.1371/journal.pone.0204129

15. Bohlen, J. (2018). A core philosophy for protecting calf health. Volume 1500, pages 1-6 in UGA Cooperative Extension Bulletin

Modular Mouse Position Surveillance System (ModMoPSS) – a home-cage based test system for laboratory mice

P. Kahnau[1], B. Urmersbach[1], P. Mieske[1,2,3], K. Diederich[1], L. Lewejohann[1,2,3]

1 German Federal Institute for Risk Assessment (BfR), Berlin, Germany; German Center for the Protection of Laboratory Animals (Bf3R),

2 Institute of Animal Welfare, Animal Behavior and Laboratory Animal Science, Freie Universität Berlin, Berlin, Germany

3 Science of Intelligence, Research Cluster of Excellence, Berlin, Germany

In most animal experiments in which mice are used, mice are removed from their familiar environment, their home-cage, and placed in a separate experimental test system. The mice are handled by humans and separated from their social group. Past studies showed that removing animals from their home-cage has a negative impact on animal welfare. Such an influence may have a negative impact on the experimental data and affect the validity of the results. Therefore, it is necessary to develop test procedures to improve the test conditions for laboratory mice and thereby the quality of the data. One promising approach offers the development of home-cage based systems, which allow automated and full-time data collection with as little human influence as possible.

In previous studies, we have already demonstrated the functionality of the Mouse Position Surveillance System (MoPSS) in preference tests with mice [1;2]. In these tests, the mice choose between different goods provided in different cages. To generate individual data from group hosed mice within the MoPSS, it is necessary to implant radio frequency identification (RFID) transponders under the skin in the neck region of the mice. Transponder implantation is performed under anesthesia and analgesia. All experiments were approved by the Berlin state authority, Landesamt für Gesundheit und Soziales, under license No. G 0069/18 and G 0182/17 and were in accordance with the German Animal Protection Law (TierSchG, TierSchVersV).

The main aim was to develop an easy to build and cost effective home-cage based system. An Arduino microcontroller is used in the MoPSS collecting data on a micro-SD card. Two RFID modules connected with two external RFID antennas are used for reading RFID tags. The physical connection between RFID modules and the Arduino is realized by a mainboard, which is built on a perfboard. For mouse tracking, two cages are connected with a tube. The two RFID antennas are installed at both ends around this tube. When a mouse passes through the tube and enters the reading range of an RFID antenna, the transponder number of the mouse, the corresponding RFID antenna number and a time-stamp are logged on the micro-SD card. When analyzing the data, we can extract time and direction of each mouse passing through the tube.

So far, we used the MoPSS as a tracking system in preference tests. However, it is also possible to measure the activity of the mice over an unlimited period of time. This can be performed with a modification of the MoPSS, the MonoMoPSS.

When measuring activity with the MonoMoPSS, a home-cage is divided into two areas by a perforated acryl plate. A tube is inserted through this plate, with an RFID antenna around it. Through this tube the mice are able to enter both areas in the cage. Each time when mice pass through the tube, transponder number and time-stamp are recorded. This activity measurement should make it possible to draw conclusions about the well-being of mice that results from activity. A decrease in activity, for example, could indicate a possible impairment of well-being.

In addition, we are developing a gate system based on our MoPSS by adding doors. The aim is to establish a home-cage based test system in which mice can independently move from the home-cage into a connected test-cage. Only one mouse at a time should be able to pass through the gate and stay within the test-cage in order to perform a test independently and undisturbed by group members.

References

1. Habedank, A., Urmersbach, B., Kahnau, p., Lewejohann, L. (2022). O mouse, where art thou? The Mouse Position Surveillance System (MoPSS)—an RFID-based tracking system. *Behavior Research Methods* 54, 676–689.

2. Hobbiesiefken, U., Urmersbach, B., Jaap, A., Diederich, K., Lewejohann, L. (2023). Rating enrichment items by female group-housed laboratory mice in multiple binary choice tests using an RFID-based tracking system. *PLoS ONE* 18(1): e0278709.

The Neurological Impact of Room Temperature in Built Environment on Thermoregulate in Children with Autism

S. M. Kanakri[1]

1Construction Management & Interior Design Department, Ball State University, Muncie, Indiana, USA. smkanakri@bsu.edu

Introduction

Room temperature and thermal comfort add dimension to the sensory interaction an individual experiences between their neurological perception and the built environment. Especially within the autism community, varying types of room temperatures can be used within the built environment to feed the sensory seeking and/or sensory avoidance needs of the user (Scheffler, Middleton, Abdus-Saboor, 2019). It is estimated that 95% of individuals with autism spectrum disorder experience sensory abnormalities while 60% of those individuals experience thermoregulate sensitivity (Schaffler, Middleton, Abdus-Saboor, 2019). Determining the correlation between thermoregulate behaviors and neural response to the thermal comfort within the built environment, these room temperatures can be implemented in spaces where sensory stimulation or sensory avoidance is needed by the user. This preliminary research has identified research questions and variables for testing the effects of room temperature within the built environment.

The purpose of this research is to identify thermal environmental stressors within the built environment for children within the ASD community and how to implement thermal comfort that contribute appropriate sensory stimulation. The literature reviews discussed the effect that room temperature has on thermoregulate as an indicator of overstimulation in children with autism spectrum disorder. This research has highlighted various methods of measuring thermoregulate in children with autism, by producing a behavioral baseline for this population, sensory processing unique to autistic spectrum conditions.

Hypothesis

If a child with autism spectrum conditions between the age of 5 and 10 is exposed to five thermal comforts situations, then there will be a statistically significant correlation between the neurological sensory process of the four defined behaviors associated with thermoregulate: rocking, hand flapping, scratching, and repetitive verbal attributes, in comparison to the control in children with autism spectrum conditions.

Research Question

What are the effects of thermal comfort present in the built environment on thermoregulate in children with autism spectrum conditions behaviors.

Rationale for This Study

Sensory stress and stimulation are complicated variables to quantify, especially when qualitative data in this situation may be inconsistent. Tools with the capability to ensure the processing of thermoregulate stimuli within the built environment include Observer XT technology. These tools are used to measure behavioral responses to an environment such as focus, stress, stimulation, and information processing (Sinclair, D., Oranje, B., Razak, K. A., Siegel, S. J., Schmid, S. 2017). Understanding the neuroscience behind sensory processing can help determine the best method for presenting thermal comfort for sensory intake.

Exposure to different room temperatures is important to develop a comprehensive view of the impact that thermoregulate stimulation has on the defined characteristics of thermoregulate (Schaffler, M. D., Middleton, L. J., Abdus-Saboor, I. 2019). The method that introduces the least number of external variables such as location, shape, or form of temperature is by presenting a temperature on the floor where participants have consistent exposure. Here, each child participates in a predetermined, neutral and activity on the floor that promotes exposure to the child's hands, feet, and forearms at the child's discretion. Reaction to thermal comfort, such as rocking, hand flapping, scratching (Baranek, G. T., Berkson, G. 1994), and repetitive verbal attributes have been defined as thermoregulate defensive behaviors for the experiment. Frequency of these behaviors will be quantitatively measured for comparison against the control.

Purpose of This Study

1. To determine a correlation between the thermal comfort and frequency of the defined thermoregulate behaviors among children with autism spectrum conditions between the ages of 5 and 10.
2. To quantify how physical attributes of thermal comfort affect the frequency of thermoregulate behaviors among children ages 5-10 with autism spectrum disorder.
3. To determine discrepancies between reporting of thermoregulate defensive behaviors and displayed behavior frequency within a controlled environment.

Methodology

Phase I: Methodology: Phase I: Parental or Guardian Survey

Participants

Participants for Phase I Methodology will be the 25 parent or guardian of the participant. The rationale for this group of participants is to identify existing thermoregulate defensive behaviors. Participants for Phase II will be 25 students diagnosed with autism spectrum conditions from Parish Elementary School located in city of Carmel, Indiana, USA between ages 5 and 10 years.

Environment

The study will take place in Health Environment Design Research Lab at Ball State's University, Muncie, IN, USA within the sensory regulated room. Environmental factors such as lighting, temperature, acoustics, and humidity will stay consistent throughout the experiment and between participants. The only accessible space in the lab will be each of the five temperature overlays covering the area of the room, carpet padding which will remain consistent throughout the experiment, an inflatable stability disc, and a predetermined neutral activity for the child to complete during the experiment.

Materials

Phase I of this study uses the quantitative collection of survey responses from parents given prior to the experiment. The survey will ask the parent or guardian to assign a score form most likely, to least likely, a child is to display each of the defined thermoregulate behaviors when exposed to each of the five thermal comfort. The purpose of this survey is to record the point of view from a primary caregiver based on the child's perceived experience with thermoregulate and displayed behaviors at home. The participant is not meant to have any input in the survey as this is from the primary caregiver's point of view. The thermal comfort present in the survey are the same ones that will be used in the experiment. Blind reactions from the participants is important to receive consistent and reliable data among participants.

Procedure

Prior to the experiment, the parent or guardian of the participant will be given a survey which will allow the parent or guardian to rank the provided description of thermal comfort used in the experiment on a five-point scale from most likely to display a given characteristic of thermoregulate behavior to least likely to display the defined characteristic of thermoregulate behavior. Each of the five thermal comfort in the experiment will be paired with each of the four defined behavioral characteristics associated with thermoregulate and ranked on the same five-point scale.

Phase II: Electroencephalogram Behavioral Observation Purpose

Behavioural data will be collected from neural activity that are directly associated with defined behavioral indicators of thermoregulate, when exposed to five varying degrees of thermoregulate stimulation. This technology provides a greater opportunity to collect data on the sensory processing effect that environmental factors have on children with autism (Sinclair, D., Oranje, B., Razak, K. A., Siegel, S. J., Schmid, S. 2017). The four behaviors that will be measured are: rocking, hand flapping, scratching, and repetitive verbal attributes. Each of these behaviors are associated with self-stimulatory behavior resulting from sensory seeking or sensory avoidance.

Procedure

A task will be assigned to the child when the child joins the lab. Each child will choose any task interested in, five different scenarios will be established. The child will be exposed for each situation for 5 minutes than have break for 10 minutes.

Measurement Techniques

The data collected from the BIOPAC EEG technology will be imported and displayed in Observer XT software. Each trial will produce a set of data points for each observed behavior that will be exported and stored for data analysis. An ANOVA test will be used for statistical analysis of the mean for these behaviors and will be compared to the mean of the control.

Ethical Statement

Strive for honesty in all scientific communications. Honestly report data, results, methods and procedures, and publication status. Fairness: I strive to treat every participant I encounter equally and reflect upon situations before judging others. I plan to apply the same code of ethics to all participants and colleagues in this project.

References

1. Baranek, G. T., & Berkson, G. (1994). Thermoregulate in children with developmental disabilities: Responsiveness and habituation. Journal of Autism and Developmental Disorders, 24(4), 457–471. https://doi.org/10.1007/bf02172128

2. Cascio, C. J., Lorenzi, J., & Baranek, G. T. (2013). Self-reported pleasantness ratings and Examiner-coded defensiveness in response to touch in children with ASD: Effects of stimulus material and bodily location. Journal of Autism and Developmental Disorders, 46(5), 1528–1537. https://doi.org/10.1007/s10803-013-1961-1

3. Cascio, C. J., Moana-Filho, E. J., Guest, S., Nebel, M. B., Weisner, J., Baranek, G. T., & Essick,

4. G. K. (2012). Perceptual and neural response to affective Thermoregulate Room temperature stimulation in adults with autism spectrum disorders. Autism Research, 5(4), 231–244. https://doi.org/10.1002/aur.1224

5. Grover, G., Zhu, S., & Twilley, I. C. (1993). Dynamic mechanical properties of carpet yarns and carpet performance. Temperature Research Journal, 63(5), 257–266.

6. Kenins, P. (1994). Influence of fiber type and moisture on measured fabric-to-skin friction.

7. Temperature Research Journal, 64(12), 722–728.

8. Kinnealey, M., & Fuiek, M. (1999). The relationship between sensory defensiveness, anxiety, depression and perception of pain in adults. Occupational Therapy International, 6(3), 195–206. https://doi.org/10.1002/oti.97

9. Lo, W.-T., Wong, D. P., Yick, K.-L., Ng, S. P., & Yip, J. (2018). The biomechanical effects and perceived comfort of temperature-fabricated insoles during straight line walking. Prosthetics & Orthotics International, 42(2), 153–162. https://doi.org/10.1177/0309364617696084

Development of a goal-directed behaviour protocol using the Barnes Maze

Ewa M. Kopeć, Lianne Robinson, Gernot Riedel

Institute of Medical Science, University of Aberdeen, Aberdeen, UK. e.kopec.23@abdn.ac.uk

Background

Behaviour is considered goal-directed when it is motivated by the possible outcome (e.g., a reward) and its cost-benefit value. Motivation is the drive, the internal determinant, behind the initiation of the action. It can be caused either by environmental stimuli, or a cognitive event. The determinants (innate or learned) can influence the probability of certain actions being taken [1]. A neuronal network hypothesized to be highly implicated in the translation of motivation into action is the limbic-ventral-striatopallidal system. It releases predominantly dopamine (supplemented with GABA-nergic inhibitory feedback). An impairment in dopaminergic functioning would thus also impair goal-directed behaviours and might explain mechanistically the traits of loss of movement initiation, emotional blunting and listlessness – a battery of debilitating negative symptoms present in a wide range of neuropsychiatric conditions ranging from Alzheimer's disease to schizophrenia. The latter emphasises the importance of a pre-clinical translatable behavioural test for goal-directed behaviour [1].

Previous studies have employed different behavioural tasks to assess goal-directed behaviour including effort-based choice tasks, progressive and fixed ratio tasks, random ratio schedules, or T-maze barrier choice procedure, in which a high effort is required to attain a higher valued reward compared to low effort resulting in a lesser reward [4, 3, 5]. Bradfield and colleagues [6] developed a task to assess goal-directed behaviour via a test of outcome devaluation in rats using operant cages. The test involves learning an association between lever presses and different outcomes, before going through a process of devaluation of one of the rewards and subsequent assessment of performance. This paradigm has recently been adapted for use with mice [2] and deficits were reported for the hAPP-J20 mouse model of Alzheimer's disease. This test demonstrated that devaluation influences outcome performance and goal-directed behaviour by assigning differing values to the rewards and thus affecting the motivation to pursue the non-devalued over the devalued outcomes.

Objective

To establish a novel test of goal-directed behaviour in the Barnes maze

Study Design

The current study aims to build upon Dhungana and colleagues' [2] work by adapting and refining the outcome devaluation paradigm to the Barnes maze. We proposed a study outline that involved three distinct phases of i) Goal directed training; ii) Devaluation and iii) Testing (see Figure 1). Animals are initially exposed to training in the goal-directed task using two possible training regimes of either 4 or 8 days in duration. During the acquisition phase the mice are trained to associate the two food rewards with different spatial locations on the Barnes maze. Following the acquisition, the mice then undertake devaluation sessions in the home cage in which one of the rewards is presented and animals are allowed to feed until satiety. The final phase of the task are the test trials performed following devaluation to ascertain the goal-directed choice behaviour of the mice. The test is performed in the Barnes maze in the absence of any reward and time spent in different areas associated with the two rewards recorded. Throughout the different phases of the study, we propose to assess relevant parameters as proxies for the expected behavioural outcomes (see Figure 1 for further details).

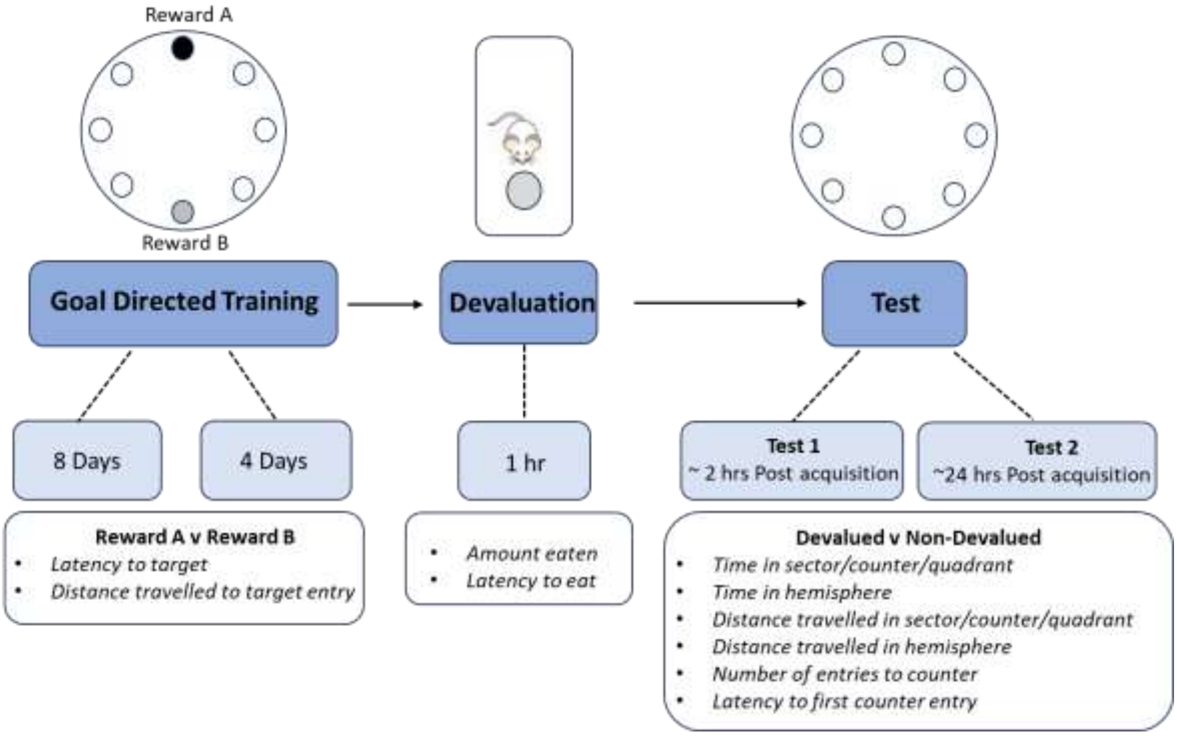


Figure 13. Goal-Directed behavioural testing in the Barnes maze: Study Design. The test consists of three distinct phases: i) Goal Directed training/acquisition; ii) Devaluation and iii) Testing. The goal-directed training is divided into two regimes utilising different durations of 4 days or 8 days. During this phase, the mice are trained to associate two food rewards with a specific target location each. Behavioural parameters measured during training include latency and distance travelled to the target on each trial as an indicator of learning. Devaluation sessions are performed post-acquisition in the home cage, with animals exposed to one of the food rewards and allowed 1 hr of free feeding until satiety with the total amount of food eaten recorded for this stage. Two devaluation sessions are conducted with a different food reward devalued in each. The final phase of the task is the test trials that are performed following devaluation. The first trial is at ~2 hrs post-acquisition and the second 24 hrs post-acquisition. During the tests, no reward or target box are presented and parameters including time spent in different areas of the arena associated with the two food rewards, entries to areas and latency/distance travelled to the first entry are all measured as indicators of performance.

Materials and Methods

Subjects

Subjects were 38 female wild-type NMRI mice aged 5-6 months (Charles River Laboratories; Margate; UK). The mice were group housed in controlled open housing conditions (Makrolon type III cages, corncob bedding, ambient temperature 21±1°C, relative humidity 50-65%, with 17–20 air changes per hour) within the Medical Research Facility at the University of Aberdeen.

Before testing, food (Special Diet Services, Witham, UK) and water were available *ad libitum* with the circadian rhythm maintained on a 12-hour light/dark cycle (lights on at 7 am) with simulated sunrise and sunset (30 min). All behavioural tests were performed during the light cycle. The experiment was performed under the European Communities Council Directive (63/2010/EU) and a project license with local ethical approval under the UK Animals (Scientific Procedures) Act (1986) and its Amended Regulations (2012). The study closely followed the ARRIVE 2.0 guidelines.

Behavioural testing

Animals were allocated into two different cohorts for behavioural testing with N=10 mice used to conduct an initial food preference test to determine whether the animals displayed a similar preference for the different food rewards. The remaining N=28 mice were assigned to the goal-directed task.

i) Food Preference test

Testing was performed over five daily sessions. Before each session, the body weights of the mice were recorded and then they were individually housed in Makrolon Type III cages (Tecniplast, Milan, Italy), lined with paper. The mice were then presented with two petri dishes containing a pre-determined weight of a food reward: i) HobNob biscuits; ii) Digestive biscuits; iii) Milk chocolate; iv) Honey Loop cereal; v) peanut butter chocolate; or vi) laboratory chow. One of the novel food rewards was presented in each testing session alongside laboratory chow. The position (left or right corner) of the rewards versus chow was counterbalanced across animals and testing sessions. The amount of food reward and laboratory chow eaten by each animal was recorded after a delay of 2hrs and 4hrs and was used to indicate a preference for the different food rewards. Throughout all of the testing sessions, the mice were given free access to water.

ii) Goal-Directed Task

Based on the results obtained in the food preference test (see Figure 2), milk chocolate and HobNob biscuits were selected as appropriate food rewards for the goal directed task. Prior to initiating the goal-directed task the second cohort of animals (N=28) were habituated to the two novel food rewards to avoid neophobia. For this purpose animals were individually housed in PhenoTyper cages (Noldus, Wageningen, NL) made of clear Perspex (30 x 30 x 35 cm) and containing a water bottle offering *ad libitum* access to water. Jars containing pre-determined quantities of HobNob biscuit and chocolate were placed in two corners of the cage with the positioning fully counterbalanced for animals. Activity of the mice was recorded by an infra-red camera built-in to the lid of each cage and analysis was recorded offline using the behavioural analysis software Ethovision XT14 (Noldus, NL). Measurements of the food intake were determined after four hours and animals were returned to their home cages at this time. Further habituation sessions to each of the food rewards were performed overnight in the home cages of the mice in the week prior to testing in the Barnes maze.

Barnes Maze

Animals were randomly assigned to one of the testing regimes: i) Group 1 – 8 days of acquisition or ii) Group 2 – 4 days of acquisition (N=14 per group). The goal-directed task was assessed using a 16-hole Barnes maze which consisted of a black open circular arena (120 cm diameter), with 16 holes (8 cm diameter) evenly spaced around the perimeter. The arena was elevated 1 m above the floor and surrounded by various distal spatial cues with both indirect and direct illumination. An escape box was located under one of the holes and offered the only means of escape. An overhead camera recorded the activity of the animal and the behavioural tracking software ANY-maze v6.03 (Ugo Basilie, Italy) was used for offline analysis.

Habituation

Animals were given one day of habituation to the arena in which curtains were drawn around the arena to obscure spatial cues and a target escape box was positioned below a random hole. Animals were released from the centre of the arena and the time taken to locate the target box was recorded, with a maximum trial time of 5 minutes.

Goal-directed training

Following habituation the mice performed goal-directed training in which the escape box was positioned either at the north or south hole. A food reward of biscuit or chocolate was placed in the box, with the location of the reward counterbalanced for all animals (ie. chocolate – north and biscuit – south or vice versa). The animal was released from the centre of the arena and the time taken to locate the box was recorded. A maximum trial time of 120s was allowed. If an animal failed to locate the escape box within the allotted time, she was guided to it by the experimenter. During acquisition, animals were given 4 trials per day (2 trials for each reward/location) for either 4 or 8 days with an ITI of ~ 10 minutes.

Devaluation

Devaluation was conducted in individual Makrolon Type III cages with free access to water. Animals were exposed to one of the food rewards for 1 hour and allowed to feed freely to satiety, with the devalued reward fully counterbalanced across animals. Each animal was exposed to two devaluation sessions: 1) immediately following completion of training and 2) ~24 hrs post-training with a different reward being presented during each session. The amount of reward consumed by an individual animal was measured.

Test

Test trials were performed ~30 minutes following the completion of the devaluation sessions. The escape box was removed, and animals were allowed to freely explore the arena for 120s, with the time spent in the areas previously associated with the rewards taken as a measure of goal-directed choice behaviour.

Throughout the goal-directed task mice were food restricted (1 – 2 g chow per day) and maintained at 80% of their body weight with access to water *ad libitum*.

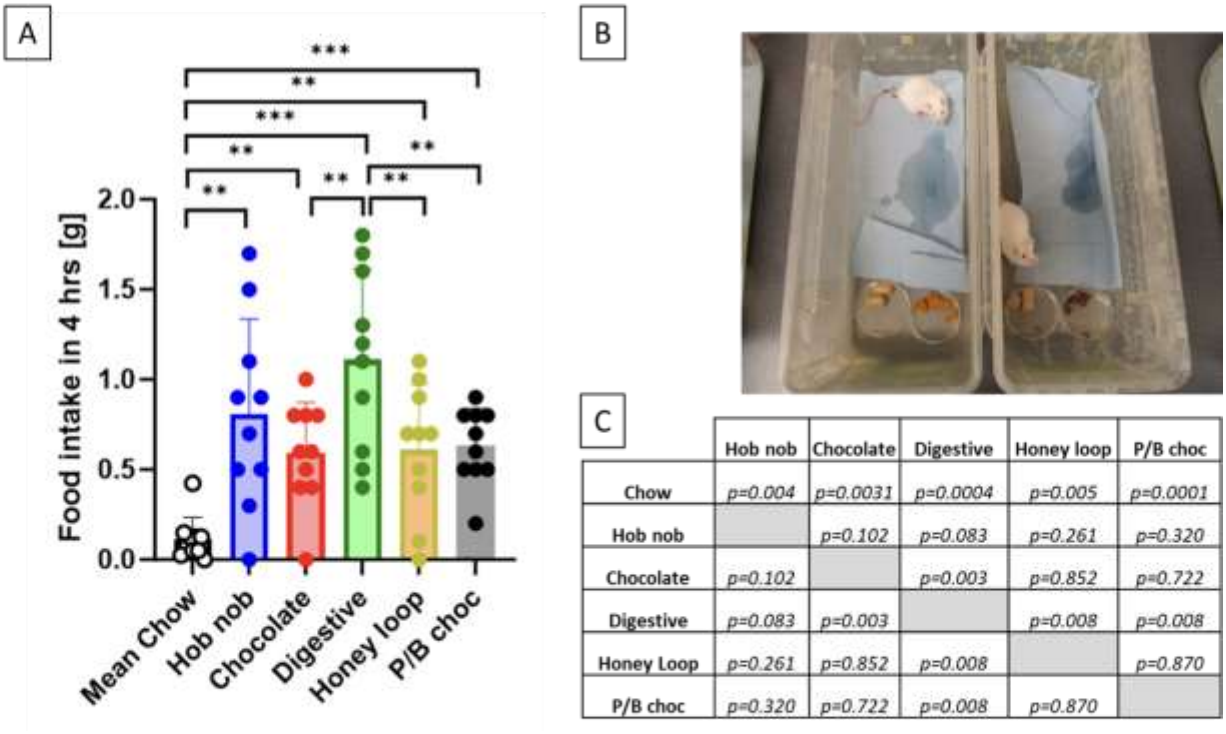
Data analysis

Various parameters were measured during each of the phases of the task (see Figure 1) and used as proxies of goal-directed behavioural outcomes. Statistical analysis using paired t-tests and other appropriate analyses were performed in GraphPad Prism v10 (GraphPad Software Inc., USA) with an assumed confidence level of 95%.

Results

Results from the food preference task observed a significant preference for all food rewards compared to laboratory chow (all p's ≤0.005) (see Figure 2A & 2C). Furthermore, a significant preference for digestive biscuits compared to all other food rewards except for HobNob biscuit was revealed (all p's≤0.003). No further differences between the other food rewards were evident, suggesting that mice display a similar preference for the rewards.

There is no data available for the goal-directed phases of the study with analysis currently ongoing. As the development of the task is still in progress and the effectiveness/outcome is still to be determined, further refinements of the task and analysis will be implemented accordingly.



	Hob nob	Chocolate	Digestive	Honey loop	P/B choc
Chow	*p=0.004*	*p=0.0031*	*p=0.0004*	*p=0.005*	*p=0.0001*
Hob nob		*p=0.102*	*p=0.083*	*p=0.261*	*p=0.320*
Chocolate	*p=0.102*		*p=0.003*	*p=0.852*	*p=0.722*
Digestive	*p=0.083*	*p=0.003*		*p=0.008*	*p=0.008*
Honey Loop	*p=0.261*	*p=0.852*	*p=0.008*		*p=0.870*
P/B choc	*p=0.320*	*p=0.722*	*p=0.008*	*p=0.870*	

Figure 14. Food intake during the food preference test. A) Mean food consumption of the different food rewards compared to laboratory chow. A significant increase in intake of all food rewards was evident relative to laboratory chow; a higher amount of digestive biscuits was eaten compared to all food rewards except for HobNob biscuit. No further differences were observed between the other food rewards. B) An example of the cage set-up for the food preference test. Chow and food rewards were positioned either on the left or right side of the cage with food reward positioning being fully counterbalanced across animals. C) Table displaying the statistical results from the paired t-tests for comparison of the different rewards. Mean+SD.

Conclusions

Based on the results of the pilot study we were successfully able to determine food rewards that could be used in the goal-directed choice task.

The current test using the Barnes maze offers a modification to the test introduced by Dhungana *et al.* [2]. An in-depth analysis of the goal-directed task is ongoing utilising appropriate analyses to determine how different factors including both type of reward and spatial location of the reward can influence the performance. These analyses offer important data towards increasing our understanding of goal-directed behaviour in the mice and will guide further refinements of the task.

References

1. Brown, R.G. and Pluck, G. (2000) 'Negative symptoms: the "pathology" of motivation and goal-directed behaviour', *Trends in Neurosciences*, 23(9), 412–417.
2. Dhungana, A. *et al.* (2023) 'Goal-Directed Action Is Initially Impaired in a hAPP-J20 Mouse Model of Alzheimer's Disease', *eNeuro*, 10(2).
3. Nunes, E.J. *et al.* (2014) 'Effort-related motivational effects of the pro-inflammatory cytokine interleukin 1-beta: Studies with the concurrent fixed ratio 5/ chow feeding choice task', *Psychopharmacology*, 231(4), 727–736.
4. Randall, P.A. *et al.* (2012) 'Dopaminergic Modulation of Effort-Related Choice Behavior as Assessed by a Progressive Ratio Chow Feeding Choice Task: Pharmacological Studies and the Role of Individual Differences', *PLoS ONE*, 7(10).
5. Yohn, S.E. *et al.* (2018) 'The monoamine-oxidase B inhibitor deprenyl increases selection of high-effort activity in rats tested on a progressive ratio/chow feeding choice procedure: Implications for treating motivational dysfunctions', *Behavioural Brain Research*, 342, 27–34.
6. Bradfield, L.A. *et al.* (2020) 'Goal-directed actions transiently depend on dorsal hippocampus', *Nat Neurosci 23*, 1194–1197.

Assessing motivation and social communication in rodent: a challenge

C. Laloux

US41-UAR2014 Biology and health platforms of Lille, Lille In vivo imaging and functional exploration platform, University of Lille, CNRS, Inserm, Lille hospital center, Pasteur institute of Lille, Lille, France.
charlotte.laloux@univ-lille.fr

Introduction

Cerebral pathologies are often associated to motivation and/or social interaction disorders. Indeed, beyond the characteristic symptomatology, patients suffering from neurological, psychiatric or cerebrovascular diseases frequently complain of lack of desire or social isolation. To better understand the origin and pathophysiology of these associated symptoms, and to find new management strategies, animal models of these diseases are often used. At the same time, the development of adapted and validated experimental paradigms to assess the components of these disorders remains a challenge in the field. Indeed, assessing motivation and social communication in rodents is far from easy.

Ultrasonic vocalization (USV) emission is an important component of social behaviour repertoire as a communicative function to transmit information to peer [1]. However, in social interaction context, the study of USV emitted by one individual is also depending of the context and the recipient individual affective or emotional state. The question here is to discuss about the protocol used to study the USVs communication in social interaction condition.

Another challenge in the field of psychiatric state assessment in rodent models is the motivation. Motivation is a complex psychological process defined as triggering behaviour to achieve a desired goal [2]. Most of the motivation assessment protocol in rodent is based on food seeking behaviour after food restriction of the animal, often in operant conditioning paradigm. Indeed, goal-directed behaviour involves a cost/benefit computation in which the anticipated benefit of the to-be-earned reward is weighed against the anticipated effort of the task at hand [3]. Based on that principle, fixed/progressive ratio and effort related choices paradigm allow the assessment of goal directed (reward motivated) behaviour in rodent. However, the food restriction degree is determinant and should be considered.

We propose here two distinct protocols that we wanted to test on the Lille functional exploration platform to assess motivation and social communication in rats, and raise the question of their relevance.

Methods

Forty adult males Wistar rats were used in 2 different cohorts (French ethical comity CEEA075, authorization number #33933).

In the first experiment, ultrasonic vocalizations were recorded in a 3 chambers social interaction paradigm on 20 rats. Two groups of rats with different social housing conditions were compared: single versus grouped housing conditions. After habituation of the tested rat to the empty arena, an intruder rat is introduced in a small bar enclosure placed in one of the 3 chambers. The microphone is located in the “meeting” chamber and record the USV emitted both animals inseparably. To avoid effect of the intruder rat state on the USV communication elicited, both the intruder and the tested rats are from the same housing condition group.

In the second experiment, operant conditioning in touchscreen chambers (Campden instrument) was used to apply Fixed/Progressive ratio (FR/PR4) protocols and Effort related choice (ERC) protocol on 20 rats. Two groups of rats with 2 different degree of food restriction were compared: one with 10% and one with 20% of food restriction as regard to baseline body weight. In FR, the animal is rewarded with sucrose palatable pellet after it makes a specified number of nose poke on the enlighten square of the screen (FR1, FR2, FR3, FR5). In PR+4 task, rats are rewarded after a certain number of nosepoke emitted, but the work requirement for each reward increases progressively with a step of 4 following previous reward delivery. The effort becomes more and more difficult, until at some point the animal refuses to complete the work requirement. The work requirement at which the animal no longer responds is taken as an index of motivation (Breakpoint). In ERC task, the animal can choose between working (nosepoke on the screen) to obtain the palatable sucrose reward via FR schedule (FR16, 32 and 40) or consuming the standard food freely available in the chamber. The only difference between FR and ERC is that during ERC, standard food is freely available in the chamber to offer the rat a choice (ERC16, 32 and 40).

Results

In social interaction test, grouped-housed rats spent less time in interaction with the intruder rat than single housed rats, confirming the importance of social housing conditions. Moreover, single-housed rats elicited significantly more USV than grouped housed rats confirming that social isolation have an impact on social communication.

In touchscreen operant conditioning for motivation assessment, from the start of training in FR1-5, the 10%-restricted rats performed much less well in the FR task in comparison to 20%-restricted rats, whatever the level of effort required. In the progressive ratio task (PR+4), the 20%-restricted animals quickly reached their breakpoint plateau and performed more trials, and so more nosepoke to be own rewarded, than their 10%-restricted counterparts. In ERC part of the task, performed only with 20%-restricted group, rats perform fewer trials in ERC than in FR at all levels of difficulty. This means that rats work well in FR, but as soon as they have a choice of free-access standard food, they work much less. The more the difficulty of obtaining palatable reward increased, the more they chose the easier option.

Discussion

These results were expected and confirm the effect of isolation on the potentiation of social interaction/communication and the effect of the degree of dietary restriction on tasks assessing food seeking motivation. Technically, the question is raised about the paradigm to assess social USV emission. USV emitted by the tested rat are dependant from the intruder rat, and it was not possible in the present paradigm to discriminate them. Here the choice was made to record USV emission of both individuals from the same group, assessing thus communication ability of the group rather than the one of one individual. Is it possible to record individual USV emission in social context without recording at the same time the USV of the intruder? in the other side, is it relevant to study vocalization of one individual without taking account the recipient ones?

Fixed/Progressive ratio (FR/PR4) and Effort related choice (ERC) protocols are classically used to assess motivation in rodent. Those protocols are based on food seeking behaviour after food restriction of the animal. Moreover, the food restriction degree and bodyweight gain curve are very difficult to control. How relevant is this paradigm to assess spontaneous motivation of the animal? Is there another way to assess motivation in research rodent models? Can we compare this motivation evaluation to the human counterpart?

The ERC is described to specifically assay the trade-off between the expenditure of effort to obtain a preferred reward and the consumption of a freely available standard reward. Here again the food restriction degree and the affective state of the animal over the task is highly impacting. In this task, animals prefer generally to work for the preferred reward and consume less of the freely available food. Does this task really assess the animal's motivation? What about the emotional or the fatigue state of the animal? I'd be happy to discuss about these question with you.

References

1. Wöhr, M., Schwarting, R.K. (2013). Affective communication in rodents: ultrasonic vocalizations as a tool for research on emotion and motivation. *Cell Tissue Res*earch **354(1)**, 81-97.
2. Ward, R.D. (2016). Methods for Dissecting Motivation and Related Psychological Processes in Rodents. *Curr Top Behav Neurosci*ence **27**, 451-70

3. Salamone, J.D., Correa, M., Farrar, A., Mingote, S.M. (2007) Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology* **191(3)**, 461-82.

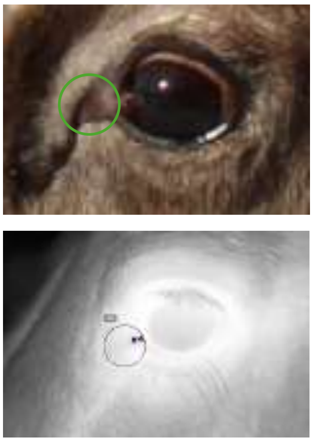
Assessment of emotional changes and pain in horses using infrared thermal measurement of the eye

Océane Liehrmann*[1], Virpi Lummaa[1], Veera Riihonen[1], Léa Lansade[2]

1 - Department of Biology, University of Turku, 20500 Turku, Finland; 2- CNRS, IFCE, INRAE, Université de Tours, PRC, 37380 Nouzilly, France.

Introduction

Measuring immediate levels of stress in animals can be challenging. Traditional methods like behavioural observations can be difficult to interpret without training, while saliva and blood sampling necessitate animal training and handling. Technological advances have led to new tools, such as infrared thermal cameras which can be used to efficiently measure external variation of the body temperature without contact. Studies conducted on humans [1] and other primates [2] have shown that emotional stimuli often lead to a drop in nose temperature, making it a potential indicator of arousal in animals. However, in mammals for which the nasal area is covered with fur, assessing temperature through infrared thermal imaging is not possible. In such cases, an alternative is to measure the temperature of the medial canthus, which is a hairless area in the eye corner. With this method a continuous drop of the medial canthus temperature during a stressor was observed in cows [3]. I also made similar observations myself with reindeer [Liehrmann et al. in prep]. Most studies using infrared thermal imaging to assess body temperature variation during emotional changes only focus on negative emotions promoted by fear. With this study we aim to fill a crucial gap by investigating the variation in horses' medial canthus temperature throughout events involving emotions of negative and positive valence, such as exposure to fearful stimuli (negative excitement), grooming sessions (positive and calming emotion), and food distribution (positive excitement). This study is the first to explore how the medial canthus temperature varies across situations promoting responses of different intensity and valence. Therefore, we attempted to assess the reliability of the medial canthus temperature as an indicator of the horse's emotional state. In addition to emotional state, medial canthus temperature could serve as an indicator of pain. Horses can express pain through changes in facial expressions, including muscle contractions in the eye area [4] but this is very difficult to assess and necessitate specific training. These muscle contractions around the eye could result in an increase of the medial canthus temperature. Therefore, this measure could serve as an indicator of pain which would be clearer and easier to use for horse owners and vets. In a second experiment we recorded the variations in eye temperature in horses when the horses presented facial expressions of pain to assess the reliability of the medial canthus temperature as an indicator of pain.



Horse medial canthus region on a photo and on a thermal imaging picture

The data was collected in May 2023. Two experiments were carried out on Welsh pony mares (female horses) bred and raised at the experimental unit of animal physiology in France (UE PAO, 37,380 Nouzilly, France, INRAE: https://doi.org/10.15454/1.5573896321728955E12). The horses involved in this study are solely used for research purposes and the experimental procedures comply with the necessary ethical permits. The experiments were conducted in France; therefore, our experimental design was reviewed by the French authorities for animal ethics. The Ethics Committee on Animal Experiments of the Val de Loire (CEEA Vdl) agreed to review the protocol and attributed positive recommendations for the conduction of the experiments. Registration Number: CE19 – 2022-0510 – 1.

Understanding the eye temperature variation during emotional changes

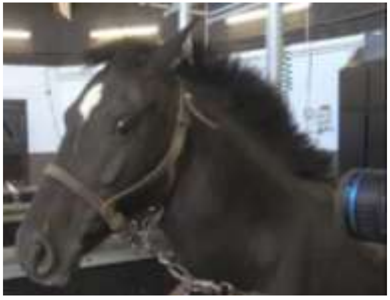
A total of 26 mares were subjected to three behavioral tests. All horses were equipped with a remotely controlled heart rate monitor made for equine use (Polar Equine RS800CX Science, Polar Oy, Finland) to obtain continuous heart rate measurements. The experiments took place in a large experimental area (3.5m x 4.5m) familiar to the horses. The focal horse was held on a lead rope by a research assistant standing on its right side. A second experimenter stood within 1m to the left side of the horse's head and video recorded the individual's eye with an infrared thermal camera (FLIR T540: FLIR Systems Inc., Wilsonville, OR, USA- cf picture) for the duration of the test. To add support to the thermal imaging data all experiments were video recorded with a 4K camera for further behavioural analyses (body language, movements, and facial expressions). Each test consisted of three phases: The control phase, lasting 1 minute under the initial test start conditions; the test phase, lasting 1 minute of stimulus; and the recovery phase, lasting 2 minutes to return to the initial conditions.

- Stimulus with negative value and high arousal (fearful stimulus): an experimenter approached the horse and shook a plastic bag for 1 minute. The goal was to maintain the horse in an alert position (head high, ears straight, tense body and avoidance behaviour), reflecting a negative emotional state, but not to have the horse panicking. Therefore, the experimenter was free to adjust the intensity of the bag shaking according to the horse behavioural reaction.
- Stimulus with positive value and low arousal (allo-grooming simulation): The experimenter holding the horse scratched it on the horse’s withers to mimic a mutual grooming action for 1 minute. We used the horse behaviour during the stimulus (leaps extension, neck extension, contact seeking) and after the stimulus (contact seeking with experimenter) to assess that the stimulus was perceived as a positive event.
- Stimulus with positive value and high arousal (appetence): An experimenter entered the experimental area with a bucket containing pellets and placed themself 1m in front of the horse so it could not reach the bucket. For 1 minute the experimenter shook the bucket to promote excitement (ears straight, leap extension, attempt to reach the bucket) and gave a small handful of pellets to the horse every 5 seconds (to prevent frustration).

To prevent any experimental bias each mare went through the 3 stimuli in a randomized order on different days.

Is the medial canthus temperature an effective indicator of pain in horses?

The mares of the research station are used for research on reproduction and undergo regular gynecological examination. These exams involve transrectal echography, which is a common practice in the horse breeding industry for monitoring the mare's menstrual cycle and pregnancy. Breeders often perform these exams themselves, and a veterinarian license is not required. During these exams, the examiner inserts their entire arm into the horse's rectum, which is a highly invasive procedure. Mares may exhibit discomfort or pain, which can be observed through the contraction of their facial muscles, as well as their overall body language. I recorded 33 mares with the thermal camera for 1 minute before the start of the echography, throughout the whole echography and 2 minutes after the echography. The heart rate and facial expressions of these mares were also recorded during the whole procedure.



Facial expression of pain during transrectal echography

The objective is to investigate whether there is an association between the medial canthus temperature and the intensity of facial expression when horses experience pain.

Analysis

To ensure rigor and impartiality, collaborative video coding by a minimum of two individuals is imperative to safeguard against potential biases and subjective judgments. Therefore, two persons will code the horse’s behaviour from the 4K videos using the software program BORIS. Both coders will train to obtain the Horse

Facial Action Coding System [5] certification (EquiFACS), a scientific observational tool for identifying and coding facial movements in horses. Both coders will use the same ethograms and will rate the occurrences and durations of specific behaviour. The repeatability of the behavioural coding between the two raters will be assessed before using the data for further analysis. After each test, the recorded heart rate variations were extracted in text form from the Polar Equine app. and sent to the computer to be used in statistical analyses. Finally, using the FLIR thermal studio software, pictures from the infrared videos will be extracted every 2 to 5 seconds (50 temperature points/test) to record the maximum temperature from the medial canthus area. The final statistical analyses will assess potential correlation between behavioural responses, heart rate response and medial canthus temperature variation.

The poster aims to present the methods and preliminary results of this study.

References

1. Ioannou, S., S. Ebisch, T. Aureli, D. Bafunno, H. A. Ioannides, D. Cardone, B. Manini, G. L. Romani, V. Gallese, and A. Merla. 2013. “The Autonomic Signature of Guilt in Children: A Thermal Infrared Imaging Study.” *PLOS ONE* **8** (11): e79440.

2. Kano, F.o, S. Hirata, T. Deschner, V. Behringer, and J. Call. (2016). “Nasal Temperature Drop in Response to a Playback of Conspecific Fights in Chimpanzees: A Thermo-Imaging Study.” *Physiology & Behavior* **155**: 83–94.

3.Stewart, M., K.J. Stafford, S.K. Dowling, A.L. Schaefer, and J.R. Webster. (2008). “Eye Temperature and Heart Rate Variability of Calves Disbudded with or without Local Anaesthetic.” *Physiology & Behavior* **93** (4–5): 789–97.

4. Dalla Costa, E., et al. (2014) "Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration." *PLoS one* **9.3**: e92281.

5. Wathan, J., Burrows, A. M., Waller, B. M., & McComb, K. (2015). EquiFACS: the equine facial action coding system. *PLoS one*, 10(8), e0131738.

Using PLS-DA to discriminate facial expressions coded from EquiFACS in the study of affective experiences

J. Lundblad[1] and P. Haubro Andersen[1]

1 Department of Animal Biosciences, Swedish University of Agricultural Sciences, Uppsala, Sweden. Johan.Lundblad@slu.se

Introduction

In 1862, Duchenne explored the impact of emotions on facial muscles, revealing two neurological pathways for intentional movements and emotional expressions [1]. The emotional component have been proposed for investigating welfare signals in animals [2], crucial for prey species like horses where discomfort indicators may be concealed in the precense of humans [3]. So far, facial expressions have been evaluated during pain [4–6], stress [7] and sedation [8] separately, thus mainly been aimed towards a compartmentalized approach. However, current insight into emotions adapt a more continuous approach [9], where affective states are not strictly associated to a base emotion. Furthermore, facial expressions of negative valence may appear in neutral states [6], suggesting that some expressions lack discriminatory power. Facial Action Coding Systems (FACS), developed for short time frames with known context, face challenges in animal studies where the observation times often are longer and thus, introduce noise in the dataset. This study explored Partial Least Square Discriminant Analysis (PLS-DA) as a method to discriminate between facial expressions during different states in horses, acknowledging the need for models adept at handling these wide dataset, using FACS for horses (EquiFACS [10]).

Material and methods

Ethical permission for the use of horses in this study was obtained from the Ethics Committee for Animal Experiments in Uppsala in accordance with EU regulations and Swedish law. Twelve horses, all Standardbred trotters, participated in the study. Facial expressions during three experimental interventions: mild temporary ischemic nociception (N), isolation stress (I) and pharmacologically induced sedation (S) were studied. The interventions were arranged in a semi-randomized cross-over design. The horses were video-filmed during 15 minutes for each intervention without an observer present. All facial expressions during the interventions were recorded on film using four wall-mounted surveillance cameras, mounted in each corner of the box to ensure video of the horse from four angles simultaneously. Facial expressions were recorded from the films using EquiFACS by two certified coders which were blinded for which intervention the horse underwent. The observations were generated using the open software ELAN[11], where onset, offset, and duration of each code in EquiFACS were annotated on the video timeline, which gave frequency for each of the facial expressions. Partial least squares-discriminant analysis was performed in SIMCA® (version 17) using standard settings. Interventions were set as class variable, frequencies of all codes were set as variables, and individual horses were set as secondary class variable.

Results

A total of 2554 code events (onset to offset) were coded. Using all EquiFACS-codes in PLS-DA, two components (t1, t2) consisting of weighted components of the EquiFACS codes were created (Figure 1). A total of 25 % of variation in the dataset was explained by the two first components, and both components were significant (Q2>0.05).

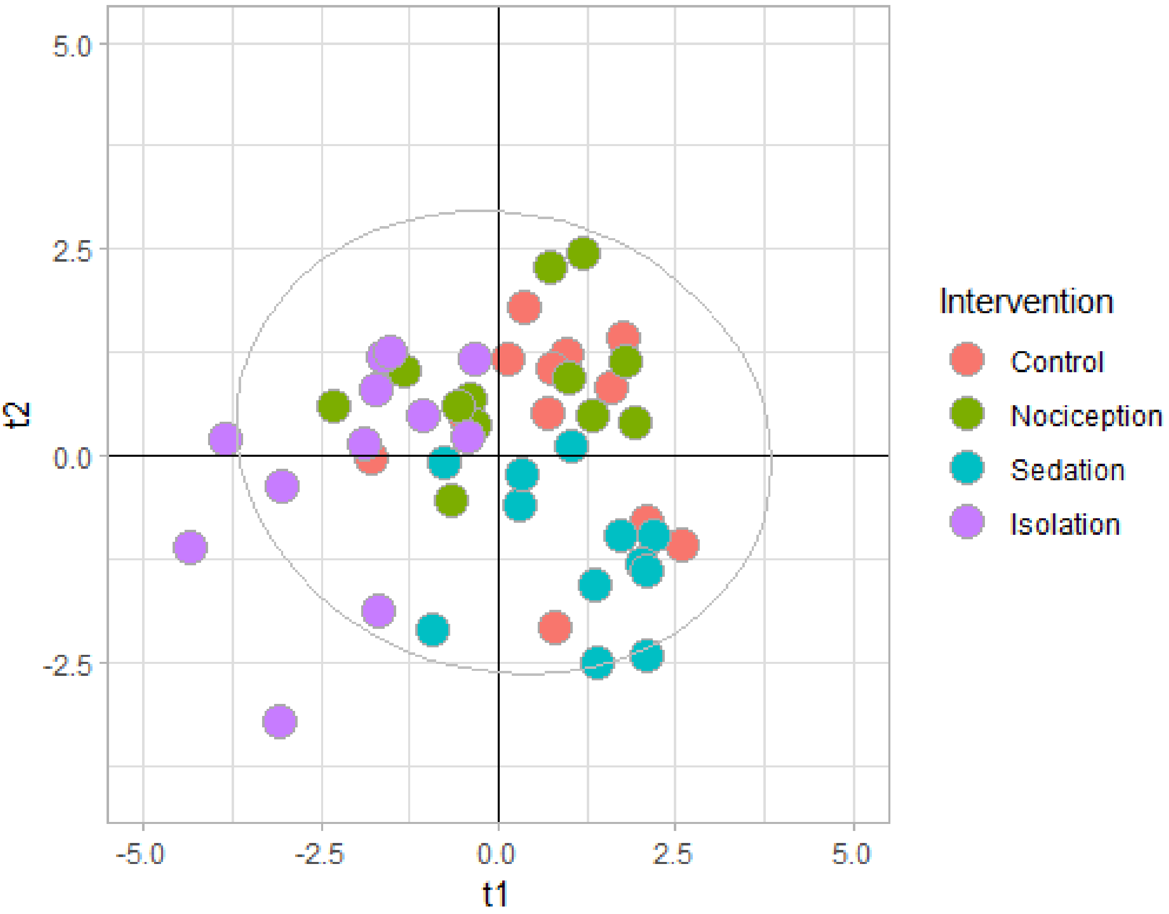


Figure 1. Scores plot of the outcome of the PLS-DA analysis. Each point corresponds to an individual horse (n=12), colored according to intervention plotted as a scatterplot over orthogonal new variables (components t1 and t2) consisting of weighted summaries of the frequencies of EquiFACS-codes designed to explain as much variation as possible while at the same time separate the interventions as much as possible. Ellipse indicates 95% confidence interval.

Furthermore, the estimated importance of each code in the discriminatory model was presented using standardized VIP values, which is shown in Figure 2. Nine action units had VIP values above 1, and could thus be deemed important contributors to the model.

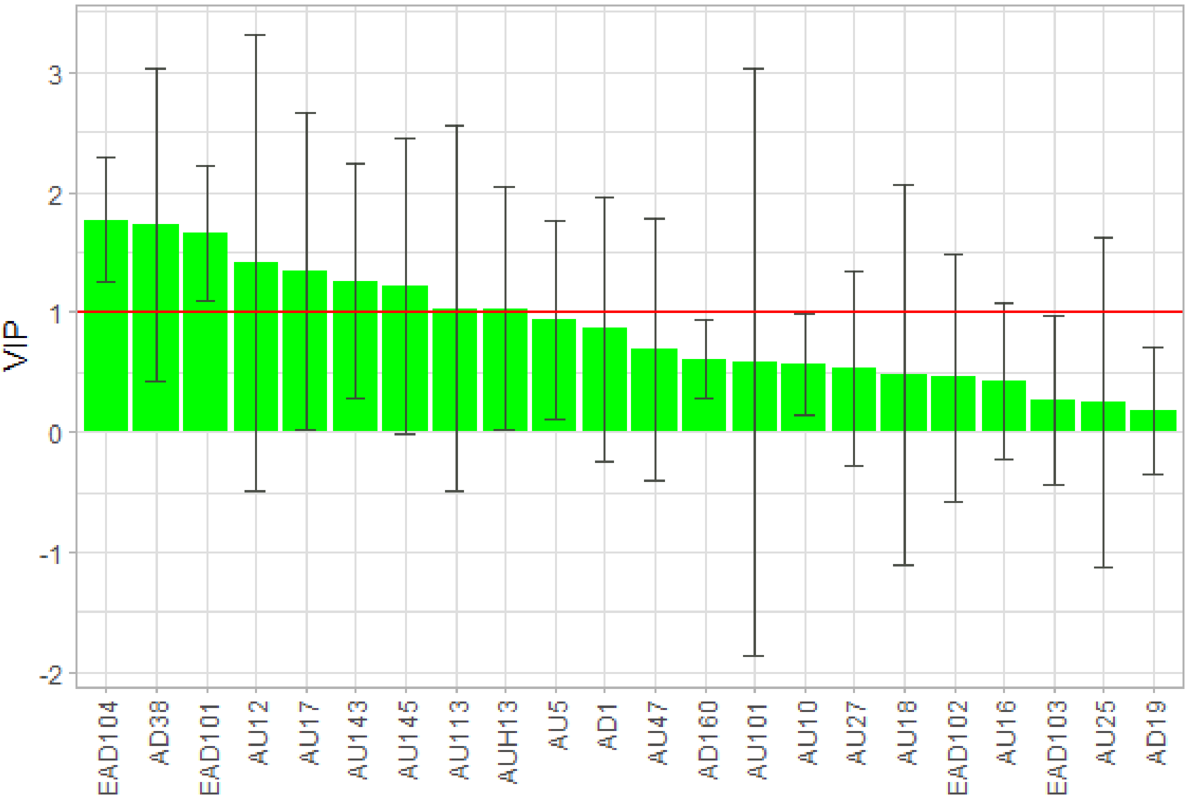


Figure 2. Variable importance in projection (VIP) plot. VIP-values calculated adding the sums of squares of the PLS loading weights adjusted with the number of explained sums of squares of each component (t1, t2), representing the contribution of each EquiFACS code to the two PLS-DA components. The average VIP is standardized to 1. Error bars indicate a 95 % confidence interval of VIP values.

Discussion and Conclusion

This study proved that using PLS-DA on frequency of EquiFACS-codes could successfully discriminate between different experiences. The distinction of facial muscle activity observed for the interventions provide valuable insights into facial codes that offer information when evaluating intricate states in horses, enhancing the biological significance of interpreting facial expressions in animals. However, while the frequencies of facial expressions pose differences between these experiences in horses, careful consideration needs to be given when evaluating facial expressions in horses during challenging situations, especially in complex situations where the context is unknown or compound.

References

1. Augustine GJ, Groh JM, Huettel SA, LaMantia A-S, White LE. Neuroscience. Sevent edition. New York, NY, US: Oxford University Press; 2024.
2. Descovich KA, Wathan J, Leach MC, Buchanan-Smith HM, Flecknell P, Farningham D, et al. Facial expression: An under-utilized tool for the assessment of welfare in mammals. Altex. 2017;34: 409–429.
3. Torcivia C, Mcdonnell S. In-Person Caretaker Visits Disrupt Ongoing Discomfort Behavior in Hospitalized Equine Orthopedic Surgical Patients. Animals. 2020;10.
4. van Loon JPAM, Van Dierendonck MC. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A scale-construction study. Vet J. 2015.

5. Dalla Costa E, Minero M, Lebelt D, Stucke D, Canali E, Leach MC. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS ONE. 2014;9: 1–10.
6. Rashid M, Silventoinen A, Gleerup KB, Andersen PH. Equine Facial Action Coding System for determination of pain-related facial responses in videos of horses. PLOS ONE. 2020;15: e0231608.
7. Lundblad J, Rashid M, Rhodin M, Andersen PH. Effect of transportation and social isolation on facial expressions of healthy horses. PLOS ONE. 2021;16: e0241532. doi:10.1371/journal.pone.0241532
8. Oliveira AR de, Gozalo-Marcilla M, Ringer SK, Schauvliege S, Fonseca MW, Trindade PHE, et al. Development and validation of the facial scale (FaceSed) to evaluate sedation in horses. PLOS ONE. 2021;16: e0251909.
9. Mendl M, Burman OHP, Paul ES. An integrative and functional framework for the study of animal emotion and mood. Proc R Soc B Biol Sci. 2010;277: 2895–2904. doi:10.1098/rspb.2010.0303
10. Wathan J, Burrows AM, Waller BM, McComb K. EquiFACS: The equine facial action coding system. PLoS ONE. 2015;10: 1–35.
11. Lausberg H, Sloetjes H. Coding gestural behavior with the NEUROGES-ELAN system. Behav Res Methods. 2009;41: 841–849.

Effects of acute sleep deprivation: A multimodal machine learning approach

S. Özsezen[1], L. Verschuren[1] and A.M. Brouwer[2]

1 Healthy Living and Work, TNO (The Netherlands Organization for Applied Scientific Research), Leiden, the Netherlands; Sylviusweg 71, serdar.ozsezen@tno.nl, lars.verschuren@tno.nl

2 Defence, Safety and Security, TNO (The Netherlands Organization for Applied Scientific Research), Soesterberg, the Netherlands; Kampweg 55, anne-marie.brouwer@tno.nl

Background

Our understanding of why sleep deprivation (SD) affects cognitive performance, and why it does so to such different extents in different people, is limited. Given that SD is unavoidable in certain professions and under certain circumstances, a better grasp on the mechanisms underlying cognitive decline through SD, and better prediction of cognitive decline, would be valuable to intervene and cope with negative effects of SD. Previous literature suggests explanations at very different levels and through various mechanisms, including psychological, neural, inflammatory, lipidomic and cellular mechanisms. Clearly, the indicated mechanisms are not mutually exclusive, and probably interdependent. The variety of suggested explanations also hints at the possibility to predict cognitive effects of SD through a combination of measures from multiple domains, from inflammatory cytokines and lipidomics to physiology reflecting arousal. Such a multimodal, comprehensive approach may also help elucidate the mechanisms underlying cognitive decline through SD. We here describe such an approach, using feature selection and machine learning techniques in an attempt to meaningfully handle and interpret a large quantity of variables.

Methods

The data reported here originate from a study that is described in greater detail by Bottenheft et al. [1] and Stuldreher et al. [2]. Of 101 participants, 53 were randomly selected for one night of sleep deprivation, and 48 slept as usual. The morning before and after the (sleep deprived) night, heart rate and electrodermal activity were recorded during 3-minute resting periods. Capillary blood was collected by a finger prick and saliva was collected in saliva collection containers. Inflammatory and lipidomic biochemical markers were assessed in plasma and saliva; cortisol was assessed in saliva. After collection of plasma and saliva, participants performed a battery of standardized cognitive tests to assess cognitive performance. Physiological and biochemical features were extracted. All features were baselined by subtracting the measurement from the second day from the measurement of first day, obtaining differences of the relevant biomarker per person. We then took a two-step approach in our analysis. Firstly, we examined how SD affects physiological and biochemical features in general. We trained three logistic regression models with elastic net penalty on subsets of data (focusing on plasma biochemical data – 749 features; focusing on saliva – 130 features; and on physiology – 40 features). The models were evaluated as to the accuracy of their prediction in an unseen sample of participants whether data comes from an individual who slept or not. Features that are relevant for distinguishing between SD or control may also be related to cognitive decline due to SD. In the second step, we use those features to try and predict change in cognitive performance over night in the SD group, again using a logistic regression model, now predicting change in cognitive performance as a continuous measure.

Results

As expected, sleep deprivation substantially decreased cognitive performance, and there was considerable variation of performance decline between sleep-deprived participants. The plasma and saliva model could distinguish between sleep deprived and control participants (90% correct for completely unseen data); whereas the physiological model could predict with less accuracy (72% on unseen data). The features that contributed significantly to the models' performance (plasma model: 114 features; physiological model: 2 features; saliva model: 6 features) were used as input in the logistic regression to predict change in cognitive performance. This did not work above chance.

Discussion

Our study showed profound effects of one night sleep deprivation on cognitive performance, lipids and inflammatory markers. Unfortunately, the attempt to relate the change in biochemical markers directly to change in cognitive performance due to sleep deprivation was not successful. As a next important but challenging step, we propose more stringent feature selection by using knowledge of the likely and unlikely relations between the biochemical markers that mark sleep deprivation and cognitive performance.

References

1. Bottenheft, C., Hogenelst, K., Stuldreher, I., Kleemann, R., Groen, E., van Erp, J., & Brouwer, A. M. (2023). Understanding the combined effects of sleep deprivation and acute social stress on cognitive performance using a comprehensive approach. *Brain, Behavior, & Immunity-Health*, 100706.2.

2. Stuldreher IV, Maasland E, Bottenheft C, van Erp JBF & Brouwer A-M (2023). Physiological synchrony in electrodermal activity predicts decreased vigilant attention induced by sleep deprivation. *Front. Neuroergon*. **4**:1199347. doi: 10.3389/fnrgo.2023.1199347.

Enhancing Cognitive Health in Older Adults through Combined Aerobic Exercise and Foreign Language Learning

Yijun Qian[1], Anna Schwartz[1], Yichi Zhang[2], Ara Jung[1], Gabi Wilds[1], Uri Seitz[2], Miso Kim[2], Arthur F Kramer[3], Leanne Chukoskie[1,2]

1 Department of Physical Therapy, Movement and Rehabilitation Science, Bouvé College of Health Science, Northeastern University, Boston, USA. qian.yiju@northeastern.edu; schwartz.ann@northeastern.edu; jung.ara@northeastern.edu; g.wilds@northeastern.edu; l.chukoskie@northeastern.edu

2 College of Art, Media and Design, Northeastern University, Boston, USA. zhang.yichi6@northeastern.edu; seitz.u@northeastern.edu; m.kim@northeastern.edu

3 The Center for Cognitive and Brain Health, Northeastern University, Boston, USA. a.kramer@northeastern.edu

Introduction

The efficacy of Aerobic Exercise (AE) and Cognitive Training (CT) programs in enhancing cognitive health among older adults has received considerable attention. While AE has consistently demonstrated its ability to induce significant neuroplasticity, thereby enhancing cognitive abilities [5] and overall well-being [3], studies on CT have yielded limited evidence regarding the transfer of learned skills to untrained tasks. This discrepancy may stem from the inherent specificity of traditional CT methods, which typically target isolated cognitive skills rather than encompassing a broader range of cognitive functions, such as attention [11] and working memory [12], reflective of daily tasks.

Recent research has highlighted the potential cognitive benefits of Foreign Language Learning (FLL), revealing improvements in brain structure and function across both young and older adults [4, 6]. Given these findings, the prospect of FLL serving as a viable strategy for mitigating cognitive aging [1, 2] appears promising, presenting itself as an alternative to conventional CT programs. FLL entails a complex learning process that engages a diverse array of cognitive skills, suggesting that it serves not only as a form of acquisition but also as a mode of cognitive training.

In recent years, research has delved into the synergistic effects of combining exercise and cognitive training, which have demonstrated greater efficacy in promoting cognitive function and enhancing learning [14] compared to interventions conducted individually. However, existing studies primarily focus on either improvement in learning or cognitive function. For example, some examine the cognitive effects of combined exercise and cognition training [15], while others assess the learning performance of combined exercise and vocabulary acquisition [16]. There is a lack of research examining both learning outcomes and changes in cognitive function resulting from combined aerobic exercise and language learning. We believe this is due to the current lack of a seamless paradigm for delivering targeted language learning, beyond mere vocabulary memorization, concurrently with exercise components.

We address this challenge by creating a virtual world tourism scenario gamified with language learning and exercise, combining auditory-based language learning with aerobic cycling. We envision that this computer-based, auditory language learning paradigm will serve as a tool to investigate several areas of interest. Firstly, it will help bridge the gap in understanding auditory-based Foreign Language Learning (FLL) in older adults. Secondly, we can explore the effects of exercise on auditory-based FLL. Lastly, we aim to examine the impacts of combining aerobic exercise (AE) and FLL strategies on improving cognitive function in older adults. While we may not be able to address all these questions simultaneously, the primary goal of our current pilot study is to assess the feasibility of this combined paradigm in foreign language learning outcomes and to observe whether short-term FLL learning affects cognitive function performance in older adults.

To test a new training method, we created an immersive virtual bike tour in Castilian Spanish. It is divided into episodes, blending exercise games to keep participants engaged. The tour offers immersive visuals and auditory cues while monitoring heart rates. The tour scripts include Spanish phrases spoken by a "biking buddy" who prompts participants to repeat after him. Our approach relies on older learners' ability to absorb language implicitly through pattern recognition, even without direct teaching. This learning environment is designed to mimic many features of how children acquire their first language in that it contains no written text, relies entirely on auditory processing, and conveys information only by the learner hearing speakers using the same structures repeatedly. Importantly, this method has been demonstrated to be a feasible way for adults to extract and acquire syntactical patterns from artificial language [8].

The primary aim of this article is to introduce the method and design of our paradigm. Additionally, we conducted a case study to test the efficacy of language learning after four 30-minute sessions among 17 older adults with minimal exposure to Spanish (10 with exercise, 7 without), recruited from the greater Boston community. Our preliminary analysis revealed significant improvements in Spanish learning in both the combined physical and language learning groups (AE+FLL) and the language learning-only group (FLL) after four sessions. In addition, our study shows that the mental efforts and frustration level is not different between groups even AE+FLL invested significant number of physical efforts than FLL only group. These findings suggest a promising approach for auditory-based Spanish learning in older adults, potentially shaping future exercise game development to comprehensively promote their health.

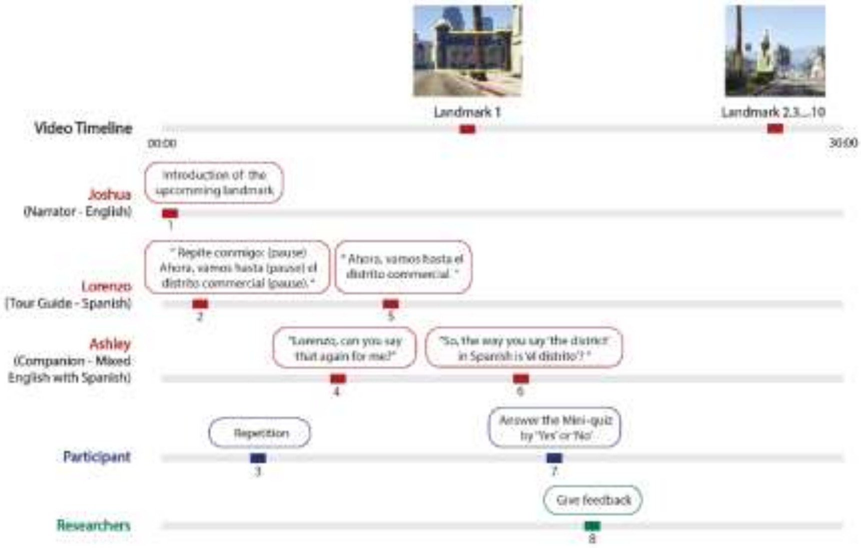

Landmark 1
Landmark 2,3....10
Video Timeline
00:00
30:00
Introduction of the upcomming landmark
Joshua
(Narrator - English)
1
" Repite conmigo: (pause) Ahora, vamos hasta (pause) el distrito commercial (pause)."
" Ahora, vamos hasta el distrito commercial. "
Lorenzo
(Tour Guide - Spanish)
2
5
"Lorenzo, can you say that again for me?"
"So, the way you say 'the district' in Spanish is 'el distrito'? "
Ashley
(Companion - Mixed English with Spanish)
4
6
Repetition
Answer the Mini-quiz by 'Yes' or 'No'
Participant
3
7
Give feedback
Researchers
8

Figure 1. This figure shows example audio-visual excerpts of the cycling and language learning game environment.

Method

Spanish learning in game

In our approach, participants extracted the meaning of the Spanish phrases by listening to cues from the English narration, observing the environment, such as highlighting landmarks, and noting what parts of the Spanish phrases are consistently repeated and which are not, relative to events on the tour. Each 30-minute virtual bike route includes 10 landmarks, such as churches, banks, or golf courses. At each landmark, a highlight appears to guide participants' attention to that landmark. The goal was to help learners recognize patterns in Spanish vocabulary as they listened and engaged with the virtual tour. Targeted Spanish knowledge was carefully selected for learners with minimal exposure to Spanish and categorized into the recognition of nouns and daily phrases, as well as several syntactical patterns in Spanish that differ from English: gender agreement with article, reversal of adjectival agreement with noun-gender, use of prepositions, and some basic contrasts involving pronouns/conjugations. During the game, participants' exposure to Spanish was entirely auditory-based with no visual text display. The structure of the Spanish was highly repetitive and used contrasting pairs to help learners identify patterns. Each phrase was presented to participants in brief chunks of a few words at a time, which they were encouraged to repeat immediately, and they would then hear the full sentence again. After they had heard each target phrase with a particular noun or noun phrase (e.g., “la iglesia”) twice, they would be prompted with a one-question comprehension quiz to receive some feedback on the patterns they were extracting, see Figure 1.

System overview

The virtual system was developed by modifying the open-world game Grand Theft Auto 5 (GTA 5) using the C# programming language. We selected GTA 5 due to its extensive and detailed geographical content, offering a remarkably authentic and high-quality representation of landmarks. Third-party mods, including gta5-real-mod, were incorporated to create a realistic 2D virtual city bike tour for participants. The narrative script and Spanish audio were generated using the Voicemaker website. We recorded footage from GTA 5 and then edited this in Adobe Premiere to incorporate AI-generated audio and visual cues of landmark highlights. The experiment consisted of four 30-minute cumulative virtual bike routes. Three non-playing characters (NPCs) accompany the participant on the tour, see Figure 1.

- Joshua (Narrator): A male native English speaker who acted as the tour guide in the game. He introduced landmarks to participants and gave detailed descriptions of landmarks in English.
- Lorenzo ("Bike Buddy"): A male native Castilian Spanish speaker who acted as the local Spanish tour guide as well as a Spanish instructor in the game. He spoke Spanish phrases to participants according to which landmarks were coming up and directed participants to repeat after him to learn.
- Ashley (Companion): The female native English speaker and travel companion, learning Spanish in the game. She requested Lorenzo to repeat Spanish sentences and actively engaged participants in the learning process, helping with in-game quizzes and providing English-accented Spanish examples.

Experiment procedures

The study consisted of six visits over approximately eighteen days. During the initial visit, participants had their blood pressure checked, received explanations, and completed consent forms. A demographic survey collected information on age, gender, education, marital status, and income, along with a Spanish proficiency pre-test. After the first visit, participants were randomly assigned to the AE+FLL Group (combining biking and language learning) or the FLL Group (language learning only). Visits 2 to 5 involved auditory-based Spanish learning, with or without stationary biking. Researchers provided heart rate monitors, adjusted bike seats, and set target heart rate ranges based on age (64% to 76% of (220 - age)). In the AE+FLL group, participants pedaled while maintaining their heart rate within the target range for 30 minutes. A supervising researcher managed equipment, monitored emergencies, and provided quiz feedback. These visits lasted 60-90 minutes each. After the 5th visit, participants returned for a 6th visit to complete additional cognitive tests and a post-test for Spanish knowledge.

Experiment environment set-up

All experimental sessions took place in the same laboratory with standardized equipment, including a stationary bike, an LG mobile phone, two computer monitors, a Polar H10 chest HR monitor, a Rhythm 24 Scosche armband HR monitor, and an EyeTech VT Mini3 eye-tracking device (for future analysis, although eye movement data is not presented in this manuscript). We used Lab Streaming Layer (LSL) to collect raw ECG data from the Polar H10 and stream it alongside data from the EyeTech VT Mini3 for future analysis. The LG phone ran the Elite HRV application, allowing participants to monitor their heart rates in real-time. We recorded experimental video and audio using OBS Studio and PsychoPy for researchers to analyze learning processes.

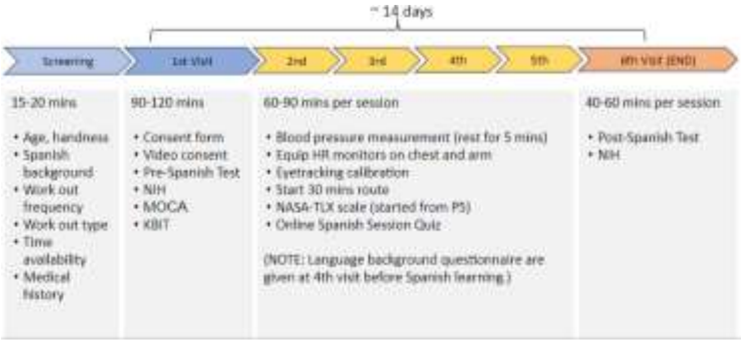


Figure 2. The experiment procedure plan for 6 sessions visits across two weeks.

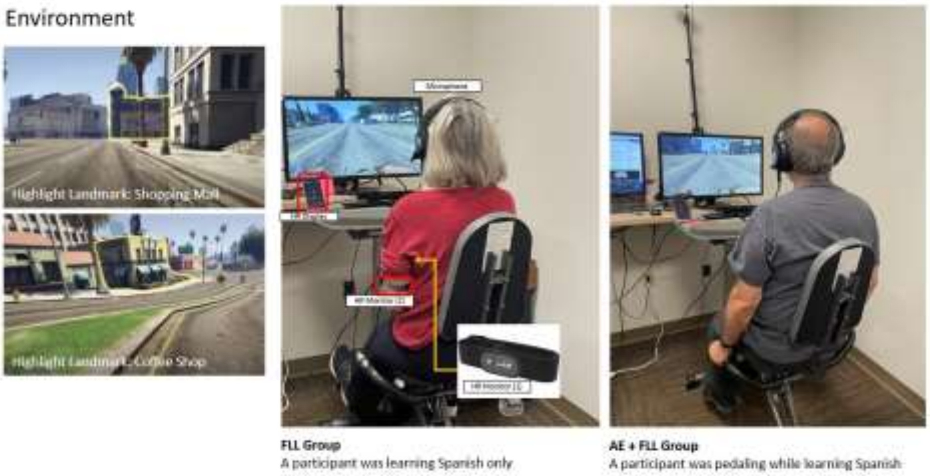

Environment
Highlight Landmark: Shopping Mall
Highlight Landmark: Coffee Shop
FLL Group
A participant was learning Spanish only
AE + FLL Group
A participant was pedaling while learning Spanish

Figure 3. The experiment environment and game environment.

Expected results

We anticipated slight improvements in learning for the AE+FLL group compared to the FLL-only group after a brief training period. Additionally, we expected that the AE+FLL group would have higher physical demands but a similar level of mental demands as the FLL group. Furthermore, we predicted that participants in the combined training group would not feel overwhelmed by the dual-tasking load.

Case Study

Participant recruitment

Participants were recruited via posters around the University, senior centers, email, and phone calls. Eligibility was determined through phone or online screening, with generally healthy individuals sought, free from significant physical impairments, visual or hearing issues. Exclusions applied to those with a history of specific neurological or heart-related conditions, except for individuals with controlled high blood pressure. All participants had minimal Spanish exposure, no prior Spanish learning experience, and hadn't engaged in over 60 minutes of moderate exercise weekly in the past 6 months. Participants meeting the cutoff standards for older adults in cognitive assessments (MoCA >= 24, KBIT-2 >= 85) were included. Seventeen participants (6 female, 11 male), aged 65 to 85 (M=73.65, SD=4.29), were randomly assigned to the biking language learning group (AE+FLL, N=10) or language learning-only group (FLL, N=7). There were no significant differences in age, global cognitive functions (MoCA), IQ (KBIT), or existing Spanish knowledge between the groups. Approval was obtained from the Northeastern University Institutional Review Board (IRB).

Data collection

Participants' accuracy rates on the Spanish pre- and post-tests were collected. These tests covered various vocabulary terms and assessed each syntactical function targeted during the experiment. Demographics like age, gender, marital status, education, and income, and language history via The Language History Questionnaire (LHQ), were collected. Cognitive assessments included the Montreal Cognitive Assessment (MoCA), Kaufman Brief Intelligence Test Second Edition (KBIT-2), and NIH toolbox tests, such as the Flanker Inhibitory Control and Attention Test (Attention), List Sorting Working Memory Test (WM), and Pattern Comparison Processing Speed Test (PS). These scores were age-adjusted and derived from standardized tests. Pre-training cognitive tests established a baseline for evaluating post-training differences. The level of cognitive loads when completing tasks was measured through NASA-TLX at the end of the 4th training session. Due to the addition of NASA-TLX after the fourth participant, cognitive load measurements were collected for only twelve participants.

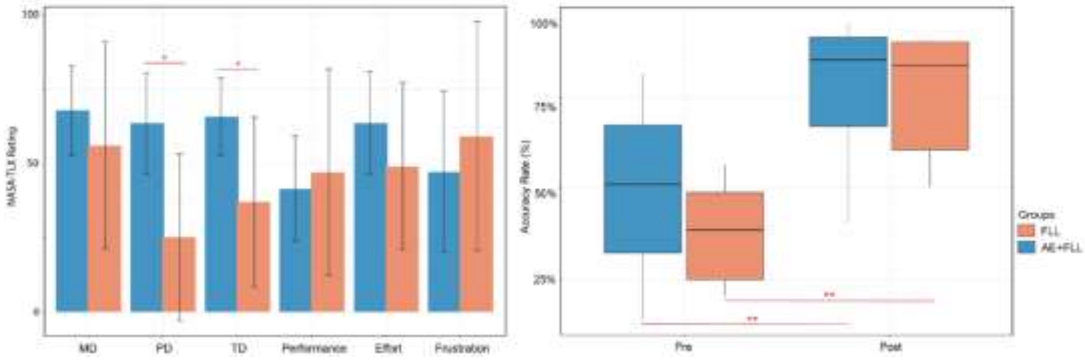


Figure 4. The graph on the left shows the NASA-TLX rating between groups. The graph on the right shows participants' Spanish learning performance in pre-training and post-training. Both pre- and post-include questions on nouns, noun gender, adjective order, prepositions, pronouns.

Results

Cognitive loads

The NASA-TLX rating, including Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Performance, Effort, and Frustration, were compared between groups. The normality of the data was assessed using the Shapiro-Wilk test. An unpaired t-test was conducted to assess the difference between MD, TD, and Performance, while the Wilcoxon rank sum test was used to assess PD, Effort, and Frustration between groups. No significant differences were found between MD, Performance, Effort, and Frustration. However, PD between groups showed a significant difference (*W = 31, p = 0.033**), and a difference was also observed in Temporal Demands (*t* = 2.4769, *df* = 6.091, *p* = 0.047*), see Figure 4 (Left).

Language Learning Performance

A paired sample t-test was used to assess increases between participants' pre- and post-test accuracy for both groups. We found that Spanish learning outcomes were significantly improved in both the AE + FLL group (*t* = -7.1815, *df* = 9, *p* < 0.00001***) and the FLL group (*t* = -6.8287, *df* = 6, *p* = 0.0004***) after four sessions of training, see Figure 3. We examined potential differences in participants' pre-test and post-test rates as a function of group using independent samples t-test. We did not find significant differences between groups' pre-test accuracy rates (*p* = 0.3783), indicating a similar baseline level for each group. We also found no significant differences between the two groups' post-test performance (*p* = 0.8627), see Figure 4 (Right).

Cognitive task performance

A paired t-test was used to conduct a within-group analysis in comparing the pre- and post-test performances in three NIH Toolbox tasks: Attention, Processing Speed (PS), and Working Memory (WM) tests. The cognitive measurements of one participant are excluded because of an unexpected event during the collection. Both groups show the trend of increased cognitive performance after four sessions but only the FLL group shows the statistical differences in PS test (*t = -5.7867, df = 6*, p = *0.001***) and only AE +FLL group shows the trend of increased working memory.

Discussion

In this study, we share a novel design combining exercise and an audio-based language learning program. Differing from traditional text-based language learning, we employed the "Listen and Repeat" approach with audio-based inspired by how children acquire their first language. The preliminary results show learning improvements regardless of whether participants' language learning was paired with exercise. This result suggests that the combination of aerobic physical exercise and auditory-based foreign language learning may be a viable offering for older adults. There has not been extensive study on learning during exercise; however, some studies suggest that the combination can enhance learning outcomes [10]. Given the potential for cognitive overload of the combination of learning and exercise we used, we are heartened to see statistically equivalent learning in groups. Other results [9, 13] suggest that over longer periods of training, the neuroplasticity boost wrought by aerobic physical activity will promote better learning, at least in studies that use more standard cognitive training paradigms.

To our surprise, the perceived mental demands and frustration level is not different between groups, in fact, the solo task group (FLL) shows a slightly higher frustration rate. This makes us believe that the combination of exercise and learning tasks is at least not too much for older adults compared to learning task alone. Consistent with previous research on exercise adding benefits to executive functions, we observed the increase in the AE+FLL group that Flanker and List Sort Working Memory [7]. Both groups showed increased attention scores after training. However, only the AE + FLL group showed an improvement post-training in our working memory task. So far, only the FLL group has shown significant improvements in processing speed. If the findings, such as the enhanced FLL learning outcomes and cognitive functions, remain consistent in a larger scale study, the inclusion of aerobic exercise alongside a cognitive training program could potentially offer older adults additional advantages. These benefits might not only pertain to their anticipated physical health benefits but could also extend to increased engagement and motivation. This combined strategy may provide a potential direction for engaging older adults in motor-cognitive activities.

Acknowledgements

We want to thank The Center for Cognitive and Brain Health and Northeastern University for funding this study.

References

1. Antoniou, M., Gunasekera, G.M., & Wong, P.C. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: A hypothesis for future research. *Neuroscience & Biobehavioral Reviews*, **37**, 2689-2698.

2. Bialystok, E., Craik, F.I., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, **45**, 459-464.

3. Biddle, S.J., Atkin, A.J., Cavill, N., & Foster, C. (2011). Correlates of physical activity in youth: a review of quantitative systematic reviews. *International Review of Sport and Exercise Psychology*, **4**, 25 - 49.

4. Bubbico, G., Chiacchiaretta, P., Parenti, M., di Marco, M., Panara, V., Sepede, G., Ferretti, A., & Perrucci, M.G. (2019). Effects of Second Language Learning on the Plastic Aging Brain: Functional Connectivity, Cognitive Decline, and Reorganization. *Frontiers in Neuroscience*, **13**.

5. Hillman, C.H., Erickson, K.I., & Kramer, A.F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Reviews Neuroscience*, **9**, 58-65.

6. Mårtensson, J., Eriksson, J., Bodammer, N.C., Lindgren, M., Johansson, M., Nyberg, L., & Lövdén, M. (2012). Growth of language-related brain areas after foreign language learning. *NeuroImage*, **63**, 240-244.

7. Pontifex, M.B., Hillman, C.H., Fernhall, B., Thompson, K.M., & Valentini, T.A. (2009). The effect of acute aerobic and resistance exercise on working memory. *Medicine and science in sports and exercise,* **41** 4, 927-34 .

8. Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, **6**, 855-863.

9. Schaefer, S., & Schumacher, V. (2010). The Interplay between Cognitive and Motor Functioning in Healthy Older Adults: Findings from Dual-Task Studies and Suggestions for Intervention. *Gerontology*, **57**, 239 - 246.

10. Schmidt-Kassow, M., Kulka, A., Gunter, T.C., Rothermich, K., & Kotz, S.A. (2010). Exercising during learning improves vocabulary acquisition: Behavioral and ERP evidence. *Neuroscience Letters*, **482**, 40-44.

11. Tomlin, R.S., & Villa, V.M. (1994). Attention in Cognitive Science and Second Language Acquisition. *Studies in Second Language Acquisition*, **16**, 183 - 203.

12. Wen, Z.(., Mota, M.B., & Mcneill, A. (2015). Working memory in second language acquisition and processing.

13. Wollesen, B., & Voelcker-Rehage, C. (2014). Training effects on motor–cognitive dual-task performance in older adults. *European Review of Aging and Physical Activity*, **11**, 5-24.

14. Anderson-Hanley, C., Arciero, P. J., Brickman, A. M., Nimon, J. P., Okuma, N., Westen, S. C., Merz, M. E., Pence, B. D., Woods, J. A., Kramer, A. F., & Zimmerman, E. A. (2012). Exergaming and older adult cognition: a cluster randomized clinical trial. *American journal of preventive medicine*, *42*(2), 109–119.

15. Takeuchi, H., Magistro, D., Kotozaki, Y., Motoki, K., Nejad, K. K., Nouchi, R., Jeong, H., Sato, C., Sessa, S., Nagatomi, R., Zecca, M., Takanishi, A., & Kawashima, R. (2020). Effects of Simultaneously Performed Dual-Task Training with Aerobic Exercise and Working Memory Training on Cognitive Functions and Neural Systems in the Elderly. *Neural Plasticity*, *2020*, 3859824. https://doi.org/10.1155/2020/3859824
16. Schmidt-Kassow, M., Kulka, A., Gunter, T. C., Rothermich, K., & Kotz, S. A. (2010). Exercising during learning improves vocabulary acquisition: behavioral and ERP evidence. *Neuroscience letters*, *482*(1), 40–44.

Effect of Sampling rate and data source on rhythmicity computation

S. Rey[1], H.R Nasser[2] and M. Cockburn[1]

[1] Animals, Products of Animal Origin and Swiss National Stud, Agroscope, Haras national suisse HNS, Avenches, Switzerland. sonia.rey@agroscope.admin.ch [2] Digital production group, Agroscope, Posieux, Switzerland.

Abstract

Rhythmicity computation is increasingly becoming a popular tool to measure the well-being of animals. By analyzing various data related to biological processes, behaviors, or activities, one can effectively assess the rhythmic patterns exhibited by animals. In our current study, we evaluate locomotor activity of horses from a group housing system using accelerometers to analyze their rhythmic patterns. We then apply a Fourier Transformation to the resulting time-series data. This allows us to calculate the Degree of Functional Coupling (DFC) [1] a proxy for rhythmicity of organisms. The DFC measures the degree to which organisms synchronize with their environment, particularly focusing on the harmonic components. This parameter measures the magnitude and intensity of the animals' circadian rhythm. Indeed, the majority of individuals exhibit circadian rhythms, indicating that their behaviors follow a 24-hour cycle. This rhythmic pattern is evidenced by the prominence of harmonic frequencies within the 24-hour cycle. Notably, these harmonics correspond to frequencies derived from dividing the 24-hour period by integers less than 24 (e.g., 24/1, 24/2, ...) [2]. Research suggests, that a good synchronicity with the environment reflects a good state of welfare. The DFC can take values ranging from zero to one, where a low value indicates a de-synchronization of the organism, whereas a value closer to one indicates more harmonic frequencies and a high synchronization of the individual with its environment [3]. The use of this parameter has been used to assess the welfare of extensively managed sheep [4] and housed dairy cows [5]. Moreover, it has been successfully used to evaluate how wild horses engage in activity and feeding [6].

However, data sources can vary giving very different information on animal rhythms. Feeding stations will measure the animals visits perhaps 3-15 times per day, whereas accelerometers can collect data at rates of up to 20Hz. Little research has addressed this issue; therefore, we investigated the effect of sampling rate and data source in a pilot study carried out on equine accelerometer data. Accelerometer devices were placed on the horse's front leg.

The x, y and z axis of equine locomotion were collected by a wireless MSR data logger at a frequency of 1 Hz. The data was collected within a different experiment, which was approved by ethical animal welfare commission of the Freiburg Cantonal authorities (2023-40-FR). The DFC was computed using the DigiRhythm R Package [2] with data settings: First we analyzed each axis of the same data set separately, as well as the squared sum of all three axis. After deciding, to proceed with the squared sum of all axis, we used this data to compare sampling rates. We used the resampling function in the DigiRhythm package to resample the data from 1 Hz to 1/60 Hz, 1/600 Hz and 1/900 Hz.

The results show that DFC values differed between both, data source and data frequency analyzed, see Figure 1. Therefore, we conclude that DFC computation is sensitive to data source and frequency, arguing that it's application without should be considered cautiously and validated in the specific setting.

We therefore encourage researchers to be aware of the impact sampling frequency and device used for data collection, and sensitize them to risk of comparing absolute DFC values between studies.

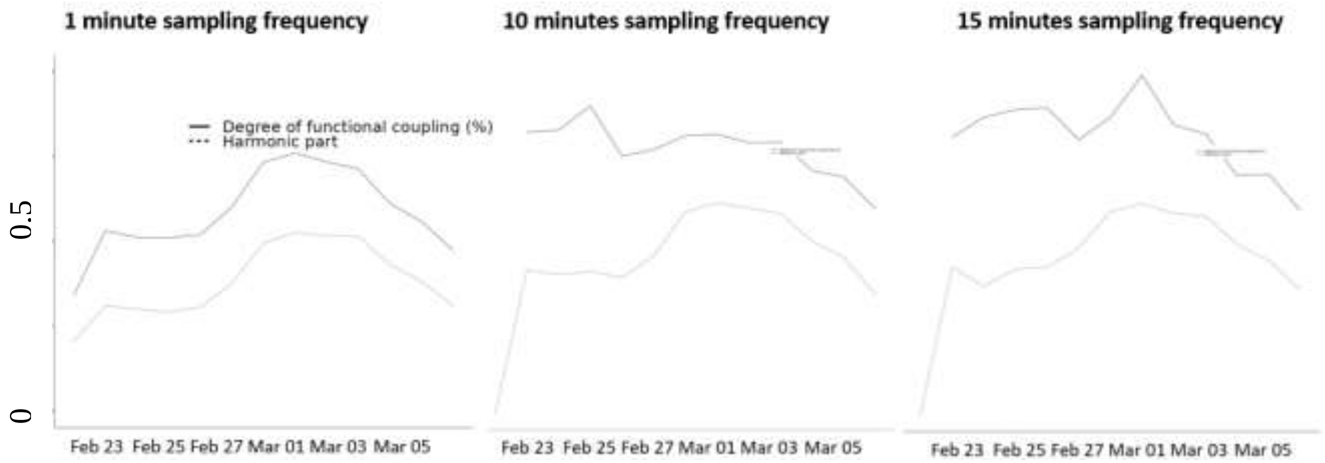


Figure 1. Degree of function coupling (DFC) (solid line) in function of time at 1 minute, 10 minutes and 15 minutes sampling frequencies respectively. Harmonic part is presented as a dashed line.

References

1. Scheibe, K. M., A. Berger, J. Langbein, W. J. Streich, and K. Eichhorn. (1999). Comparative Analysis of Ultradian and Circadian Behavioural Rhythms for Diagnosis of Biorhythmic State of Animals. *Biological Rhythm Research* **30**(2):216-233.

2. Nasser, H.-R., Schneider, M., and Cockburn, M. (2023). DigiRythm: an R package for rhythmicity assessment using the degree of functional coupling. Manuscript submitted for publication.

3. Sinz, R., and Scheibe, K.M. (1976). Systemanalyse der multioszillatorischen Funktionsordnung im zirkadianen und ultradianen Frequenzbereich und ihr Indikationswert für Belastungswirkungen, dargestellt am Beispiel verschiedener Licht-Dunkel-Verhältnisse bei der Intensivhaltung von Schafen. *Acta Biologica et Medica Germaniae* **35**: 465–414.

4. Sarout, B. N. M., Waterhouse, A., Duthie, C. A., Poli, C. H. E. C., Haskell, M. J., Berger, A., & Umstatter, C. (2018). Assessment of circadian rhythm of activity combined with random regression model as a novel approach to monitoring sheep in an extensive system. *Applied Animal Behaviour Science* **207**: 26-38.

5. Fuchs, P., F. Adrion, A. Z. Shafiullah, R. M. Bruckmaier, and C. Umstätter. (2022). Detecting ultra-and circadian activity rhythms of dairy cows in automatic milking systems using the degree of functional coupling—A pilot study. *Frontiers in Animal Science* **3**: 839906.

6. Berger, A., K.-M. Scheibe, K. Eichhorn, A. Scheibe, and J. Streich. (1999). Diurnal and ultradian rhythms of behaviour in a mare group of Przewalski horse (Equus ferus przewalskii), measured through one year under semi-reserve conditions. *Applied Animal Behaviour Science* **64**(1):1-17.

The aWISH project: Automated animal welfare monitoring for broilers and pigs at the slaughterhouse

T.B. Rodenburg[1], N. van Staaveren[1], M.F. Giersberg[1], N. Van Noten[2], K. van Langeveld[1,2], R. Klont[3], B. de Ruiter[3], M. Bouwknegt[3], B. Peijenburg[3], K. Mulder[4], F.A.M. Tuyttens[2]

[1]Animals in Science and Society, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands. t.b.rodenburg@uu.nl

[2] ILVO (Flanders Research Institute for Agriculture, Fisheries and Food), Merelbeke, Belgium

[3] VION Food Group, Boxtel, The Netherlands

[4] Connecting Agri & Food B.V., Uden The Netherlands

Animal welfare monitoring of farm animals on commercial farms is difficult to realize. Yet, collecting objective data on specific welfare indicators can be valuable to assess how a given farm or chain is performing and can be used to provide targeted advice. All farm animals that are kept for meat production will arrive at a slaughterhouse at the end of the growing period. Sensor technology placed at the slaughterhouse can provide information on specific welfare indicators, and perhaps even aggregated overall welfare scores, and can be used to provide feedback to farmers, catching teams (in the case of poultry), transporters and slaughterhouses.

Within the aWISH project, we aim to develop methodology for automated animal welfare monitoring for broilers and pigs at the slaughterhouse. We do this by setting up six pig and broiler pilots in the different countries participating in the project. Each pilot focuses on connecting data collected at the farm (either routinely or by researchers) to data collected automatically at the slaughterhouse and investigating whether the slaughterhouse data can provide valid information regarding the on-farm welfare of the animals involved. For instance, a broiler flock with a high incidence and severity of footpad lesions is likely to have problems with wet litter, low activity levels and a poor indoor climate during the growing period. Within aWISH, we aim to combine all this information through an interactive dashboard and we investigate whether in a future application the slaughterhouse data alone can be used to provide feedback to the farmer regarding animal welfare.

The Dutch aWISH pilot takes place at a pig slaughterhouse in The Netherlands. Here, we collaborate with a specific chain that produces pork with one star within the Dutch ‘Better Life’ label, managed by the Dutch Society for the Protection of Animals. Within that chain, approximately 12 farms will be followed for 2 cycles. A number of welfare indicators will be recorded on farm at the beginning and at the end of the growing period. Data collection will be non-invasive and observational only, so no approval from an animal ethics committee is required. At slaughter, sensor data will be collected on the same pigs regarding vocalisations produced at unloading, wounds on the carcass and on ears and tail and tail length. Each farm has climate sensors installed, that for instance allow to monitor ammonia levels in the air. For one cycle, the farmers will receive climate advice from a trained adviser and for the other cycle, they will not (in randomized order). It is expected that the sensor data collected at slaughter can be connected to the welfare data collected on-farm. Furthermore, climate advice is expected to result in better welfare outcomes measured at slaughter.

This project has received funding from the European Union’s Horizon Europe Research and Innovation Program under Grant Agreement No. 101060818.

Contact-free Measurement of Behavioural Readouts in the Lizard *Anolis carolinensis*.

N. Röhrdanz[1], A. Menon[1] and P.Wulff[1]

1. Institute of Physiology, Christian-Albrechts-Universität zu Kiel, Kiel, Germany. n.roehrdanz@physiologie.uni-kiel.de

Abstract

The green anole lizard (*Anolis carolinensis*) is an interesting model organism to study in cognitive neuroscience. It shows a brain of reduced complexity compared to mammals. Some brain structures like the hippocampus are already present in the lizards' brain while other brain structures like the neocortex seem to be missing. In this study, we aimed to establish a contact-free method to measure different behavioural readouts in the lizards: Respiratory rate, head direction, movement/ freezing and skin colour change. We used a virtual-reality based experimental testing chamber to present the stimuli in a virtual natural context. This setup also allowed recording of the animals' behaviour in 360°. Data of the animals posture was extracted with DeepLabCut[1]. 23 points on the lizards' body were tracked. Eight animals were recorded over a total of 13-17 sessions while different acoustic and visual stimuli were presented (neutral or predator). The lizards responded differently to the different stimulus-categories. These results indicate that our recording setup allows reliable contact-free measurements of different behavioural readouts of the green anole lizard from 360° videos. The extraction of the respiratory rate from video data in reptiles is novel and provides a great advantage against wearable sensors. We anticipate this method to prove useful in the analysis of arousal, fear and stress in other species, too.

Introduction

In physiology much of the research is done with mice or rats as model organisms. However, some scientific questions may profit from looking at different model organisms. It would for example be quite informative to study how cognitive functions, typically associated with neocortical regions in mammals, are achieved in circuits where the neocortex is absent. An example of a brain without a neocortical region is the reptilian brain. Even though reptiles do not have a neocortex there is evidence that lizards can solve tasks that in mammals require the neocortex [2]. At the same time the reptilian hippocampus is most likely a homologue to the mammalian hippocampus[3]. It is structured in three layers [4] and may have the same subregions as the mammalian hippocampus[5] . How do reptiles achieve mnemonic computations such as long-term consolidation or reversal learning that depend on hippocampal-neocortical dialog in mammals?

Classical experiments to study hippocampal function or other cognitive abilities involve complex behavioural paradigms like different types of mazes for the test of spatial navigation, working memory or reversal learning. In the past many behavioural experiments on lizards seemed to have failed (see [6,7] for overview).

The investigation of cognitive function in reptiles through paradigms that require active behaviour is more challenging compared to mice. Lizards show different exploration patterns and motivators. To overcome these limitations, we focus on the identification of passive behavioural readouts that could be used to gauge cognitive function e.g. in memory tasks. There are some behaviours that are promising candidates for readouts of arousal in Anolis *carolinensis*. Locomotion and head movement increased in the presence of predators[8–10]. Furthermore the respiratory rate was the first response that was conditioned to light in a fear conditioning experimental task [9]. Another readout that is special in the *Anolis carolinensis* is its ability to change body colour. *Anolis carolinensis* are able to change their body colour from a bright green to a dark brown. The body colour is associated to levels of arousal [11]. The skin colour is green, when the arousal level is low and brown, when it is high [11].

While all of these readouts seem to be good candidates of passive readouts they still have to be validated. A strong readout for arousal or fear like the freezing response in mice has not yet been reported. A key challenge is to find methods to analyse the behaviours of interest. The use of sensors for the measurement of the respiratory rate interferes with the experimental paradigm. Standard tracking methods for the analysis of movement are restricted to the mouse or rat model and can't track other animal models.

In this study we present a way to analyse the above mentioned readouts of arousal in *Anolis* from recorded video data by using deep-learning based animal tracking methods (DeepLabCut [1]). The contact free tracking of user defined body parts reduces the animals stress levels and enables the study of previously hidden behaviours. We used a virtual-reality based setup combined with a 360° camera to study the free-moving lizards' response to different audio-visual stimuli.

Methods

All experiments were performed in accordance with the German law on animal protection and approved by the Animal Care and Ethics Committee of the Christian-Albrechts-University, Kiel.

Anolis carolinensis
A total of 8 adult, female *Anolis carolinensis* were used in this study (SVL = 4.77 cm, weight = 3.05 g).

Setup
In this experiment we used a virtual-reality-based experimental setup (see Figure 1A), which allowed us to present different audio-visual stimuli in a semi-natural context. The setup consisted out of 4 17" computer screens (Dell 1703Fpt, 1280x1024 px). The monitors were placed around a transparent sphere made out of two transparent half-spheres (d = 35cm). In the centre of the sphere a 360° camera was placed (Insta360 one X2, 5.7k resolution). The camera was water-cooled using a custom build water-cooler in order to keep temperature inside the sphere constant. In the lower half of the sphere a symmetric branch pattern made from cork was glued onto the sphere. In the upper sphere air slits and an indirect lighting were positioned.

Stimuli
The aim of this study was to identify valid readouts for arousal in the *Anolis caroinensis*, we therefore chose different auditory and visual stimuli that should either have a neutral or a negative (predator) valence(see Figure 1C).

All visual stimuli were shown with a size of 5 cm, which is above the visual acuity of *Anolis* [12] regardless where inside the sphere the animals were located. All auditory stimuli were played at 65 dB SPL,which is 30-35dB above the hearing threshold [13]. All stimuli were presented for 5 seconds.

As stimuli with negative valence we used the contour (stationary, moving vertical and moving horizontal) of a red-tailed hawk and calls from the red-tailed hawk.The red-tailed hawk is a natural predator of *Anolis carolinensis* and previous observations indicated that the behaviour of *Anolis carolinensis* changes as response to the call of the red tailed hawk with increased distance to the sound source and signs of increased vigilance [8,10].

Furthermore we also used images of another anole lizard to investigate changes in arousal in social situations.

Procedure
Each animal participated in 13-17 sessions with one session a day. During the first three sessions the animals were put inside the VR-setup for 20 min to habituate. Those three days were followed by the experimental sessions. The experimental sessions had a duration of 30 minutes each. During the first 10 minutes no stimuli were presented. From minute 10 to 30 stimuli of one category and one sensory input were presented 10 times. The interstimulus interval was 90 seconds with a random offset between -30 to 30 seconds. The experimental sessions started with the neutral stimuli. The stimulus type for the session was drawn randomly. Each stimulus type was presented in two sessions. The predator sessions followed the neutral sessions to prevent anterograde interference. The last two sessions were the ones where another *Anolis carolinensis* was presented.

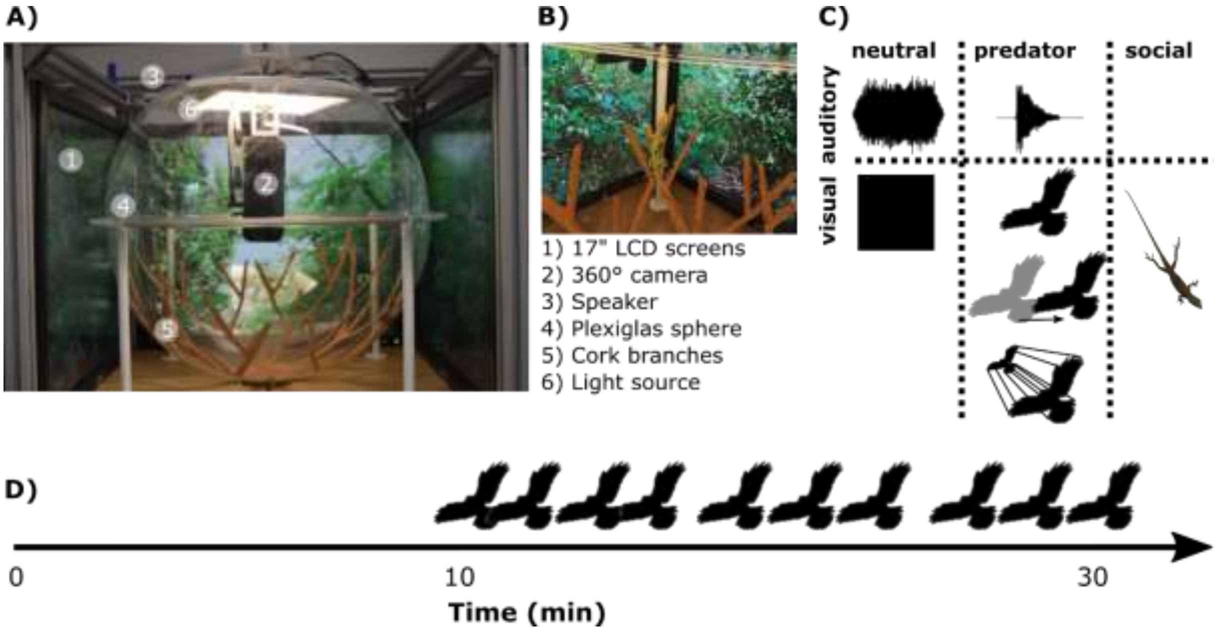


Figure 1: Setup and experimental design. Panel A) shows the virtual-reality-based experimental setup. Panel B) shows an example of a stimulus presentation. The stimuli used in this experiment are shown in panel C). The session time schedule is visualized in panel D).

Quantification and statistical analysis

We preprocessed 360° videos by making them planar videos with the animal in the centre (FOV = 75°, distortion = 0, 1440x1440px). Afterwards 23 bodyparts were tracked with the DeepLabCut software **[1]** (see Figure 2A-B).
Head movement: We calculated head movement from the extracted DeepLabCut[1] bodypart coordinates. First we calculated the signed angles between the $\overrightarrow{chest - neck}$ and $\overrightarrow{neck - nose}$ vectors which results in the head direction (see Figure 2C). Head movement was then calculated as change in head direction over time.

Respiratory rate: The green anole lizards do not have a diaphragm, thus respiration is achieved by thorax movement alone. This movement results in a change of thorax width. We calculated the thorax-width as euclidean distance between the left and right outer chest (see Figure 2C). We then calculated the respiratory rate from the length of the respiratory cycles.

Leg movement: Since we preprocessed the 360°videos so that the *Anolis* was always in the centre of the frame, movement couldn't be simply extracted from the movement of one bodypart in space over time. Instead we calculated the change in angles between the $\overrightarrow{left\ hind\ hip - right\ hind\ hip}$ and $\overrightarrow{right\ hind\ hip - right\ hind\ knee}$ vectors (see Figure 2C).

Skin colour: The skin colour in *Anolis carolinensis* is linked to arousal[11]. Aroused *Anolis* show a brown skin colour while otherwise the skin colour is green. In order to reduce dimensionality we converted the RGB-recorded videos to grayscale. Since green is brighter than brown this conversion results in ordinal data. We calculated the anoles skin colour by averaging the grayscale values in the chest region (surface between neck, right outer chest, tailbase, left outer chest; see Figure 2C).

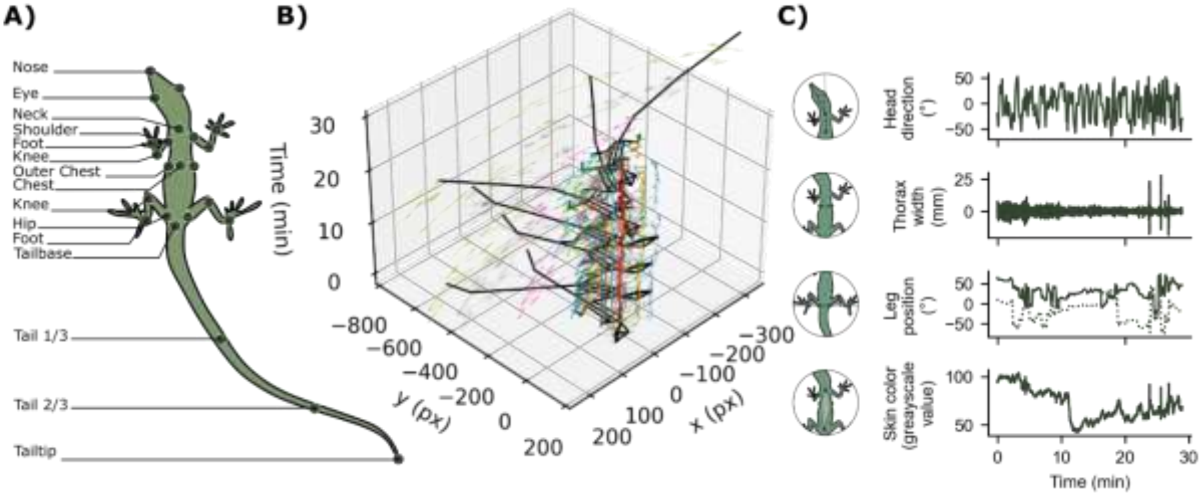


Figure 2: Extraction of behavioural readouts of *Anolis carolinensis* from video data. Panel A) shows the labeling scheme that was used to train the DeepLabCut[1] algorithm. 23 points are tracked in the videos and can be later used for the extraction of behavioural readouts. In Panel B) an example of the spatiotemporal data from one session is shown. The principles of the extraction of the behavioural readouts that are reported whithin this paper are shown in Panel C).

To analyze the effect of the different auditory or visual stimuli on the *Anolis carolinensis*, we analyzed changes in each of the readouts that occoured during the presentation of the stimuli. The percentage of change was used as measure. Therefore we calculated the realtive change of the 15 seconds before stimulus onset to the 30 seconds after stimulus offset. The likelihood parameter in DeepLabCut[1] indicates the condfidence ot the model in the prediction of the tracked bodypart. Data points where the likelihood parameter of the DeepLabcut[1] bodypoints was below .9 for the respiratory rate and .8 for the other measures were excluded. Whithin animal data was averaged per animal and condition. For each behavioural readout we calculated a repeated measures Anova with paiwise post-hoc tests.

Results

Head movement

There are differences in “change in head movement” between the different stimulus type combinations (F(8,56) = 2.529, p < .0210, η^2 = 0.248, sphericity corrected). Head movement is increased after the presentation of “Neutral image” (t(7)= 3.740, p = 0.0036), “Neutral sound” (t(7)= 3.603, p = 0.0043), “Predator image” (t(7)= 2.660, p = 0.0162), “Predator sound” (t(7)= 2.437, p = 0.0225) and “Predator audiovisual” (t(7)= 2.486, p = 0.0444).

Respiratory rate

There are no differences in “respiratory rate” between the different stimulus type combinations (F(8,56) = 0.852, p < . 6377, η^2 = 0.100). The respiratory rate does not change significantly during the presentation of any stimulus except the ”Predator sound” (t(7) = 2.684, p = .0157). The changes of the respiratory rate are smaller compared to the other readouts and the variance is high.

Leg Movement

There are differences in “leg movement” between the different stimulus type combinations (F(8,56) = 4.822, p < .0123, η^2 = 0.350, sphericity corrected). Leg movement is increased after the presentation of “Neutral image” (t(7)= 3.629, p = 0.0042), “Neutral sound” (t(7)= 4.700, p = 0.0011), “Neutral whitenoise” (t(7)= 3.811, p = 0.0159), “Predator image” (t(7)= 3.492, p = 0.0051) and “Predator sound” (t(7)= 2.334, p = 0.0262). Leg movement is increased more during the presentation of neutral stimuli compared to stimuli of predators (t(7)=3.359, p = .006), this might be a weak indicator for the existence of a freezing response in *Anolis carolinensis*.

Colour change

There are differences in “colour change” between the different stimulus type combinations (F(8,56) = .896, p < .0001, η² = 0.143, sphericity corrected). Colour change is increased after the presentation of “Neutral sound” (t(7) = 3.054, p = 0.0092) and “Pred image” (t(7) = 2.239, p = 0.0332).

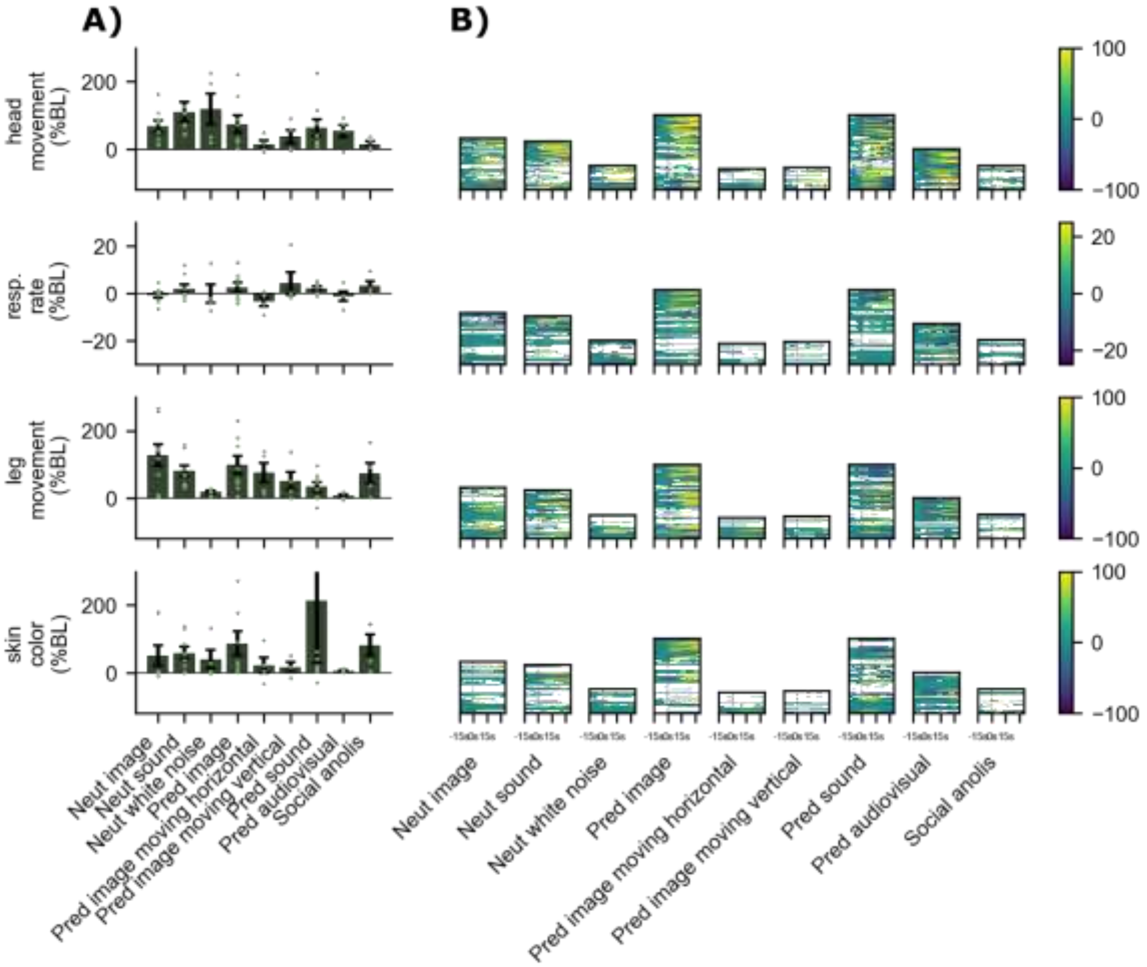


Figure 3: Event related changes in the different behavioural readouts. Panel A) shows percentual change in each readout. Change is calculated as M(x[stimulus onset + 5 seconds : stimulus onset + 35 seconds]) / M(x[stimulus onset -15 sec : stimulus onset]). In Panel B) the baseline corrected (15 seconds before stimulus onset) timecourses for each trial in each condition is shown. White areas represent unsufficient tracking data, that was filtered out because the likelihood of the points was too low.

Discussion

New deep-learning-based video tracking methods like Deeplabcut[1] can be used for precise tracking of user defined body parts of the animal as well as detection and quantification of different behaviours.

The presented methods allow the contact free observation of user defined active and passive animal behaviours that are hidden to standard tracking algorithms. We have demonstrated the use of those algorithms for the quantification of behaviour of the lizard *Anolis carolinensis*. We were able to quantify different movement- as well as exploration-related behaviours. Furthermore, we presented a unique way to extract respiratory rate. *Anolis* showed increased leg movement after stimulus presentation of neutral valence, this might be weak evidence for the presence of the freezing response in *Anolis*. Head movement is increased after most of the stimulus presentations which might be due to increased vigilance. The respiratory rate increased after the presentation of the predator call. This could be a possible preparation for a flight response when the location of the predator is unknown.

In the future the data could be further processed with other promising deep-learning algorithms. This would give us the opportunity to quantify more complex behaviours or identify new behavioural readouts. For supervised algorithms the SimBA[14] toolkit is a promising option as it goes hand in hand with the DeepLabCut[1] data. For unsupervised algorithms B-SOiD[15] might be an interesting option. Supervised algorithm will be useful if the ethogram of interest is known while unsupervised algorithms may help to identify behaviours of relevance.

Passive behavioural readouts are best suited for the investigation of cognitive processes in lizards as they are not requiring active participation of the animals.

References

1. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21: 1281–1289. doi:10.1038/s41593-018-0209-y
2. Leal M, Powell BJ. Behavioural flexibility and problem-solving in a tropical lizard. Biol Lett. 2012;8: 28–30. doi:10.1098/rsbl.2011.0480
3. Naumann RK, Ondracek JM, Reiter S, Shein-Idelson M, Tosches MA, Yamawaki TM, et al. The reptilian brain. Curr Biol. 2015;25: R317–R321. doi:10.1016/j.cub.2015.02.049
4. Luzzati F. A hypothesis for the evolution of the upper layers of the neocortex through co-option of the olfactory cortex developmental program. Front Neurosci. 2015;9. doi:10.3389/fnins.2015.00162
5. Tosches MA. From Cell Types to an Integrated Understanding of Brain Evolution: The Case of the Cerebral Cortex. Annu Rev Cell Dev Biol. 2021;37: 495–517. doi:10.1146/annurev-cellbio-120319-112654
6. Burghardt GM. Learning processes in reptiles. 1977.
7. De Meester G, Baeckens S. Reinstating reptiles: from clueless creatures to esteemed models of cognitive biology. Behaviour. 2021;158: 1057–1076. doi:10.1163/1568539X-00003718
8. Shiho T, Sakai O, Iwai N. Exploration of aversive bioacoustics for the effective management of invasive green anoles (Anolis carolinensis). J Nat Conserv. 2022;68: 126215. doi:10.1016/j.jnc.2022.126215
9. Davidson RE, Richardson AM. Classical conditioning of skeletal and autonomic responses in the Lizard (Crotaphytus collaris). Physiol Behav. 1970;5: 589–594. doi:10.1016/0031-9384(70)90085-5
10. Cantwell LR, Forrest TG. Response of *Anolis sagrei* to Acoustic Calls from Predatory and Nonpredatory Birds. J Herpetol. 2013;47: 293–298. doi:10.1670/11-184
11. Greenberg N. Ethological Aspects of Stress in a Model Lizard, Anolis carolinensis. Integr Comp Biol. 2002;42: 526–540. doi:10.1093/icb/42.3.526
12. Fleishman LJ, Yeo AI, Perez CW. Visual acuity and signal color pattern in an *Anolis* lizard. J Exp Biol. 2017; jeb.150458. doi:10.1242/jeb.150458
13. Brittan-Powell EF, Christensen-Dalsgaard J, Tang Y, Carr C, Dooling RJ. The auditory brainstem response in two lizard species. J Acoust Soc Am. 2010;128: 787–794. doi:10.1121/1.3458813
14. Nilsson SR, Goodwin NL, Choong JJ, Hwang S, Wright HR, Norville ZC, et al. Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals. Animal Behavior and Cognition; 2020 Apr. doi:10.1101/2020.04.19.049452
15. Hsu AI, Yttri EA. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. Nat Commun. 2021;12: 5188. doi:10.1038/s41467-021-25420-x

Rate and Patterns of Inter-Game Variability in Team Sports: Implications for Sample Size

Adam Sewell[1], Matthew T. Robins[2], Shelley A. Ellis[1] and Andrew J. Callaway[1]

[1]Bournemouth University and [2]University of Kent

Introduction

Although the importance of variability has been widely reported in human movement [1, 2], the topic remains relatively unexplored in team sports. Examining the variability between games may reveal patterns or trends contributing to a deeper understanding of team behavior throughout a competitive season. Time-lag analysis has been proposed as a method to quantify variability in ecological communities over relatively small-time intervals (i.e between 10 and 40) [3]. However, it has yet to be utilised in team sports which would address the limitations posed by traditional time series analysis approaches that require large sample sizes. Therefore, the aims of the study are: 1) introduce time-lag analysis to explore inter-game variability within team sports, and 2) examine implications for sample size on the rate and patterns of inter-game variability.

Method

Sample

The sample consisted of 38 games for a single team from the 2022-23 English Premier League season. The data was obtained from the Football Ref open-source platform (https://fbref.com/en/). Six performance indicators were selected as common indicators used to evaluate football performance. The performance indicators comprised: possession, shots on target, crosses, pass accuracy, aerial duels won (%), and, number of tackles. Ethical approval was granted by Bournemouth University (ref: 53858).

Time-lag analysis

The time-lag analysis process was carried out using the methodology previously outlined [3]. Firstly, performance indicators were standardised to allow for comparison across metrics using z-scores. Secondly, the Euclidean distance was calculated to quantify the dissimularity between increasing lags. In team sports, a 'lag' denotes the number of games between each pairwise comparison. For example, a lag of one would compare consecutive games (e.g. game 1 with game 2, game 2 with game 3, and so forth). A lag of two would compare games with a one-game gap (e.g. game 1 with game 3, game 2 with game 4), while a lag of three would extend the gap to two games (e.g., game 1 with game 4). This pattern would continue until the maximum lag is achieved (e.g. game 1 with game game 38). Thirdly, the Euclidean distance and square root of the time-lag were plotted for each sample size. Time-lag transformation aimed to reduce the bias of larger time lags where fewer data points are present. Linear and non-linear regression techniques summarised the trend in the temporal dataset. Each regression was visually and statistically interpreted using the theoretical patterns associated with time-lag analysis (see Figure 1). The coefficient of variation and probability was used to evaluate the goodness-of-fit and significance. If the regression was not statistically significant ($p > 0.05$), it implied the profile exhibited fluctuations or stochastic variation over time. When a profile maintained a coefficient of variation of less than 0.001 for five consecutive instances, it was interpreted that no meaningful trends or patterns in team behavior could be observed for that window. Data analysis and visualisation was conducted in R Studio (version 2021.09.01) using the Vegan and GGPlot2 packages.

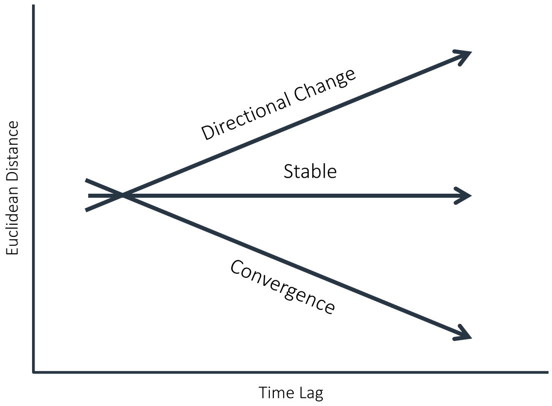


Figure 1. Theoretical patterns associated with a time–lag regression analysis (adapted from [2]).

Results

Figure 2 shows the time-lag regression of the six performance indicators across increasing sample sizes. Results show significant trends were consistently observed within sample sizes between 8 and 30 games (range: $R^2 = 0.01 - 0.51$; $p < 0.047$). In contrast, shots on target exhibited a stochastic profile in the majority of samples ($p > 0.057$).

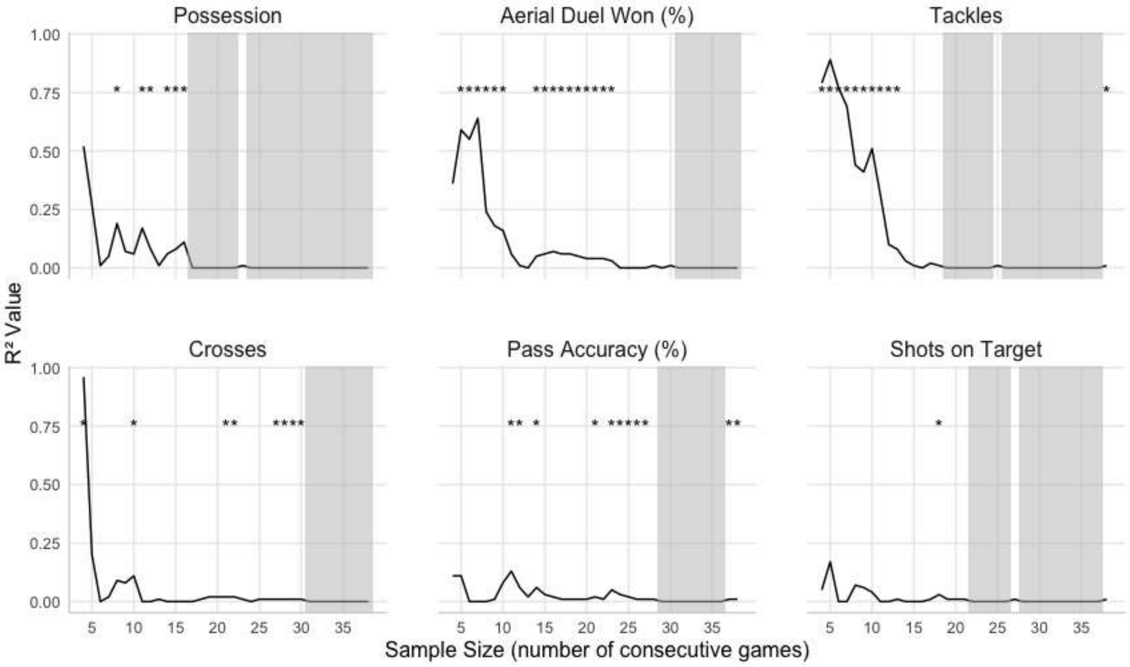


Figure 2. Time lag regression for typical performance indicators in football at increasing sample sizes (between 3 to 38 games) during the 2022-23 Premier League season.

*Denotes $p < 0.05$ and grey shading indicates a R^2 value < 0.001 for five consecutive games.

The number of tackles, possession, and aerial duels won showed directional change between 4 and 13 games (R^2 range 0.08 - 0.89; $p < 0.021$). In constrast, directional change was only identified in larger sample sizes for pass accuracy (between 21 and 27 games) and crosses (between 27 and 30 games). While the profile of the number of tackles remained linear, possession and aerial duels changed to a non-linear pattern between 14 and 16 games and, 14 and 23 games respectively. The non-linear trend identified was an inverted U.

Tackles was chosen as an exemplar to illustrate the implications of sample size on the rate of inter-game variability. Figure 3 shows the changing rate of inter-game variability as the sample size increased from 6 to 38 games. Directional change was observed between 6 and 13 games indicated by a significant regression ($p < 0.011$) and a positive gradient (coefficient range = 0.33 - 1.60). A decrease in the rate of directional change was observed, evidenced by the deminishing gradient of the linear regression (coefficent = 1.03 to 0.33). Furthermore, there was an increase in stochastic variation between 14 and 18 games, as indicated by a decrease in the coefficient of variation (R^2 = 0.08 - 0.01). No visual or statistical difference was inferred above a sample size of 19 games as the coefficient of variation maintained a value below 0.001 over a window of five games (see Figure 2).

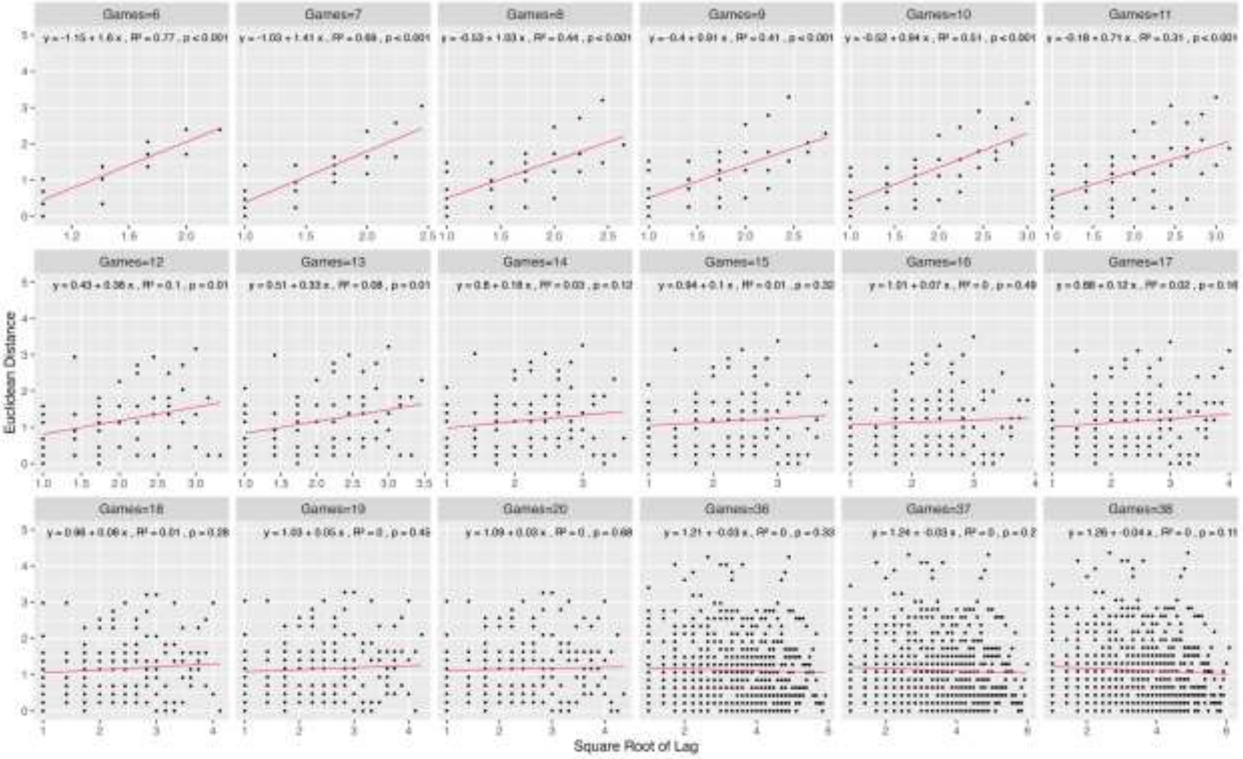


Figure 3. Time-lag analysis for number of tackles per game at increasing sample sizes between 6 and 38 games for a Premier League team during the 2022-23 season.

Discussion

The findings of the current study demonstrate how time-lag analysis, which has been transfered from ecology [3], can be a useful tool to understand the temporal patterns in sports behavior over short time intervals. Directional change was more evident in smaller samples (< 13 games), suggesting systematic changes in behavior may be more evident in shorter time periods. When analysing the rate and patterns of inter-game variability for crosses and pass accuracy, systematic change was only observed in larger sample sizes, occurring after 27 and 21 games respectively. These findings offer valuable insights for coaches, emphasising that certain aspects of performance may require more time for behavioral changes to take effect.

The use of time-lag analysis was also able to identify more complex non-linear trends for possession and aerial duels. The inverted U pattern found in this study has also been observed in Ecology when investigating the potential impact of an experimental intervention on an ecosystem [3]. These findings provide evidence that a divergence-convergence dynamic may also exist in team sports. Divergence represents a change in performance, whereas convergence indicates that performance may be returning to a more consistent pattern. In coaching, recognising these divergence-convergence patterns could offer insight into when to refine technical or tactical strategies, enabling effective periodisation.

A stochastic profile, as demonstrated in shots on target in the current study, is characterised by a random or unpredictable pattern. Opposition strength [4] and match location [4, 5] have been shown to impact a team's sports performance, which may account for the higher variation in aspects of performance seen in the current study. Further investigation should consider: (1) establishing good practice when identifying linear and non-linear trends, (2) methods to normalise the rate of change across different time intervals, (3) optimisation of the output visualisation for a coaching context, (4) explore how situational factors may influence rate and patterns of inter-game variability in team sports, and (5) examine inter-game variability through the lens of ecological dynamics to ascertain possible (dys)function of inter-game variability.

Conclusion

Findings show directional change occurred more frequently in smaller sample sizes (< 13 games). In contrast, larger sample sizes (> 25 games) were commonly stochastic due to the higher variance in sports performance. In conclusion, time-lag analysis has been shown to be a useful tool when exploring inter-game variability and understanding rate and patterns of inter-game variability in samples between 6 and 25 games.

References

1. Buttfield, A., & Ball, K. (2020). The practical application of a method of analysing the variability of within-step accelerations collected via athlete tracking devices. *Journal of Sports Sciences*, **38(3)**, 343-350.

2. Cowin, J., Nimphius, S., Fell, J., Culhane, P., & Schmidt, M. (2022). A proposed framework to describe movement variability within sporting tasks: A scoping review. *Sports Medicine-Open*, **8(1)**, 85.

3. Collins, S. L., Micheli, F., & Hartt, L. (2000). A method to determine rates and patterns of variability in ecological communities. *Oikos,* **91(2)**, 285-293.

4. Liu, H., Gómez, M. A., Gonçalves, B., & Sampaio, J. (2016). Technical performance and match-to-match variation in elite football teams. *Journal of Sports Sciences,* **34(6),** 509-518.

5. Carmichael, F., & Thomas, D. (2005) Home-field effect and team performance. *Journal of Sports Economics*. **6,** 263-281.

Movement: a Python Toolbox for Analysing Pose Tracking Data

N. Sirmpilatze[1,2], C.H. Lo[2], B.D. Peri[3], D. Sharma[4], S. Miñano[1], S. Keshavarzi[3], A.L. Tyson[1,2].

1 Sainsbury Wellcome Centre, University College London, London, United Kingdom. n.sirmpilatze@ucl.ac.uk

2 Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom.

3 Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom.

4 Department of Chemistry, University College London, London, United Kingdom.

Abstract

While pose estimation tools like *DeepLabCut* and *SLEAP* are increasingly used to extract body part positions from videos of animal behaviour, processing the resulting pose tracks lacks a standardised, user-friendly approach. To address this, we are developing *movement*, a Python package that offers a consistent and modular interface for analysing pose tracks. By leveraging the scientific Python ecosystem, especially *xarray* and *napari*, *movement* facilitates tasks such as data cleaning, visualisation, and motion quantification.

Background

In recent years, the study of animal behaviour has profited from markerless pose estimation tools, like *DeepLabCut* [1], *SLEAP* [2] and *LightningPose* [3]. Relying on deep neural networks, these tools detect user-defined keypoints in video frames, thereby capturing the positions of specific body parts (poses) in 2D or 3D. This process results in the generation of pose tracks, which are sequential collections of poses over time, with each track corresponding to an individual in the video [4].

The extraction of pose tracks is often just the beginning of the analysis. Researchers use these tracks to investigate various aspects of animal behaviour, such as kinematics and spatial navigation [4]. Typically, these analyses involve custom, project-specific scripts that may be hard to reuse across different projects and are rarely maintained after the project's conclusion. Existing general-purpose, open-source tools for analysing pose tracks are limited: *DLC2Kinematics* [5] is specific to *DeepLabCut* outputs while *PyRat* [6] seems to have ceased active development. The lack of a graphical user interface (GUI) in these tools further hinders their accessibility and broader adoption. While several mature Python packages adeptly handle mobility analysis for geospatial trajectories [7], [8], they lack the necessary functionalities for many studies of animal behaviour.

Aim

In response to these challenges, we identified the need for a versatile and user-friendly tool that is compatible with a range of pose estimation frameworks and supports interactive data exploration and analysis. To meet this need, we are developing *movement*, a free and open-source Python package that we hope will streamline measuring animal behaviour.

Key features

A unifying interface for pose tracks

A fundamental aspect of *movement* is its ability to process pose tracks from various sources with a consistent interface. This is achieved by designing a unifying data structure with *xarray*, a prominent scientific Python library known for handling labelled multi-dimensional arrays [9]. In *movement*, pose tracks are structured as arrays with four dimensions—time, space, individuals, and keypoints—each annotated with descriptive labels (see example in Figure 1). This approach can accommodate data from a variety of pose estimation tools, whether in 2D or 3D, tracking single or multiple individuals. Moreover, using a mature standard such as *xarray* brings numerous advantages, including a suite of built-in functionalities for efficient data handling. Users can easily perform operations along specific dimensions and select data points by label, e.g. apply smoothing over time for a particular individual. Additionally, *xarray* boosts performance by enabling vectorization across dimensions and supports parallel processing through the *dask* library [10].

At present, *movement* offers input/output functionalities for *DeepLabCut*, *SLEAP*, and *LightningPose*, and we plan to expand this support to other widely adopted tools in the community. The aim is to achieve interoperability with all leading tools in animal pose estimation and behaviour classification, facilitating format conversions between them. While *movement* is not designed for behaviour classification or action segmentation, it may extract features useful for these tasks.

With a standardized *xarray* data structure for pose tracks in place, we can focus on streamlining their processing and analysis. We are actively developing several key processing steps for upcoming releases of *movement*. One common need is to remove inaccurate or implausible keypoint predictions, such as those that come with low confidence values or violate the smoothness of physical motion. This step often needs to be followed by interpolation of missing or dropped values. Other essential processing steps we're focusing on include smoothing, resampling in space or time, and coordinate system transformations, like shifting to an egocentric frame of reference.

The processed pose tracks are instrumental for calculating various kinematic variables frequently reported in behavioural studies, including velocity, acceleration, and head direction. Our initial focus for *movement* is to integrate efficient and validated methods for these commonly used variables, with plans to expand our offerings based on community feedback. In the long term, we envision *movement* evolving to include specialised modules for applications like pupillometry and gait analysis, further broadening its scope and utility in behavioural research.

Ease of use

Another key aspect of movement is its user-friendliness, with a design that accommodates researchers with varying coding skills and computational resources. Available through the Python Package Index (PyPI), movement is cross-platform and lightweight, optimised for use on standard laptops without requiring a dedicated GPU. We strive to avoid dependencies that may compromise this goal. The package is supported by comprehensive documentation on our website, to ensure accessibility and usability.

To further enhance usability, we are integrating a GUI using *napari*, a popular multi-dimensional image viewer for Python [11]. This integration leverages *napari*'s existing layer types to represent the data—i.e. 'Tracks' for pose tracks, 'Shapes' for regions of interest (ROIs) and 'Image' for video frames. In this way, pose tracks can be viewed and analysed in their spatial context, such as the arena in which the animal is moving and the location of objects within it. This capacity, combined with the ability to interactively select subsets of pose tracks along any dimension, should enable users to easily identify data anomalies as well as interesting patterns in the animal's behaviour. We thus hope to facilitate both quality control and hypothesis formulation through the *napari* interface.

Robustness and openness

Movement is being developed by a team that includes full-time research software engineers who are committed to maintaining the project long-term. We test all functionality, targeting 100% test coverage, and strive to ensure scientific accuracy and reproducibility by validating results against ground truth datasets and simulations.

Our development process is fully transparent and hosted on GitHub, fostering community engagement, scrutiny and contributions. Our website includes statements on *movement*'s mission, scope, and roadmap, as well as detailed contribution guidelines, all of which are living documents and are being updated as the project evolves. The source code is free to redistribute and modify under the BSD 3-Clause licence.

Call to action

This abstract does not represent a full list of current and potential capabilities of *movement*, which is still in its early stages of development. We are actively seeking feedback from the community to shape the project's future direction. We invite you to join our Zulip chat forum to share your comments and suggestions for improvement. Your input is crucial and will be a key consideration in our planning for future updates and releases.

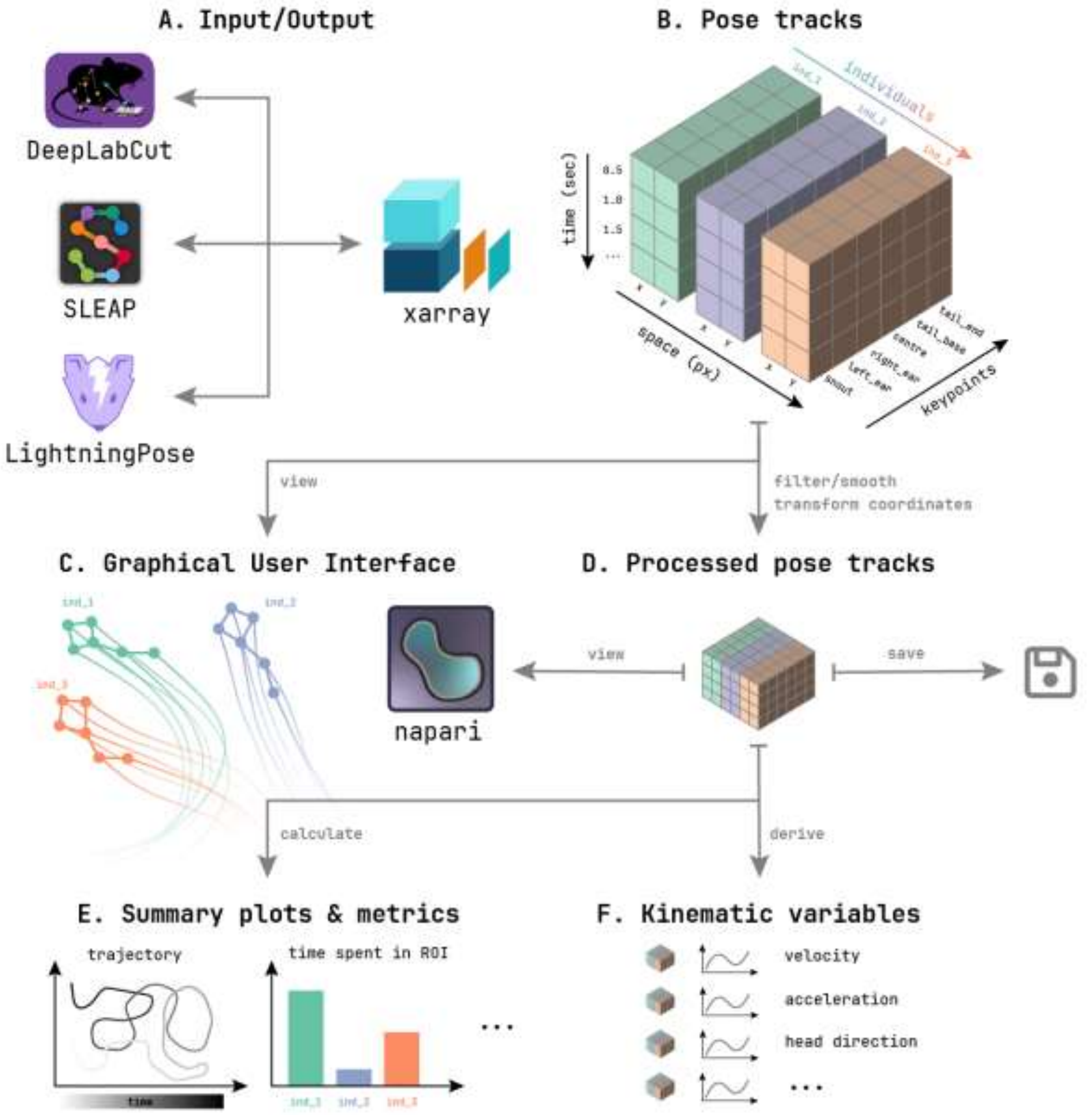


Figure 15. An overview of *movement*'s design and functionality. A. Pose tracks from various frameworks, including *DeepLabCut*, *SLEAP*, and *LightningPose*, are imported into an *xarray* data array. B. This example illustrates pose tracks of three individuals, each with six keypoints, tracked in 2D (x, y) at a rate of 2 frames per second. C. Our graphical user interface (GUI), developed using *napari*, enables interactive visualisation of pose tracks. D. Pose tracks can undergo various processing steps like filtering, smoothing, and coordinate transformations, preparing them for downstream analysis. E. Movements are quantified within the context of the animal's environment, with results presented as summary plots and metrics. F. Kinematic variables such as velocity, acceleration and head direction are extracted from pose tracks and stored in dedicated *xarray* objects for further plotting and analysis.

Acknowledgements

We wish to thank Loukia Katsouri, Laura Schwarz, Sanna Titus, Dimokratis Karamanlis and Shailaja Akella for providing sample data for testing *movement.* We have sought assistance from OpenAI's ChatGPT in editing this document for language and phrasing.

References

1. Mathis, A. et al (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* **21**, 1281–1289. doi: 10.1038/s41593-018-0209-y

2. Pereira, T. D. et al (2022). SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods* **19**, 486–495. doi: 10.1038/s41592-022-01426-1.

3. Biderman, D. et al (2023). Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling, and cloud-native open-source tools. *bioRxiv* [Preprint]. doi: 10.1101/2023.04.28.538703.

4. Pereira, T. D., Shaevitz, J. W. & Murthy, M. (2020). Quantifying behavior to understand the brain. *Nature Neuroscience* **23**, 1537–1549.

5. Mathis, M. et al (2020). DLC2Kinematics: a post-deeplabcut module for kinematic analysis. *Zenodo*. doi: 10.5281/zenodo.6669073.

6. De Almeida, T. F., Spinelli, B. G., Hypolito Lima, R., Gonzalez, M. C. & Rodrigues, A. C. (2022). PyRAT: An Open-Source Python Library for Animal Behavior Analysis. *Frontiers in Neuroscience* **16**. doi: 10.3389/fnins.2022.779106.

7. Graser, A. (2019). MovingPandas: Efficient Structures for Movement Data in Python. *GI_Forum* **7**, 54–68. doi: 10.1553/giscience2019_01_s54

8. Pappalardo, L., Simini, F., Barlacchi, G. & Pellungrini, R. (2022). scikit-mobility: A Python Library for the Analysis, Generation, and Risk Assessment of Mobility Data. *Journal of Statistical Software* **103(4)**, 1–38. doi: 10.18637/jss.v103.i04.

9. Hoyer, S. & Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software* **5(1)**, 10. doi: 10.5334/jors.148.

10. Rocklin, M. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. *Python in Science Conference (Austin, Texas, 2015)*, 126–132. doi: 10.25080/Majora-7b98e3ed-013.

11. Ahlers, J. et al (2023). napari: a multi-dimensional image viewer for Python. *Zenodo*. doi: 10.5281/zenodo.3555620.

Unsupervised Machine Learning Analysis of Freezing Reveals Postures Correlating with Task Performance and Fear Level

Nancy J. Smith-Vickery[1,2], Daniel Weatherill[3], Andrew Wikenheiser[2], and Michael S. Fanselow[1,2,4]

Staglin Center for Brain & Behavioral Health[1], Department of Psychology[2], Department of Radiological Sciences[3], Department of Psychiatry and Behavioral Sciences[4], University of California, Los Angeles, USA

Abstract

Women are more prone to anxiety disorders but are significantly underrepresented in pre-clinical research. Recent progress in automated behavioral assessment, specifically through animal pose estimation, enables efficient analysis of datasets using unsupervised machine learning (UML) that are being adopted in many behavioral neuroscience applications. We employed a cluster-analysis-based UML algorithm to process data from markerless pose estimation, tracking freezing postures in a series of shock avoidance and fear conditioning experiments in rats. This approach enabled us to study how female and male rats responded to fear stimuli when modifying the intensity and density of foot shock presentations. We tested the hypothesis that sex differences in avoidance task performance at a low shock intensity are due to differences in fear level, reflected in differences in freezing postures. The UML algorithm categorized freezing postures into eight subtypes, composed of a mixture of freezing bouts from both female and male rats. Stratified K-fold cross-validation confirmed the consistency of the eight identified clusters of freezing postures. Notably, when the shock intensity was low, female rats punished for freezing exhibited lower performance on the task compared to their male counterparts, who were similarly penalized for freezing. Controlling for size differences between animals, the freezing postures of these females were enriched in two specific clusters. The Predatory Imminence Continuum (PIC) theory relates threat proximity to defensive behavior and intensity. The PIC categorizes defensive responses based on increasing threat into pre-encounter, post-encounter, and circa-strike modes. Freezing in response to a predator situates an animal in the post-encounter mode, indicating fear. However, merely observing freezing as a binary response does not reveal the extent of fear experienced by the animal. To gain insight into what describes the different postural clusters, we extracted positional, orientational, and postural features from the raw keypoint data and, first, trained a K-nearest neighbors (KNN) classifier to predict cluster identity. We then used permutation feature importance analysis, which revealed that different mixtures of postural information were important for classifying different clusters. Studying the behaviors before and after each freezing bout may help gauge the animal's fear level. For instance, grooming before freezing may suggest low fear, while being startled before freezing indicates panic-like state, based on PIC categorization. Therefore, where an animal lies within the post-encounter mode of the PIC and its posture during a freezing bout should be a function of its PIC position prior to freezing as well as how quickly it shifts into a post-encounter state. We suggest the different clusters of freezing postures are correlates of position along a continuum of fear. Indeed, immediately prior to freezing, we found unique patterns of behaviors falling into the different PIC modes mentioned above amongst the different clusters. We also found the rate of decrease in these behaviors prior to the freezing bouts differed between clusters. Importantly, up until, now freezing has been seen as an all-or-none fear metric. Our study is the first to utilize different freezing postures as a graded metric of fear magnitude. Using markerless pose estimation and cluster-analysis-based UML offers unique insights into the behavioral expression of fear in species that communicate in ways not easily understood by humans. Overall, our findings provide valuable insights into rodent behavior and may eventually aid the diagnosis and treatment of anxiety-related disorders.

Introduction

Females are twice as likely as males to be diagnosed with anxiety-related disorders [1], yet they are underrepresented in preclinical research [2]. Recent advancements in automated behavioral assessment, mainly through animal pose estimation [3-4], allow efficient analysis of datasets using unsupervised clustering approaches [5]. We set out to test the hypothesis that differing shock intensities (*i.e.*, threat levels) produce varying levels of fear, which in turn differentially affects behavioral flexibility needed to override innate defense responses to learn an instrumental avoidance response. To do so, we designed an avoidance model that manipulated shock intensity (i.e., nature of the threat) to look for differences in rats' ability to learn an avoidance response and correlate these to graded differences in freezing postures to possible differences in levels of fear. Employing a customized unsupervised machine learning (UML) algorithm [6] and cluster analysis [7], we processed coordinate data [8] from markerless pose estimation and human observed categorical behavior [9-10] to track freezing postures [11] in fear avoidance learning experiments [12]. Our study explored how female and male rats responded to fear stimuli under different shock presentations, revealing nuanced differences detected through cluster-analysis-based UML. Our study is the first to utilize different freezing postures as a graded metric of fear magnitude. Using markerless pose estimation and cluster-analysis-based UML offers unique insights into the behavioral expression of fear in species that communicate in ways not easily understood by humans. Overall, our findings provide valuable insights into rodent behavior and may eventually aid the diagnosis and treatment of anxiety-related disorders.

Method

Behavioral scoring

Multiple coders independently scored pre-recorded videos logging the appropriate behavioral code (Figure 1). Using the VLC media player, with the jump-to-time extension shown in Figure 1b allowed for precise and efficient frame identification. This process was repeated for over 1.5 million frames and were hand-scored by three independent raters. Importantly, coders remained blind to the contingency and shock level of the animals. Figure 1a visualizes the scores assigned by coders (highlighted in yellow, blue, and pink) for analyzing inter-rater-reliability, with discrepancies resolved through discussions [9].

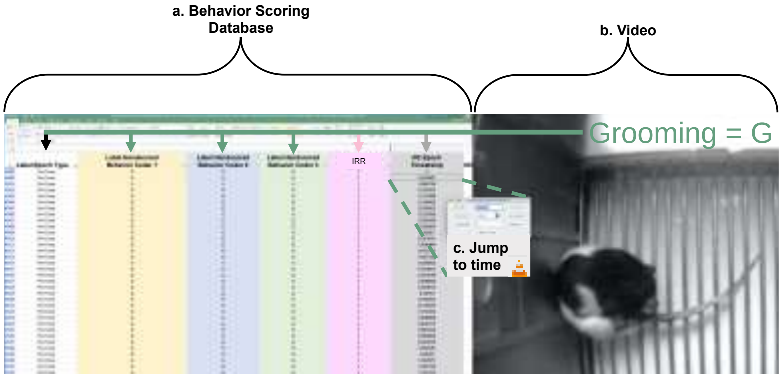


Figure 1. Behavioral coding diagramed to show behavioral scoring database (a) to hand score behavior observed by watching videos of rats. Observers used VLC and the jump to time tool to score frame by frames behavior (b).

Cluster Analysis

Videos and data exported from EthoVision XT [13] were processed and fed into DeepLabCut [3] for frame extraction using a custom script (Figure 2). Formatted and cropped videos had each frame manually processed for key body point placement, enabling posture capture, artificial neural network (ANN) training, model evaluation, and refinement for accurate pose estimation (Figure 2b). Data was further processed in BehaviorDepot [14] to smooth additional key points and interpolate missing data Figure 2c. The processed data was then inputted into a custom Spyder script that translated frames on the Cartesian plane, rotated frames for consistent posture comparison, and scaled all frames to the size of a reference rat to control for sexual and body dimorphism as shown in Figure 2d. Once scaled, the workflow concluded with dimensionality reduction and clustering analysis (Figure 2d).

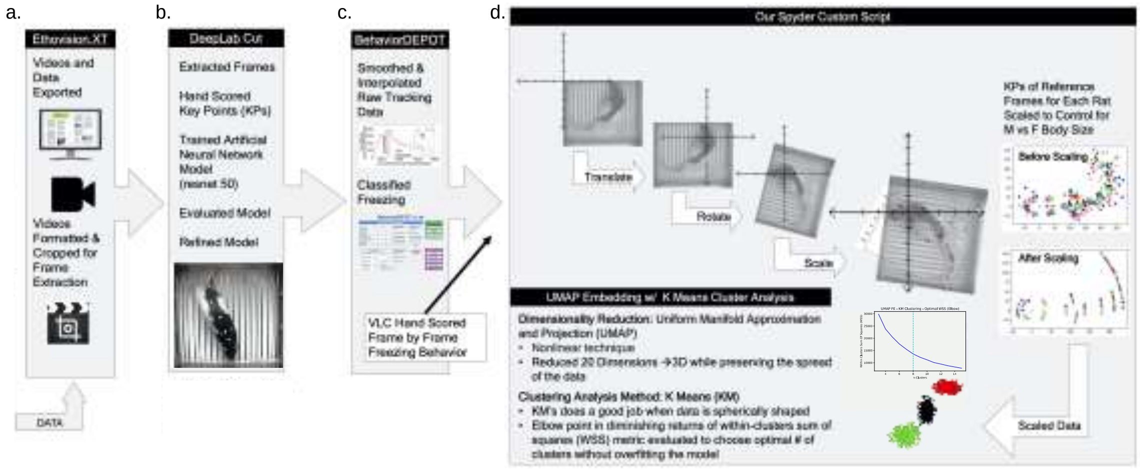

a.
b.
c.
d.
Ethovision XT
Videos and Data Exported
Videos Formatted & Cropped for Frame Extraction
DATA
DeepLab Cut
Extracted Frames
Hand Scored Key Points (KPs)
Trained Artificial Neural Network Model (resnet 50)
Evaluated Model
Refined Model
BehaviorDEPOT
Smoothed & Interpolated Raw Tracking Data
Classified Freezing
VLC Hand Scored Frame by Frame Freezing Behavior
Our Spyder Custom Script
Translate
Rotate
Scale
KPs of Reference Frames for Each Rat Scaled to Control for M vs F Body Size
Before Scaling
After Scaling
Scaled Data
UMAP Embedding w/ K Means Cluster Analysis
Dimensionality Reduction: Uniform Manifold Approximation and Projection (UMAP)
Nonlinear technique
Reduced 20 Dimensions → 3D while preserving the spread of the data
Clustering Analysis Method: K Means (KM)
KM's does a good job when data is spherically shaped
Elbow point in diminishing returns of within-clusters sum of squares (WSS) metric evaluated to choose optimal # of clusters without overfitting the model
UMAP FE - KM Clustering - Optimal WSS (Elbow)
Within Clusters Sum Of Squares (WSS)
n Clusters

Figure 2. Cluster-analysis-based unsupervised machine learning (UML) algorithm workflow. The workflow involves assessing fear responses in male and female rats through two avoidance tasks, Avoid-if-Freeze and Punished-if-Freeze, with varying shock intensities as consequences if the avoidance response is not met. The idea is that it is easier to freeze when afraid and much more challenging to inhibit freezing when afraid. (a) Prepossessing of EthoVision videos and data from the last training session was cropped for frame extraction for markerless pose estimation by (b) DeepLab Cut, where raw x- & y- coordinate data was used to train an artificial neural network on pose estimation before refining model through (c) smoothing and interpolation of raw data using BehaviorDEPOT. (d) Interpolated raw key point data was scaled to control for body size differences, and a cluster-analysis-based unsupervised machine learning (UML) algorithm was applied to identify freezing patterns to analyze the level of fear in rats.

Results and Discussion

Our study utilized a cluster-analysis-based UML algorithm to analyze data from markerless pose estimation, specifically tracking freezing postures in rats across shock avoidance and fear conditioning experiments. We aimed to investigate how both female and male rats respond to fear stimuli, particularly when varying the intensity and density of foot shock presentations. Our hypothesis centered on sex differences in avoidance task performance at low shock intensity linked to varying fear levels, reflected in freezing postures. The UML algorithm categorized freezing postures into eight distinct subtypes, which were confirmed consistently through cross-validation (Figure 3 a-b). Notably, when shock intensity was low, female rats penalized for freezing showed decreased task performance compared to similarly penalized males. Controlling for size differences (Figure 2d), these females exhibited enrichment in two specific freezing posture clusters (Figure 3 a-b).

We integrated the Predatory Imminence Continuum (PIC) theory, a well-established framework that associates threat proximity with defensive behavior, into our study. Fear plays a crucial role in guiding organisms towards adaptive behaviors according to the level of threat they encounter [12]. By restricting choices, fear aids survival by directing behavior towards strategies likely to avoid predation [12, 15]. However, fear's downside lies in its potential to induce behavioral inflexibility, constraining actions to rigid patterns even in situations with low or no threat, thereby hindering flexible decision-making [12, 15-17]. This inflexibility characterizes anxiety disorders, where inappropriate fear and anxiety reactions prevail [17]. This integration helped us to better understand freezing responses. We revealed that freezing behavior is not binary but falls along a continuum of fear, with different postural clusters representing varying fear levels. Our analysis included extracting positional, orientational, and postural features to train a K-nearest neighbors (KNN) classifier, which highlighted the importance of different postural information for cluster classification (Figure 4 a-i). Very preliminary analysis, not pictured here, on behaviors before and after freezing bouts were studied to gauge fear levels, with grooming suggesting low fear and startle responses indicating panic-like states.

We propose a significant shift in understanding freezing postures, suggesting that they correspond to positions along the PIC continuum (Figure 5). Unique behavioral patterns were observed before freezing across different PIC modes and clusters. Importantly, our study disrupts the traditional view of freezing as an all-or-none fear metric, instead presenting freezing postures as a graded measure of fear magnitude. This finding has profound implications for the field of psychology and animal behavior.

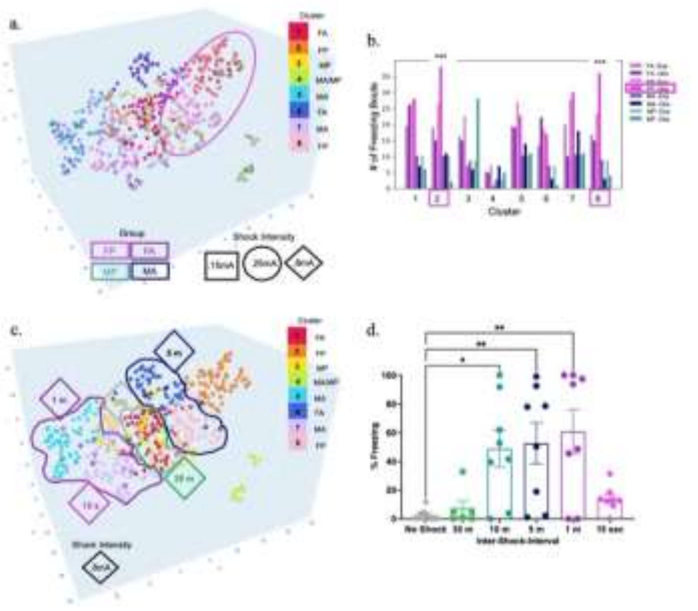


Figure 3. Clustering analysis of freezing rats: UMAP & K Means identify 8 distinct clusters. (a) UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction and K Means clustering on 447 frames of rats freezing revealed 8 distinct clusters within the dataset. In an avoidance task, male and female rats were assigned to groups that needed to freeze to avoid foot shock or refrain from freezing to avoid foot shock. (b) Enrichment of female rats' punishment for freezing (FP, pink) were enriched in clusters 2 and 8. Other groups: males shocked for freezing (MP, green), females required to freeze to avoid punishment (FA, purple), and males required to freeze to avoid punishment (MA, navy blue). Our UML algorithm also examined raw key point data from 48 freezing frames obtained from a separate experiment with different timings for shock administration (e.g., every 15 seconds, 1 minute, 5 minutes, 10 minutes, 30 minutes, and no shock control. (c) Those freezing frames were overlayed onto the eight existing clusters to visually represent the delineation of potential fear dimensions of the eight clusters. (d) The freezing percentage varied significantly compared to the control group with no shock, showing that the rats' fear levels changed based on the timing between each shock delivery. As the time between shocks decreased, indicating a more imminent threat, the animals' fear level increased, leading to a higher percentage of freezing. In simple terms, when animals feel more afraid, they freeze more. ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$, ns p values > .05. Error bars represent ±SEM.

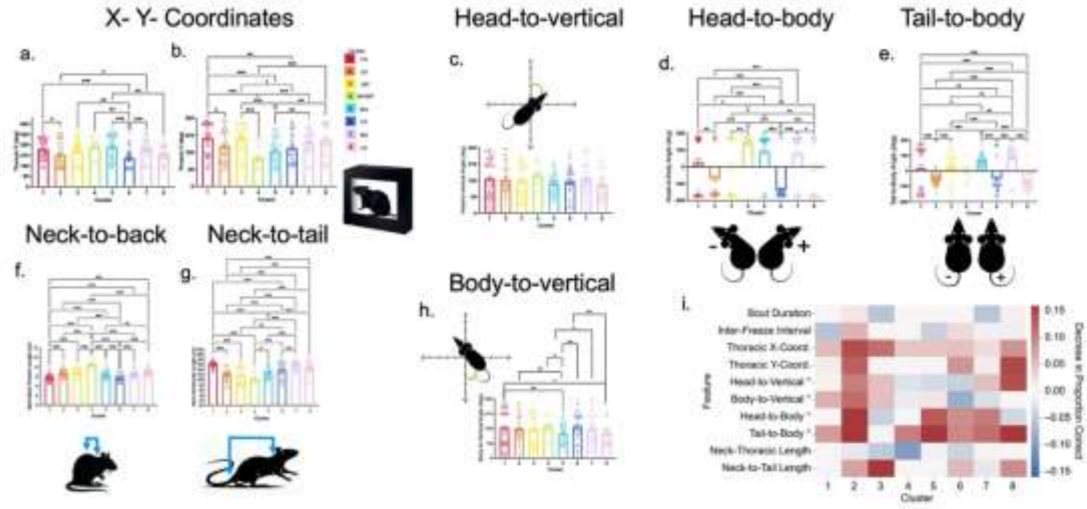


Figure 4. Postural diversity in cluster classification. (a-h) No one feature is completely associated with any one cluster. (i) A heatmap of postural diversity in cluster classification (red = important to cluster). Behavioral features include bout duration and inter-freeze interval; positional features include thoracic X-coord. and thoracic Y-coord.; orientational features include head-to-vertical ° and body-to-vertical °; postural features include head-to-body °, tail-to-body °, neck-thoracic length, and neck-to-tail length.

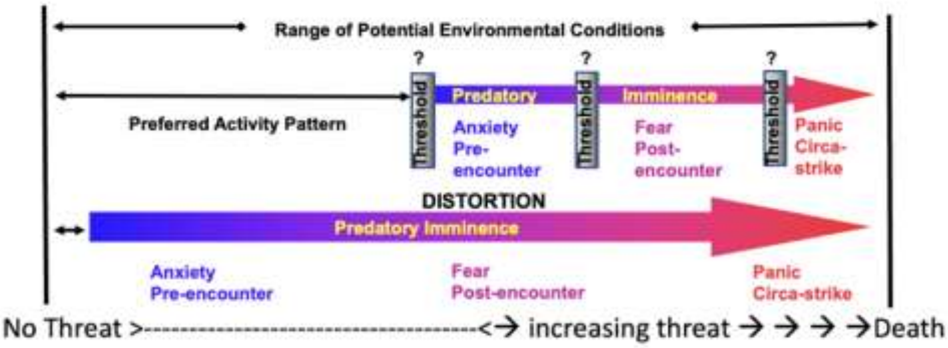


Figure 5. Predatory Imminence Continuum (PIC). Anxiety disorders may result from a distortion of the PIC, where defensive behaviors intrude into daily activities. Thresholds denote boundaries between behavioral modes.

References

1. Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*, 62(6), 617-627. https://doi.org/10.1001/archpsyc.62.6.617

2. Dalla, C., & Shors, T. J. (2009). Sex differences in learning processes of classical and operant conditioning. *Physiol Behav,* 97(2), 229-238. https://doi.org/10.1016/j.physbeh.2009.02.035

3. Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9), 1281-1289. https://doi.org/10.1038/s41593-018-0209-y

4. Mathis, M. W., & Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Curr Opin Neurobiol*, 60, 1-11. https://doi.org/10.1016/j.conb.2019.10.008

5. McVey, C., Hsieh, F., Manriquez, D., Pinedo, P., & Horback, K. (2023). Invited Review: Applications of unsupervised machine learning in livestock behavior: Case studies in recovering unanticipated behavioral patterns from precision livestock farming data streams. *Applied Animal Science*, 39, 99-116. https://doi.org/10.15232/aas.2022-02335

6. Hozumi, Y., Wang, R., Yin, C., & Wei, G.-W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Computers in Biology and Medicine*, 131, 104264. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104264

7. Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, 25, 1104-1109. https://doi.org/https://doi.org/10.1016/j.phpro.2012.03.206

8. Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6), 90-95.

9. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. https://www.ncbi.nlm.nih.gov/pubmed/843571

10. Holly, K. S., Orndorff, C. O., & Murray, T. A. (2016). MATSAP: An automated analysis of stretch-attend posture in rodent behavioral experiments. *Scientific Reports*, 6(1), 31286. https://doi.org/10.1038/srep31286

11. Fanselow, M. S. (1980). Conditioned and unconditional components of post-shock freezing. *Pavlov J Biol Sci*, 15(4), 177-182. https://doi.org/10.1007/BF03001163

12. Bolles, R. C., & Riley, A. L. (1973). Freezing as an Avoidance Response. *Learning and Motivation*, 4, 268-275

13. Noldus, L. P., Spink, A. J., & Tegelenbosch, R. A. (2001). EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behav Res Methods Instrum Comput*, 33(3), 398-414. https://doi.org/10.3758/bf03195394

14. Gabriel, C. J., Zeidler, Z., Jin, B., Guo, C., Goodpaster, C. M., Kashay, A. Q., . . . DeNardo, L. A. (2022). BehaviorDEPOT is a simple, flexible tool for automated behavioral detection based on markerless pose tracking. *Elife,* 11. https://elifesciences.org/articles/74314

15. Fanselow, M. S., & Lester, L. S. (1988). A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior. In R. C. Bolles & M. D. Beecher (Eds.), *Evolution and learning* (pp. 185–212). Lawrence Erlbaum Associates, Inc.

16. Perusini, J. N., & Fanselow, M. S. (2015). Neurobehavioral perspectives on the distinction between fear and anxiety. *Learn Mem*. 22(9):417-25. doi: 10.1101/lm.039180.115. PMID: 26286652; PMCID: PMC4561408

17. Fanselow, M. S. (2023) Negative valence systems: sustained threat and the predatory imminence continuum. *Emerg Top Life Sci*. 6(5):467-477. doi: 10.1042/ETLS20220003. PMID: 36286244; PMCID: PMC9788377.

Failed Negative Geotaxis in *Drosophila melanogaster*: Measurement Techniques and Implications

J. Vasu[1], K. Vu[1], R. Hartman[1]

Hartman Behavioral Neuroscience Lab, Loma Linda University, School of Behavioral Health[1], United States of America

Abstract

Drosophila Melanogaster, the common fruit fly, is commonly used in behavioral neuroscience research. Negative Geotaxis is a natural inclination to climb, commonly observed in *Drosophila Melanogaster*. While negative geotaxis has been extensively studied as models for aging, neurodegeneration, and drug effects, researchers have failed to observe errors in this behavior. There is currently no research on Failed Negative Geotaxis (FNG), when flies fail to exhibit their natural climbing behavior and fall. As the common factors that can contribute to locomotion deficits, investigating how FNG relates to age, sex, and climbing speed on a standardized RING assay provides a novel behavioral measurement without adaptation difficulties. We have developed a method for recording these errors, FNG, using a traditional Rapid Iterative Negative Geotaxis (RING) assay [1]. FNG changes were significantly associated with sex, age, and climbing speed in a pilot Drosophila sample. Our study aims to introduce FNG as a new dimension in *Drosophila* behavior analysis, capturing a specified deviation from typical negative geotaxis. This research not only contributes novel mechanisms in assessing locomotion deficits, but does so with a common place behaivoral neuroscience assay.

Methods

Drosophila (N = 170) were housed in a temperature-controlled (23.33°C) 12-hour day/night cycle box, segregated by age (days post-eclosure) and sex. They were placed in plastic vials, each containing ten individuals, with markings at one centimeter intervals. After forcing the flies to the vial bottom, a one-minute recording captured their climbing behavior. A diagram for the RING assay can be seen in Figure 1. The height climbed within the first four seconds was measured for each fly, and the averages were calculated. Average max height climbed for each vial was recorded at the one minute marker. Additionally, instances of Failed Negative Geotaxis (FNG) were recorded if a fly fell or failed to climb during the one-minute assessment. This procedure was replicated three times, and the resulting averages for speed, max height climbed, and FNG were documented.

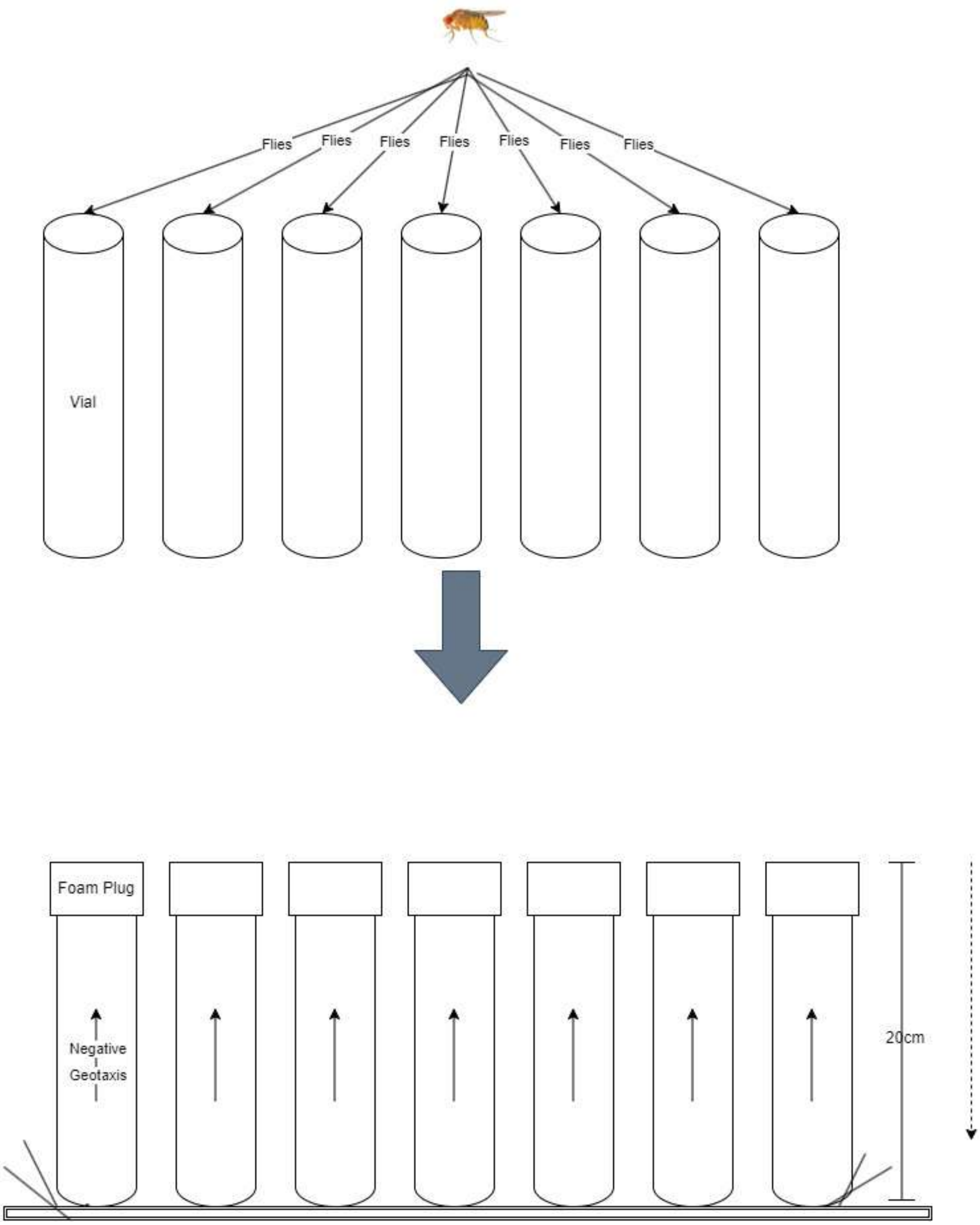
Results

A multiple linear regression was run on average FNG, age, max height climbed, and the interraction effect of average height climbed and age. Analysis revealed that as age in *drosophila* increased by one day, the instances of FNG significantly decreased by 0.28, when max climbing height was set equal to zero, 95% CI [-0.55, -0.01], p = .04. The interaction between average height climbed and FNG was contingent upon age, indicatingf a modertating effect of age on the association between max height climbed and instances of FNG, p = .05.

Ethical Statement

The Hartman Behavioral Neuroscience Lab is ethically governed by the Institutional Animal Care and Use Committee (IACUC). Even though *drosophila* are not under the oversight IACUC, all procedures adhered to IACUC polices.

Figure 1- Rapid Iterative Negative Geotaxis (RING) Diagram



Note. Flies are placed in vials and tapped lightly onto a surface to aggregate them to the bottom of the vials. Flies begin to climb shortly after.

References

1. Dilliane, C. C., Suelen, F., Jaqueline, V., & Welligton, L. B. (2017). Valeriana officinalis and melatonin: Evaluation of the effects in Drosophila melanogaster rapid iterative negative geotaxis (RING) test. *Journal of Medicinal Plants Research*, *11*(44), 703–712. https://doi.org/10.5897/JMPR2017.6492

CCK+ CA1 interneurons differentially contribute to hippocampus-dependent behaviour

K. Balueva[1], P. Wulff[1]

1 Department of Physiology, Christian-Albrechts-Universität zu Kiel, Germany. P.wulff@physiologie.uni-kiel.de

The mammalian hippocampus is well studied for its role in the formation of declarative memories. Whereas glutamatergic principal cells encode information in the firing patterns of variable cell ensembles, GABAergic interneurons shape this process by regulating principal cell activity and the composition of active ensembles. These GABAergic interneurons differ in their morphology, their inputs and synaptic target domains on principal cells as well as their activity during specific network states [1]. Accordingly, different types of interneurons may contribute differentially to hippocampus-dependent learning. Recent studies on interneuron types expressing either the marker gene parvalbumin or somatostatin have indeed revealed their specific involvements in hippocampus-dependent behavior [2–4], supporting this notion. In contrast, very little is known about the behavioural relevance of another type of GABAergic interneurons in the hippocampus, which expresses the neuropeptide cholecystokinin (CCK). CCK-expressing interneurons (CCKIs), which like parvalbumin- or somatostatin-expressing interneurons, comprise different morphological subtypes, receive serotoninergic input from the raphe nuclei, signal via anxiety-related alpha2 subunit containing GABA-A receptors and are pre-synaptically modulated via endocannabinoids, which allows for close plastic interactions with postsynaptic principal cells [5–7]. Due to the molecular composition of their input and output interfaces it has long been speculated that CCKIs may be involved in mediating subcortical influences including mood or motivation [6]. The paucity of hard evidence regarding the behavioural function of CCKIs, is likely due to the difficulties related to the genetic targeting of these cells, because also pyramidal cells express CCK. This has prevented the CKKI-specific transgenic expression of molecular tools to gain control over these cells.

To directly probe the behavioural relevance of CCKIs in the CA1 region of the mouse hippocampus, we have implemented an intersectional virus-based method to specifically express the inhibitory DREADD hM4Di in these neurons. We found that our approach allows highly specific targeting of hippocampal CCKIs. Specific CNO-mediated chemo-genetic silencing of CA1 CCKIs caused dis-inhibition of CA1 pyramidal cells as assessed by enhanced novelty-induced expression of the immediate early gene zif268. To test for hippocampus-related behaviour we investigated spatial maze learning and contextual fear learning as well as spontaneous recognition memory for objects or conspecifics. We found that silencing of CCKIs improved contextual learning and object recognition but impaired social recognition. We suggest that CCKIs in the hippocampal CA1 region are differentially tied into circuits that underly specific memory circuits.

All experiments were performed in accordance with the German law on animal protection and approved by the Animal Care and Ethics Committee of the Christian-Albrechts-University, Kiel.

References

1. Klausberger T, Somogyi P. Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. Science. 2008;321: 53–57. doi:10.1126/science.1149381

2. Murray AJ, Sauer J-F, Riedel G, McClure C, Ansel L, Cheyne L, et al. Parvalbumin-positive CA1 interneurons are required for spatial working but not for reference memory. Nat Neurosci. 2011;14: 297–299. doi:10.1038/nn.2751

3. Ognjanovski N, Schaeffer S, Wu J, Mofakham S, Maruyama D, Zochowski M, et al. Parvalbumin-expressing interneurons coordinate hippocampal network dynamics required for memory consolidation. Nat Commun. 2017;8: 15039. doi:10.1038/ncomms15039

4. Lovett-Barron M, Kaifosh P, Kheirbek MA, Danielson N, Zaremba JD, Reardon TR, et al. Dendritic inhibition in the hippocampus supports fear learning. Science. 2014;343: 857–863. doi:10.1126/science.1247485

5. Harris KD, Hochgerner H, Skene NG, Magno L, Katona L, Bengtsson Gonzales C, et al. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. PLoS Biol. 2018;16: e2006387. doi:10.1371/journal.pbio.2006387

6. Freund TF, Katona I. Perisomatic inhibition. Neuron. 2007;56: 33–42. doi:10.1016/j.neuron.2007.09.012

7. Castillo PE, Younts TJ, Chávez AE, Hashimotodani Y. Endocannabinoid signaling and synaptic function. Neuron. 2012;76: 70–81. doi:10.1016/j.neuron.2012.09.020

Acknowledgements

The program committee are very grateful for all the hard work done by the reviewers of the papers in this Proceedings:

Andrew Spink
Anne-Marie Brouwer
Annika Bremhorst
Bas Rodenburg
Chiara Canori
Elsbeth Van Dam
Frances Wiseman
Gernot Riedel
Giulia Pedretti
Hans Theuws
Ivo Stuldreher
Jason Rogers
Kevin Ike
Khiet Truong
Kyle Roddick
Lars Lewejohann
Lianne Robinson
Liezl Maree
Loes Ottink
Maarten Reijnders
Malou van der Sluis
Marie Schneider
Maykel Van Miltenburg
Mike Toscano
Mona Giersberg
Paola Valsecchi
Peter Juhas
Poja Shams
Reinko Roelofs
Richard Brown
Rick D'eath
Ruud Tegelenbosch
Szu-han Wang
Tenzing Dolmans
Thilo Womelsdorf
Timon Daniels
Tiziano Travain
Tobias Heffelaar
Vivek Kumar