Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience

Nastacia L. Goodwin[1,2,a], Simon R. O. Nilsson[1,a], Jia Jie Choong[1,4] and Sam A. Golden[1,2,3]

Abstract

The use of rigorous ethological observation via machine learning techniques to understand brain function (computational neuroethology) is a rapidly growing approach that is poised to significantly change how behavioral neuroscience is commonly performed. With the development of open-source platforms for automated tracking and behavioral recognition, these approaches are now accessible to a wide array of neuroscientists despite variations in budget and computational experience. Importantly, this adoption has moved the field toward a common understanding of behavior and brain function through the removal of manual bias and the identification of previously unknown behavioral repertoires. Although less apparent, another consequence of this movement is the introduction of analytical tools that increase the explainabilty, transparency, and universality of the machine-based behavioral classifications both within and between research groups. Here, we focus on three main applications of such machine model explainabilty tools and metrics in the drive toward behavioral (*i*) standardization, (*ii*) specialization, and (*iii*) explainability. We provide a perspective on the use of explainability tools in computational neuroethology, and detail why this is a necessary next step in the expansion of the field. Specifically, as a possible solution in behavioral neuroscience, we propose the use of Shapley values via Shapley Additive Explanations (SHAP) as a diagnostic resource toward explainability of human annotation, as well as supervised and unsupervised behavioral machine learning analysis.

Addresses

[1] University of Washington, Department of Biological Structure, Seattle, WA, USA

[2] University of Washington, Graduate Program in Neuroscience, Seattle, WA, USA

[3] University of Washington, Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, WA, USA

[4] University of Washington, Department of Electrical and Computer Engineering, Seattle, WA, USA

Corresponding author: Golden, Sam A (sagolden@uw.edu)
(Goodwin N.L.), (Nilsson S.R.O.), (Choong J.J.), (Golden S.A.)

[a] Denotes equivalent contribution.

Current Opinion in Neurobiology 2022, **73:**102544

This review comes from a themed issue on **Neurobiology of Behavior 2022**

Edited by **Tiago Branco** and **Mala Murthy**

For complete overview of the section, please refer the article collection - Neurobiology of Behavior 2022

Available online 26 April 2022

https://doi.org/10.1016/j.conb.2022.102544

Introduction

Ethology, the study of behavior and social organization from a biological perspective, is a central tenant of modern behavioral neuroscience research. Invariably, ethological observation and analysis is plagued by the significant effort and time required to manually annotate complex behavior. These careful efforts are rewarded, however, by the ability to identify and categorize common behaviors in model species, as well as variations due to disease or experimental perturbation. This ability is an integral component of behavioral model reproducibility and extendibility. More recently, there has been a concerted effort to combine these ethological observations with machine learning–based approaches for automated behavioral classification, in combination with the study of brain function, under the umbrella of computational neuroethology [1].

The combination of automated classification with ethological analysis circumvents several issues inherent to manual annotation, namely observer drift and bias, long analysis times, and poor inter-rater-reliability both within and between research groups [1–3]. Frame-by-frame behavioral analysis also allows for the acquisition of behavioral annotation with sampling frequencies and accuracies that match modern neural recording and manipulation techniques. Thus far, machine learning

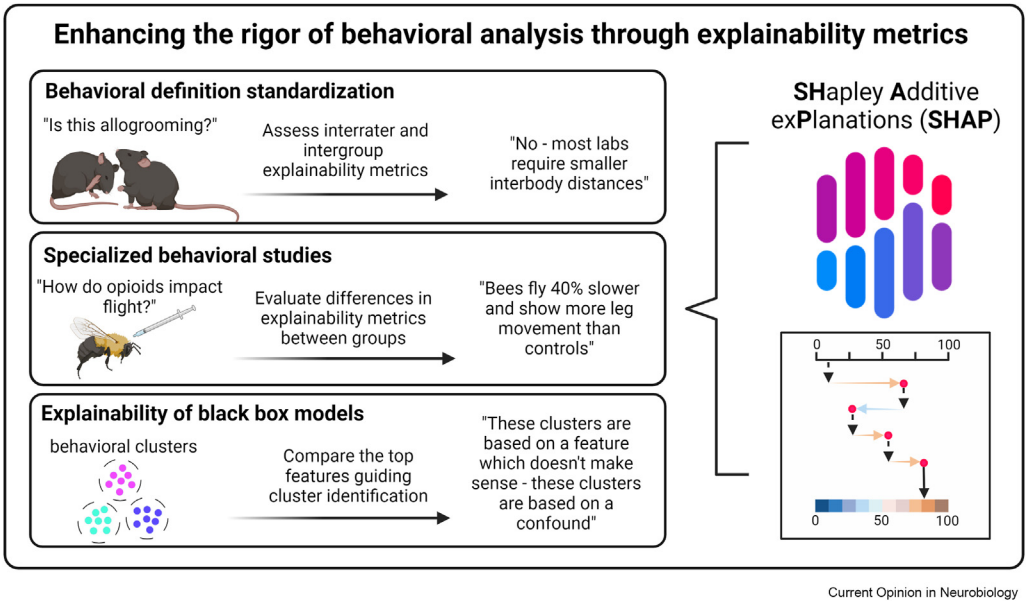
algorithms have uncovered previously unknown behavioral repertoires in model organisms and have confirmed and extended foundational assumptions of ethology through the incorporation of neural data [4–6].

With the increased adoption of new open-source platforms for automated behavioral classification, machine learning techniques are now accessible to a wide array of behavioral neuroscience labs despite variations in budget and computational experience. A major advantage of these machine learning approaches is the identification of behavioral latent motifs—the individual behavioral components that make up a gestalt behavioral definition. The machine learning models that predict these motifs and behaviors can be shared between and iterated on by research groups to develop a common understanding of behavior. In this current opinion we focus on three main consequences of this last statement, and their applications toward behavioral neuroscience (Figure 1).

First, despite repeated calls by researchers for more transparent and operationalized reporting [7,8] and an equally strong edict by the National Institutes of Health (NIH) in the form of enhanced reproducibility through rigor and transparency [9], behavioral neuroscience has yet to enter a 'golden era' of behavioral definition standardization. Pragmatically, it is common for observational behavioral metrics to be reported as loose subjective definitions within a sentence or two in the methods. More recently, predominantly because of journals adopting the NIH rigor and reproducibility standards, observational behavioral metrics are reported with more objective operationalized definitions including information regarding a behavior's timing, transitions, and constraints. Despite these efforts, however, the challenges of manual behavioral annotation inherently prevent consistent use and transfer of operationalized behavior definitions. Here, we propose that the widespread adoption of machine learning–based automated behavior classification will circumvent issues inherent to manual annotation, allowing for more standardized behavioral definitions both within and across research groups. More specifically, we highlight the importance of developing and reporting explainability metrics when using these tools.

Second, while the standardization of assays and behavioral definitions is considered integral to reproducibility, studying underlying mechanisms often and seemingly paradoxically requires deviation from standardized behavioral procedures. Indeed, with the use of increasingly complex neural recording and manipulation methods, there is an equal need to repeat that "the neural basis of behavior cannot be properly characterized without first allowing for independent detailed study of the behavior itself" [10]. Although it is pragmatically attractive to operationally define a standardized behavior as a proxy for its neural correlates and control,

Figure 1



Enhancing the rigor of behavioral analysis through explainability metrics. SHapley Additive exPlanations (SHAP) is an open-source explainability platform under continuous development by the AI field. SHAP has several advantages for computational neuroethology, including promoting standardized behavioral definitions, providing objective and specific metrics of changes in behavior between groups, and in increasing the transparency of unsupervised clustering results.

this can introduce bias when expanding a study population. The need for extensions of standardized behaviors are highlighted by recent observations considering sex as a biological variable in stress-related procedures, which have revealed, for example, that female rats demonstrate 'darting', a novel active fear-coping behavior not typically seen in males [11], while female California mice show increased social vigilance [12]. Similarly, driven by technological considerations, the transition from head-fixed neural recording to freely moving neural recording has identified divergent mechanisms between the same constrained or unconstrainted behaviors [13]. Questioned directly, as neural data increase in fidelity and size and as behavioral procedures become more specialized, how can the field of behavioral neuroscience ensure accurate reporting of these behavioral procedures in a way that ensures reproducibility? The need for behavioral specialization is at odds with the movement toward more generalized standardization - but aligns well with approaches in machine learning explainability.

Third, to negate the above concerns, significant effort has been spent developing unsupervised approaches for behavioral analysis that may be especially suited for identifying previously unnoticed behavioral motifs in model species. However, these unsupervised models are often opaque, and users cannot readily determine the reasoning that guides their decisions. In commercial sectors, such a lack of transparency is unacceptable by both end users and regulatory agencies. To retain scientific rigor and reproducibility, behavioral neuroscience should be at the forefront of demanding transparency via explainable models when possible, or via post-hoc explainability methods when black-box models cannot be avoided.

Here, we propose the adoption of Shapley values using Shapley Additive Explanations (SHAP) as an open-source resource for the explainability of human annotation as well as supervised and unsupervised behavioral classification results in preclinical behavioral neuroscience. We present several examples of the utility for using Shapley values within the context of the above three concerns. Shapley values are, of course, only one possible solution, and we provide a summary of alternatives that can be used when adopting both supervised and unsupervised machine learning into the analytical portfolio of behavioral neuroscience labs.

A brief machine learning primer

Before diving into the details of explainability, transparency, and universality, a brief primer on machine learning provides valuable context (See Box 1). First, critically, accurate behavioral classification requires equally accurate and precise animal tracking data. Typically, frame-by-frame positional data are obtained via 2D or 3D pose estimation [14—18], algorithms such as optical flow [19], or histograms of oriented gradients [20]. The fields of pose estimation and video tracking are rapidly expanding and beyond the scope of this article (but see Refs. [21,22] for excellent reviews on the state of the field). Ultimately, it is important for users to understand that with any machine learning application, the quality of the input (tracking and video annotations) directly impacts the quality of the output. The adage 'garbage in, garbage out' is especially accurate in this case. Significant improvements to accessibility and use of open-source software have occurred in the last several years ensuring that users, with care and validation, can generate consistently acceptable pose-estimation models using open-source tools. Once quality tracking has been obtained, users can move on to using machine learning algorithms that typically fall into supervised or unsupervised classes, which can be used entirely separately or in combination for different analytical purposes.

Supervised machine learning techniques are often sufficient for basic behavioral analysis, and can more simply tell researchers when and how often a predefined behavior of interest is occurring. For each individual frame of a video, users provide the algorithm with (*i*) many calculated features (i.e., nose to nose distance), and (*ii*) the 'answer' as to whether the behavior of interest is occurring. Different supervised methods such as support vector machines, random forests, and gradient boosting machines then look for mathematical rules which best delineate the features of negative versus positive frames. For more detailed explanations of their application, there are numerous reviews available [1,2,23].

When using unsupervised techniques, users provide no 'answer' for the algorithm, but rather set rules for similarities and differences within the data. The art of tuning these hyperparameters can be difficult, and generally requires a deeper knowledge of statistics and the models being used. When wielded correctly, unsupervised algorithms output different clusters (which can be adjusted to consist of individual frames or entire behavioral bouts). Researchers then select and look at several examples from each cluster to try to establish definitions that differentiate the clusters behaviorally. Importantly, users must be careful to tune hyperparameters based on objective metrics like realistic physiological parameters, to prevent any biases regarding expected behaviors and predictions that fall outside of explainable reasoning.

These efforts are championed by the burgeoning field of Explainable Artificial Intelligence (XAI), which calls for the use of explainable methods for both high- and low-stakes decision-making. Succinctly, XAI proposes that if a model lacks validation, or if a greater understanding of how a model works is required, XAI provides tools to

Box 1. Definitions of interest (indicated with bold italic in main text).

Term	Description
Coalition game	Coalition game theory deals with situations where the outcome is known, and rules dictate how actors interacted to achieve the outcome, and how the proceeds of the game should be divided amongst the actors. In the context of machine learning, the outcome is the models' decision, and the actors are the features.
Counterfactual	The minimum change within a feature value that produces an alternative decision by a machine learning model.
Deterministic vs stochastic	Stochastic algorithms produce different results when implemented twice on the same data set. Deterministic algorithms give the same results when implemented twice on the same data set. Stochasticity is favorable on larger data sets and is present in most prevalent machine learning techniques in behavioral neuroscience (e.g. decision trees, deep neural networks). Algorithm stochasticity can typically be reduced by increasing the number of observations (i.e. behavior events). Stochastic methods may be considered disadvantages in regulated settings, as the same algorithm can produce different outcomes for the same observation when implemented twice.
Ensemble model	The grouping of multiple learning algorithms. The individually weak learning algorithms are combined to create a strong or accurate behavior classification (e.g. random forest algorithm).
Entropy and Gini impurity	Measurements that evaluate the randomness of target labels after a feature split. For example, if the behavior of interest requires a minimum velocity, features that measure body-part velocities will decrease entropy and impurity and be concluded to have greater importance for accurate classification.
Feature battery	Numerical measurements that together describe the activities and morphologies of the actor(s) in video frames. A battery can be composed of several hundred values that represent data from both a single video frame (e.g. the volume or rotation of an actor) and aggregate data that include preceding video frames (e.g. the velocity or path of a body-part).
Feature classes	A subgroup of features in the feature battery that measure similar activities or morphologies in the video frame. For example, two values that represent the velocities of two different body-parts both belong to a class of velocity features.
Feature value	A single metric value representing a feature in a feature battery. For example, a feature value can be the millimeter distance between two body-parts.
Interpretability calculation	A calculation which attempts to make the machine's decision processes understandable for humans. An interpretability calculation can provide a numerical output that makes the decision on a single video frame interpretable (also referred to as 'local interpretability' or 'explainability'), or an aggregate output that make the decisions on all video frames interpretable (also referred to as 'global interpretability').
Model decisions	The final output of a machine learning model. A classifier decision may be presented as a single value, or a vector of values, ranging between 0 and 1. For example, the probabilities that a behavior is present in an image or that a body-part exists in different parts of an image are model decisions.
Permutation importance	The decrease in model performance when the values associated with a feature are randomly shuffled.
Shapley additivity axiom	Dictates that the explanation algorithms' output value relates to the decision value. Within a decision tree ensemble algorithm, the additivity axiom ensures that the grand total contribution of a feature is the sum of its contribution to each tree.
Shapley value	In machine learning, Shapley values is a method for summarizing the cooperation and contribution of the features to the model decision.
Solution concept	Formal rules within game theory that dictate the strategies of the players to produce rational game outcomes.

let you do so. Recent reviews provide detailed explanations of the predominant concepts and challenges within the XAI field [24—29]. Here, we will focus our discussion on explainability methods that are either intrinsic (built into the model) or applied post-hoc, and we define 'explainable' as how readily users can understand the relationship between feature values and the model's decisions.

Critically, the principles of behavioral neuroscience are not changed by the incorporation of these concepts and techniques. While most of these algorithms were not designed with the field in mind, great strides have been made to modify canonical machine learning tools for use in ethology. Common principles of behavioral design remain—training sets should consist of individuals from each experimental group, and training a classifier on the behavior of a very small number of animals will not provide a generalizable model. Just as machine learning algorithms rely on good tracking data, they also rely on solid experimental design with construct and predictive validity toward the biology they are examining. More specifically and pragmatically, as more labs begin adopting machine learning tools for behavioral classification, we note the opportunity to set a precedent for the incorporation of accessible explainability metrics.

Looks can be deceiving

A scientific approach solicits face, construct, and predictive validity within its methods and results [30]. In pose-estimation and behavior classification, face and predictive validity are immediately achieved through intuitive visualizations with overlays and other performance metrics, while construct validity typically is not a requisite. Construct validity is essential, however, as the models and human observers may study different features to reach the same conclusion. For example, behavioral classification models can be sensitive to minor perturbations in experimental set-ups between training and testing (changes in animal sizes, age, sex, bedding material, presence/absence of experimental equipment such as head stages, wires, and operant modules). Explainability metrics can be used as a diagnostic tool to identify confounding variables and avoid such pitfalls a priori.

Tools for transparency and explainability

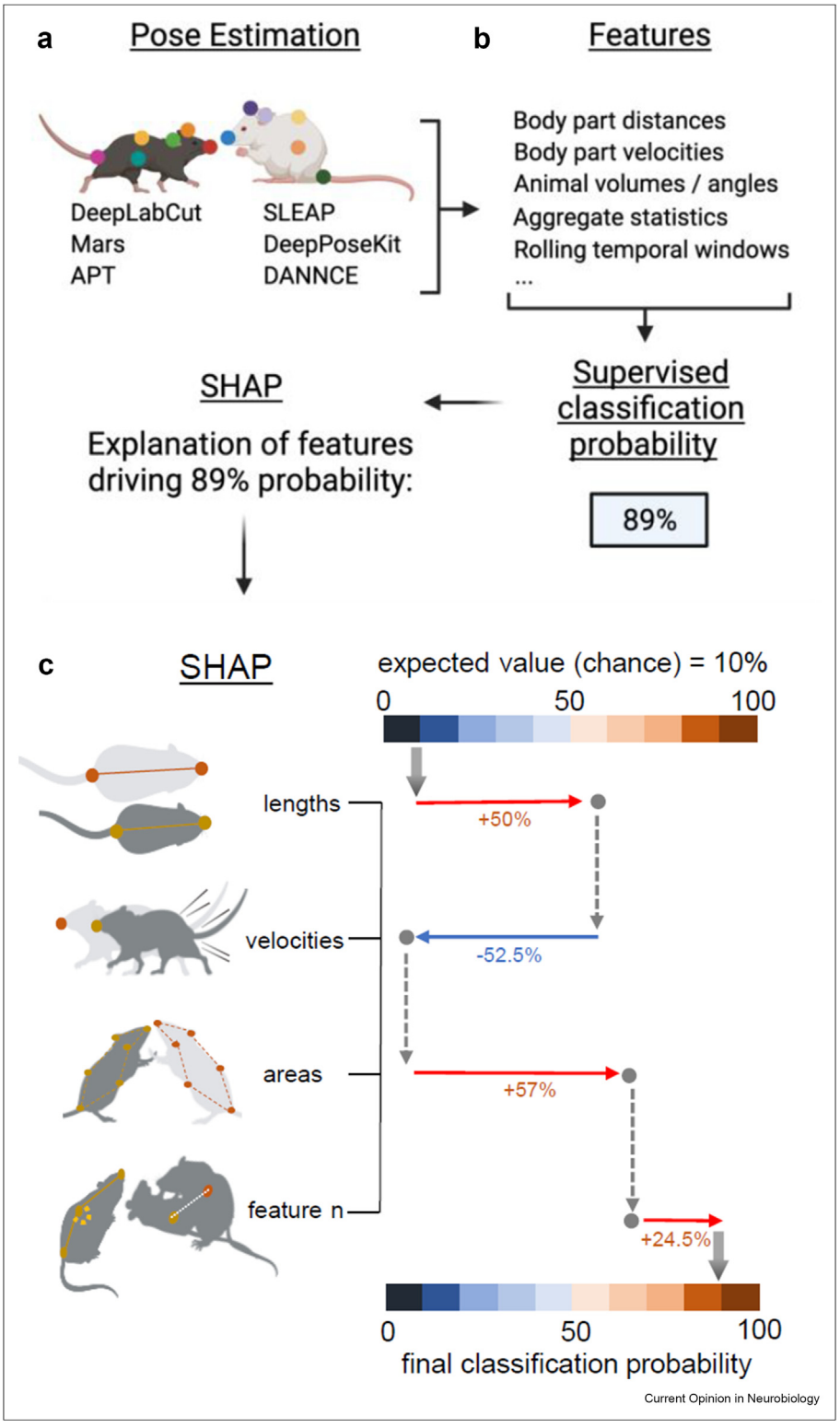
The use of XAI is rapidly expanding through the inclusion of both inherently explainable algorithms [24] and post-hoc explainability methods [31,32]. Here, we support the use of post-hoc explainability methods for computational neuroethology in behavioral neuroscience due to their relative simplicity of use, flexibility, and generalizability to diverse platforms. Primarily, despite a decade of calls for increased training, only ~15% of neuroscience graduate programs in the United States require computational coding courses [33—37]. Within the field of behavioral neuroscience, this generally functionally translates into the development of machine learning platforms that are lab specific. Second, once a platform is created and has face validity there is little incentive—or funding support—to further develop and incorporate explainable algorithms. Optimally, newly developed platforms should be created with inherent explainability in mind. Functionally, incorporation of post-hoc explainability methods that do not require de-novo platform development are more likely to be widely adopted within the field.

Explainability methods aim to provide accurate and verbalizable accounts for how known ***feature values*** contribute to ***model decisions***. The output from ***explainability calculations*** enable researchers to communicate and compare the results of disparate classifiers and supports experimenters in making informed decisions for machine model implementation and use. Interpretability is an active and emergent research area, and several initiatives in academic and commercial sectors are focused on devising new algorithms. These initiatives focus primarily on explaining decisions, promoting human/AI interactions, and the user's ability to evaluate and improve the trustworthiness and ***generalizability*** of different models.

Explainability methods can be ***deterministic or have gradations of stochasticity***, and provide local (e.g. individual video frames) and/or global explanations (aggregate feature weights within a model). The methods may be algorithm-specific or agnostic, and be intrinsic to the model or require additional processing. The core methodologies of most interpretability approaches will nevertheless seem familiar to experimentalists: features undergo some class of iterative ablation or permutations, the change in performance after such manipulation is evaluated, and the manipulation's impact is succinctly summarized [38]. How the permutation is performed, evaluated, and summarized differs between methods with varying relatedness to human processes and intuition for complex outcomes [39].

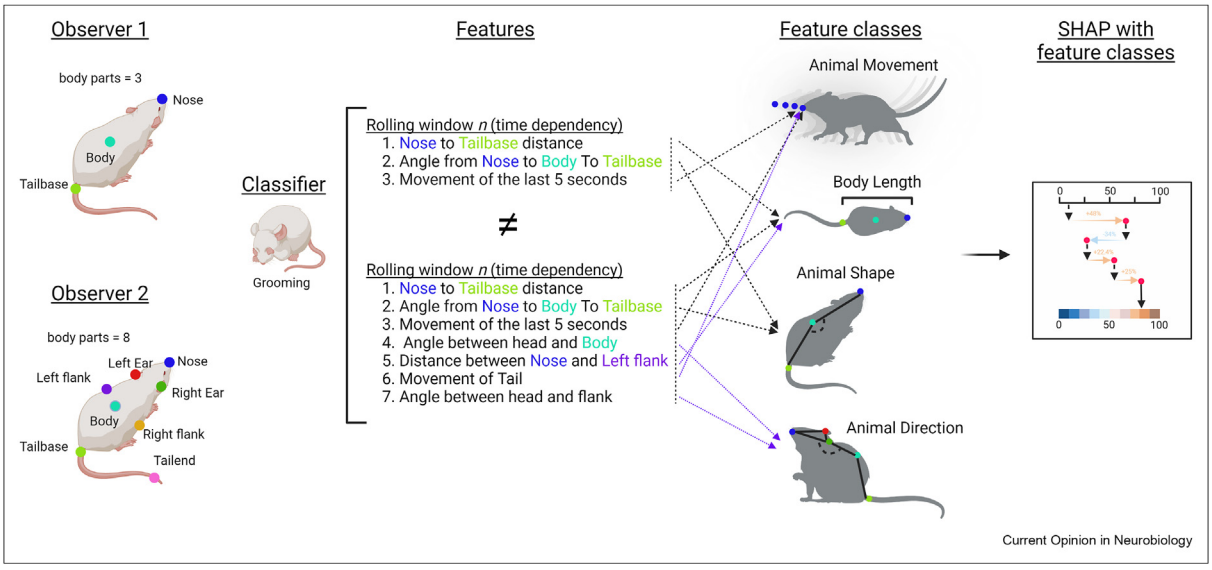
Although the advancement of behavior research may necessitate divergent experimental protocols and computational methods, any practical explainability methods obligate a certain level of cross-lab standardization. The most meaningful explainability metrics would also provide intuitive information on strength, directionality, and potentially linear and non-linear relationships between the known feature values and model decisions at the level of individual observations. Meaningful explanations in behavioral science also strongly benefit from methods that are ubiquitously used, transparent, and open source. Accessible explainability methods also compute feature contributions in ways that agree with human intuition. Recent XAI

Figure 2



Shapley value explainable artificial intelligence pipeline for behavioral neuroscience. (**a**) Many automated behavior analysis platforms rely on pose estimation data from raw video as input, followed by the calculation of feature batteries based on the position of animals within and across frames. (**b**) These feature batteries are used as input to supervised machine learning algorithms, which output a probability score of the behavior of interest occurring in a specific video frame. (**c**) The contribution of the different features values toward the output probability score can be decomposed post-hoc into a verbalizable description. For example, features measuring animal lengths may contribute to an increase in the behavior classification probability, while features measuring animal velocities contributes to a decrease the classification probability of the same behavior. Shapley values ensures that the combined description of all feature contributions is accurate, rational, and related to the final classification probability through its additivity and efficiency axioms.

Figure 3



Adapting SHAP to behavioral neuroscience. Standardization in behavioral neuroscience remains elusive, and labs may differ in pose estimation techniques as well as specific features that they calculate per video frame. Despite these differences, binning features into broader feature categories allows for the direct comparison of these classifiers across labs.

methods achieve this by exploiting ***solution concepts*** from collaborative game theory, using axioms defined by *fair* and *rational* payoffs in N-person games [38]. The axioms vary by solution concept, but some are more desirable and generally accepted within the context of machine learning classification. For example, it may be desirable that:

(i) Features with identical effects on classification probabilities are assigned identical importance scores (symmetry),
(ii) Features without effect on classification probabilities are allocated zero importance (null or dummy features),
(iii) The contribution of a feature within an ensemble model should equal the sum of its contributions to each model in the ensemble (additivity), and
(iv) The grand total of feature contributions should equal the grand total of the model output (efficiency).

The Shapley value is a solution concept that uniquely satisfies these rules [40] and is, therefore, an appealing tool for understanding and presenting explainability metrics within machine learning applications.

In brief, the Shapley value is a ***coalition game theoretical approach*** for fairly distributing the proceeds of a game to its players. To understand the contribution of feature X in a machine learning model, the output of a model involving features S is subtracted from the output of a model involving features $S + X$. This calculation is subsequently repeated for all possible permutations of features in the ***model***. Feature X is subsequently allocated the proceeds according to its contribution within each model permutation. The Shapley value holds theoretical properties that can make interpretations readily understandable, including additivity among features in relation to a model's prediction. This is known as the ***Shapley additivity axiom***.

Shapley values are a method for presenting intuitive explainability metrics while maintaining agnosticism to the model's structure. Its use in machine learning, however, has been prohibited by its computational costs; calculating Shapley values for all feature permutations necessitates exponential time and unfeasible runtimes for even smaller data set. However, through a suite of approximation algorithms, the SHAP library [27] performs Shapley value calculations using model agonistic approaches (KernelSHAP), tree-based models (TreeSHAP [41]), linear models (LinearSHAP), deep neural networks (DeepSHAP [32]) in polynomial time making it accessible to machine learning applications and behavioral neuroscience. The optimal SHAP algorithm is determined by the input model; TreeSHAP, LinearSHAP and DeepSHAP are optimized for decision trees, linear model, and neural networks, respectively, while KernelSHAP remain model agnostic.

Shapley values in behavioral analysis

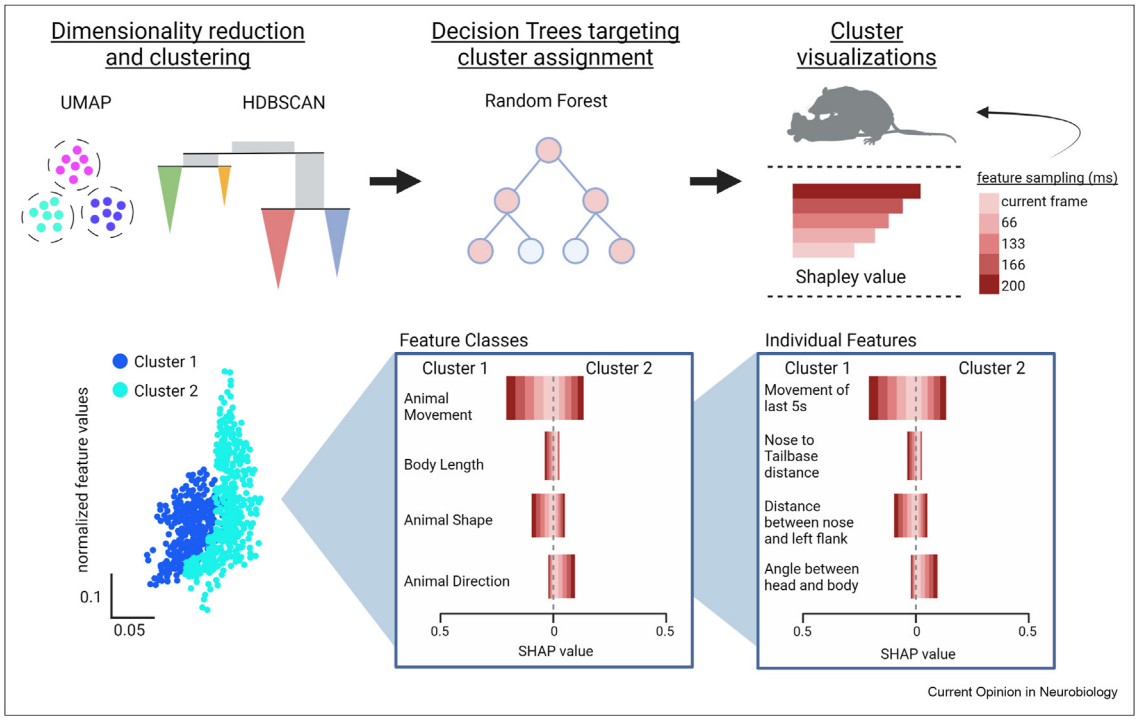
Computational developments [42–44] have made it possible to efficiently calculate comprehensive ***feature***

batteries for accurate and generic machine learning pipelines in behavioral neuroscience. Within such generic applications, feature batteries are comprised of hundreds of keypoint-to-keypoint metric distances, angles, and movement velocities and path tortuosities in various temporal and spatial windows. At this resolution, individual feature values have limited relevance for explaining human annotator conduct and classification probabilities, and presenting feature explainability metrics for each individual feature does not make the model's decisions explainable.

Here, the Shapley value ***additivity axiom*** can help make model decisions more explainable (Figure 2). Before the evaluation of individual feature contributions for a binary behavioral classifier, there are two known values. First, we know the average probability (*expected value*) that the behavior is occurring in a video frame. For example, the expected value is 5% if the behavior was annotated as present in 5% of the video frames used to create the classifier. Second, we also know the outcome prediction for an individual frame. For example, the model predicts that the behavior is present, in a previously unseen frame, with 85% probability. When using TreeSHAP and Shapley values, the difference between these two numbers (the *outcome probability* minus the *expected value* equals 80%) is attributed to the combination of feature values, and this difference value is distributed among the features according to their contribution within the ensemble. The Shapley value output in this example is a vector which length equals the number of features, and which sum equals 80.

To increase interpretability and universality of Shapley values representing a given supervised classification model, we can collapse Shapley value contributions of features that measure similar, general, and often colinear characteristics of the behavior of interest into interpretable biologically based categories while maintaining consistency and accuracy. To understand random forest classifiers, we propose the aggregation of Shapley values, for example derived from the open-source TreeSHAP package, into different behavioral classes and sampling frequency sub-classes (Figure 3). In Figure 3, for example, we have selected ***classes*** and ***sampling frequencies*** for explaining supervised predictions of complex social behavior in two interacting mice. Ultimately, these classes and sampling frequencies are dictated by the selection of model organism and behaviors of interest. The opportunity to bin feature importance scores into user-defined superordinate classes—while maintaining explainability—is a result of the mathematically

Figure 4



Supervising the unsupervised. Current methods for understanding unsupervised clusters are subjective—namely watching clustered bouts and trying to find and describe the differences. SHAP provides the opportunity for an objective description of feature differences contributing to cluster alignment via the creation and analysis of supervised classifiers for each cluster.

proven Shapley value additivity and efficiency axioms [40]. These axioms ensure that whichever Shapley values are collapsed, their aggregated grand total is a valid and interpretable representation of their combined contribution within the classification scenario. This attribute is appealing for behavioral neuroscience use-cases, and other experimental scenarios, that require a high-degree of flexibility.

Importantly, the use of these categories ensures that classifiers that target the same behavior, but use different feature sets, can be directly compared. A common reason that classifiers use different feature sets is a consequence of the initial pose-estimation scheme selected by experimenters. For example, two research groups interested in the same behavior may use pose-estimation schemes that identify 5 versus 8 body parts, respectively, due to their experimental needs; regardless, by binning their classification features within the same biologically determined classes and sampling frequencies, the use of Shapley values allows for direct comparisons between their classifiers.

We argue that explainability scores, such as Shapley values, can address several general concerns pertaining both supervised machine classifications and human annotations in behavioral neuroscience settings.

(i) As Shapley values can be used in a generalized manner, without needing identical pose-estimation key-points or feature sets, they provide a quantifiable, rather than qualitative, description of an operationalized behavioral definition that can be used across labs, institutes, and annotators.
(ii) Shapley values may reveal construct validity issues, where machine models produce accurate classifications through features that are independent correlates of the true features of interest.
(iii) Shapley values are widely accepted, understood, and under continual development and scrutiny in the greater computer science and AI fields.

In addition to its utility for supervised classification explainability, Shapley values can also be used a priori or post-hoc in unsupervised learning contexts to investigate or explain latent structures within embedding vectors (Figure 4). For example, to understand the discriminating features of different behavioral clusters in dimensionality reduced space, the cluster assignments can function as target variables for supervised classifiers. For example, the cluster 'A' classifier data set would contain all bouts for clusters A, and B, with cluster A bouts annotated as positive, and all other bouts annotated as negative. Users can then objectively compare which features drive the classification of a behavior as A or B. These supervised classifiers are straightforward to make with current open-source tools and are optimal for Shapley value calculations using the TreeSHAP library. Shapley values are then calculated for each cluster, and this analysis gives quantitative data of cross-cluster differences that can support and supplement qualitative visual analysis of the different behavioral events. Conversely, Shapley values have also been used as input to dimensionality reduction and clustering algorithms which, as opposed to raw feature values, intrinsically represent the importance of the features for the target variable and may thus produce more insightful cluster representations [27]. A further caveat in unsupervised analysis of raw feature values is the independent scales across the different features. This shortcoming is addressed when clustering Shapley values as all features are standardized to a single unit scale, which is the probability of the behavior [27].

Alternatives to Shapley values

Shapley values are one paradigm amongst a continuously growing number of approaches for XAI. The fundamental priority is not choosing a specific algorithm but rather that such algorithms become regularly used in behavioral neuroscience as machine learning approaches to behavior analysis become more common. Here, we have promoted the use of Shapley values because of its wide use [32] and highly active development [45], in conjunction with its relative ease of implementation and generalizability of post-hoc explainability to machine learning platforms in behavioral neuroscience. However, promising methods that address annotator differences within inherently interpretable models through program synthesis and wavelets are also being developed and are likely to significantly enhance inter-annotator model explainability [46,47].

Currently, popular and computationally inexpensive methods for interpretability in classification scenarios are aggregate gini impurity and entropy measures that provide information gain following inclusion of the feature into decision ensembles. These measures are intrinsically used to create decision trees and, therefore, often calculated by default. Most prevalent machine learning libraries provide aggregate feature importance through gini impurity and entropy measures, highlighting their popularity and common use (*Scikit*: [48]; MLlib [49]; *Xgboost:* [50]; Tensorflow [51]).

Conversely, ***permutation importance*** calculations provide a global explainability measure by calculating information loss following the scrambling of features [53]. LIME is another prevalent approach that provides local explainability metrics by fitting inherently interpretable linear regression models on perturbations around the original feature values [54]. LIME is closely related to SHAP but does not satisfy game theory axioms [32].

Another increasingly popular approach for local explainability that explores causality use ***counterfactuals***. Here, feature values are titrated and a causal relationship between the feature value and the outcome is concluded when concomitant changes in model decisions are observed [55].

Inherently explainable algorithms also can provide increased explainability for machine learning applications [24]. The conceptual definitions of 'inherently explainable' vary, but typically involve low-dimensional data sets (i.e. few features) that are analyzed using decomposable algorithms that can be verbalized without further advanced statistical analysis [56]. Notably, a decision tree algorithm can be accurate and considered 'inherently explainable' if carefully crafted using a reduced number of select and highly predictive features. Select features could be identified using experimenter domain-knowledge, or through preprocessing using advanced statistical and machine learning methods [47]. The inclusion of inherently explainable models should be a goal for developers. However, we propose a more general use of post-hoc explainability for behavioral classification for several pragmatic reasons:

(i) The output of prevalent machine learning toolkits for behavior analysis [5,57–61] are all directly amenable to post-hoc explainability methods. These tools, however, can currently not be used out-of-box to produce inherently explainable models.
(ii) A carefully curated feature-set can provide inherent interpretability for supervised applications. However, the same curated feature-set could bias and obstruct detection of experimenter-undefined novel behaviors in unsupervised applications. This disconnect could be addressed by using different feature-sets for supervised versus unsupervised learning.
(iii) Two inherently explainable machine learning models, that classify different behaviors, may not be directly compared as they will rely on different input data and feature sets.

Notably—although smaller feature-sets can produce both accurate and inherently explainable models—this does not equal computational savings for behavioral neuroscience use-cases. Many features computed by generic classification tools [57,58] are not necessary for accurate classification, but are required for post-hoc applications such as unsupervised learning and aggregate and conditional descriptive statistics. Thus, although fewer features are used by inherently explainable models, the users of inherently explainable models may still have to perform the same extensive calculations as those used by less transparent approaches.

Future directions

Ongoing efforts aim to develop, catalog, and promote XAI for deep learning and other machine learning methods [25,26]. While such work tackles technical topics of comprehensive importance, our own efforts explore the potential of well-developed XAI methods for practicable and wide use in behavioral neuroscience experiments. New XAI methods are developed and documented at pace, and the behavioral neuroscience field should be ready to adapt as superior and accessible methods become available.

We have focused on supervised applications, but there is a pressing demand to explain the biological relevance of behaviors detected through unsupervised algorithms. Unsupervised applications parse behavior with reduced or no experimenter oversight or ground-truth, and labels are then assigned by experimenters post-hoc [59–61]. The assignment of behavior labels may depend on qualitative assessments through visualizations and inspection of a reduced number of features. Future unsupervised applications would benefit from accurate, in-depth, and verbalizable statistical explanations for why the algorithm dissociated the clustered behaviors and what each clustered behavior represents, in addition to the qualitative appraisals and metrics on their validity [62]. This will enhance experimenter confidence in the algorithms and increase their validity and reproducibility within the greater behavioral neuroscience community.

Another ongoing direction, and area of active development, for both pose estimation pipelines and behavioral classification (DLStream [63]; DLC [64]; Google ML Kit [65], Tensorflow Lite [66]) are real-time applications. Real-time behavioral classification, based on real-time pose-estimation or other behavioral tracking, is a critical component for impending closed-loop recording and stimulation platforms that predict current and future behaviors. Within such applications, an in-depth understanding of the workings of real-time classifiers, partly through explainability measures, is critical as experiments depend on one-shot inference without opportunities for post-hoc re-analysis or model retraining. Further, from the technical perspective of computational expense during real-time predictions, enhanced explainability metrics allow for the winnowing of unneeded features without the loss of needed predictive power.

Lastly, supervised classification scenarios are becoming more complex and varied as they are more widely adopted. Future supervised classification scenarios will likely involve shifting multi-animal, multi-object environments, as well as user-defined spatial regions of interest and experimental subjects that share partly overlapping features (age, sex, strain) [67]. To understand the generalizability of machine learning models in

such complex, less stable settings, enhanced explainability may help.

Conclusion

The use of explainability metrics in combination with both supervised and unsupervised methods have clear implications for the transparency and universality of behavioral classification across research groups. Explainability metrics, if reported when using machine learning classification, provide discrete and quantifiable descriptions of behavior and remove the subjectivity of standard operationalized definitions. Shapley values, as described, provide a simple and effective metric that is easy to understand and apply in non-specialized settings.

Moving forward, for this purpose, efforts must be made to establish portals like the Research Resource Identification (RRID) database [68]. RRID reporting is now widely accepted, and often required for publication, due to its alignment with the Materials Design Analysis Reporting Checklist for Authors [69] that sets the minimum reporting criteria for studies in the life sciences and is endorsed by major journal publishers. Already spanning antibodies, reagents, model organisms, and even software, it is equally important to begin reporting behavioral classification metrics.

Early efforts to organize such a system have been led by open-source organizations such as the OpenBehavior [70] project, which provides a logistical framework to begin developing and supporting the hosting and dissemination of behavior-centric open-source pipelines. Alternatively, organizations like The Jackson Laboratory (JAX), already world-leaders in genetic database across inbred and outbred mice strains, have also turned their attention to behavioral phenotyping. The Mouse Phenome Database (MPD) provides primary phenotype data for the laboratory mouse and provides tools to analyze and visualize these data in relationship to JAX genetic databases [71]. Recently, the MPD has housed data generated using unsupervised analysis of grooming behavior [72]. Platforms like OpenBehavior.org and MPD are primed to not only host this behavioral analysis, but as machine learning becomes more common, also the provide access to the explainability metrics paired with these data.

Open-source initiatives now enable most behavioral neuroscience labs to perform automated and accurate detection of both experimenter-defined and novel behaviors in video-recordings. How these algorithms produce their predictions are typically not quantified, leading to a lack of transparency for cross-lab comparisons and machine model construct validity. Parallel strides in AI explainability have produced open-source algorithms that permit accurate and stable calculations of detailed contributions of individual features and their interactions for machine model decisions at the level of individual observations. These explainability algorithms have begun to be incorporated into many tools aimed at behavioral neuroscientists [14,57], and we anticipate that metrics based on Shapley values or other post-hoc explainability metrics will be an important part of these efforts.

Funding and disclosure

The authors declare that they do not have any conflicts of interest (financial or otherwise) related to the text of the article. The research was supported NIDA R00DA045662 (SAG), NIDA P30DA048736 (SRON and SAG), NARSAD Young Investigator Award 27082 (SAG), NIDA T32NS099578-04 (NLG), NIH F31MH125587-01 (NLG). SRO Nilsson's current affiliation is Deep Labs Research, CA, USA. We thank Ian Covert and Thomas Stearns for their thoughtful comments on the manuscript draft. Figures created with BioRender.com.

Conflict of interest statement

Nothing declared.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. Datta SR, Anderson DJ, Branson K, Perona P, Leifer A: **Computational neuroethology: a call to action**. *Neuron* 2019, **104**:11–24.
2. Anderson DJ, Perona P: **Toward a science of computational ethology**. *Neuron* 2014, **84**:18–31.
3. Egnor SER, Branson K: **Computational analysis of behavior**. *Annu Rev Neurosci* 2016, **39**:217–236.
4. Vogelstein JT, Park Y, Ohyama T, Kerr RA, Truman JW, Priebe CE, Zlatic M: **Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning**. *Science* 2014, **344**:386–392.
5. Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, Abraira VE, Adams RP, Datta SR: **Mapping sub-second structure in mouse behavior**. *Neuron* 2015, **88**: 1121–1135.
6. Rudolf J, Dondorp D, Canon L, Tieo S, Chatzigeorgiou M: **Automated behavioural analysis reveals the basic behavioural repertoire of the urochordate Ciona intestinalis**. *Sci Rep* 2019, **9**:2416.
7. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, *et al.*: **A call for transparent reporting to optimize the predictive value of preclinical research**. *Nature* 2012, **490**:187–191.
8. **Reality check on reproducibility**. *Nature* 2016, **533**. 437–437.
9. NOT-OD-15-103: *Enhancing reproducibility through rigor and transparency*. National Institutes of Health; 2015.
10. Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D: **Neuroscience needs behavior: correcting a reductionist bias**. *Neuron* 2017, **93**:480–490.
11. Gruene TM, Flick K, Stefano A, Shea SD, Shansky RM: **Sexually divergent expression of active and passive conditioned fear responses in rats**. *Elife* 2015, **4**.

12. Greenberg GD, Laman-Maharg A, Campi KL, Voigt H, Orr VN, Schaal L, Trainor BC: **Sex differences in stress-induced social withdrawal: role of brain derived neurotrophic factor in the bed nucleus of the stria terminalis**. *Front Behav Neurosci* 2014, **7**.

13. Meyer AF, O'Keefe J, Poort J: **Two distinct types of eye-head coupling in freely moving mice**. *Curr Biol* 2020, **30**: 2116–2130.e6.

14. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M: **DeepLabCut: markerless pose estimation of user-defined body parts with deep learning**. *Nat Neurosci* 2018, **21**: 1281–1289.

15. Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, Couzin I: **DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning**. *eLife* 2019, **8**, e47994.

16. Pereira TD, Aldarondo DE, Willmore L, Kislin M, Wang SS-H, Murthy M, Shaevitz JW: **Fast animal pose estimation using deep neural networks**. *Nat Methods* 2019, **16**:117–125.

17. Dunn TW, Marshall JD, Severson KS, Aldarondo DE, Hildebrand DGC, Chettih SN, Wang WL, Gellis AJ, Carlson DE, Aronov D, *et al.*: **Geometric deep learning enables 3D kinematic profiling across species and environments**. *Nat Methods* 2021, **18**:564–573.

18. Karashchuk P, Rupp KL, Dickinson ES, Walling-Bell S, Sanders E, Azim E, Brunton BW, Tuthill JC, Anipose: **A toolkit for robust markerless 3D pose estimation**. *Cell Rep* 2021, **36**: 109730.

19. Bohnslav JP, Wimalasena NK, Clausing KJ, Dai YY, Yarmolinsky DA, Cruz T, Kashlan AD, Chiappe ME, Orefice LL, Woolf CJ, *et al.*: **DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels**. *Elife* 2021, **10**, e63377.

20. Dolensek N, Gehrlach DA, Klein AS, Gogolla N: **Facial expressions of emotion states and their neuronal correlates in mice**. *Science* 2020, **368**:89–94.

21. Mathis MW, Mathis A: **Deep learning tools for the measurement of animal behavior in neuroscience**. *arXiv:190913868 [cs, q-bio]* 2019.

22. Pereira TD, Shaevitz JW, Murthy M: **Quantifying behavior to understand the brain**. *Nat Neurosci* 2020, https://doi.org/10.1038/s41593-020-00734-z.

23. Goodwin NL, Nilsson SRO, Golden SA: **Rage against the Machine: advancing the study of aggression ethology via machine learning**. *Psychopharmacology* 2020, https://doi.org/10.1007/s00213-020-05577-x.

24. * Rudin C: **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**. *Nat Mach Intell* 2019, **1**:206–215.

Dr. Rudin calls for the use of inherently interpretable machine learning models whenever possible, rather than using post-hoc explainability methods on black box models. Dr. Rudin states that, "trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society.

25. Das A, Rad P: **Opportunities and challenges in explainable artificial intelligence (XAI): a survey**. *arXiv:200611371 [cs]* 2020.

26. Shahroudnejad A: **A survey on understanding, visualizations, and explanation of deep neural networks**. *arXiv:210201792 [cs]* 2021.

27. * Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I: **From local explanations to global understanding with explainable AI for trees**. *Nat Mach Intell* 2020, **2**:56–67.

Presents the use of Shapley values for understanding decision processes of machine learning ensemble models. Their method is becoming a de facto standard in explainable artificial intelligence.

28. Doshi-Velez F, Kim B: *Towards a rigorous science of interpretable machine learning*. 2017. arXiv:1702.08608.

29. Vu M-AT, Adalõ T, Ba D, Buzsáki G, Carlson D, Heller K, Liston C, Rudin C, Sohal VS, Widge AS, *et al.*: **A shared vision for machine learning in neuroscience**. *J Neurosci* 2018, **38**:1601–1607.

30. Markou A, Chiamulera C, Geyer MA, Tricklebank M, Steckler T: **Removing obstacles in neuroscience drug discovery: the future path for animal models**. *Neuropsychopharmacology* 2009, **34**:74–89.

31. Shapley LS: **Stochastic games: Proc Natl Acad Sci U S A** 1953, **39**:1095–1110.

32. Lundberg S, Lee S-I: **A unified approach to interpreting model predictions**. *arXiv:170507874 [cs, stat]* 2017.

33. Goldman MS, Fee MS: **Computational training for the next generation of neuroscientists**. *Curr Opin Neurobiol* 2017, **46**:25–30.

34. Grisham W, Lom B, Lanyon L, Ramos L: **R: Proposed training to meet challenges of large-scale data in neuroscience**. *Front Neuroinf* 2016, **10**.

35. Pevzner P, Shamir R: **Computing has changed biology—biology education must catch up**. *Science* 2009, **325**:541–542.

36. Juavinett AL: **The next generation of neuroscientists needs to learn how to code, and we need new ways to teach them**. *Neuron* 2022, **110**:576–578.

37. *Society for neuroscience: surveys of neuroscience departments and programs*. Surveys of Neuroscience Departments and Programs; 2016.

38. * Covert IC, Lundberg S, Lee S-I: *Explaining by removing: a unified framework for model explanation*. 2020. arXiv:2011.14878 [cs.LG].

Covert et al. surveys popular methods for explainable AI, and describe foundational cross-method similarities as well as differences in algorithm behavior, and discuss variations in how the influence of features are summarized to the end-user.

39. Miller T: **Explanation in artificial intelligence: insights from the social sciences**. *Artif Intell* 2019, **267**:1–38.

40. Osborne MJ, Rubinstein A: *A course in game theory*. MIT Press; 1994.

41. Lundberg SM, Erion GG, Lee S-I: **Consistent individualized feature attribution for tree ensembles**. *arXiv:180203888 [cs, stat]* 2019.

42. Lam SK, Pitrou A, Seibert S: **Numba: a LLVM-based Python JIT compiler**. In *Proceedings of the second workshop on the LLVM compiler infrastructure in HPC - llvm '15*. ACM Press; 2015:1–6.

43. McKinney W: *Pandas: a foundational Python library for data analysis and statistics*. 2011.

44. RAPIDS: *Collection of libraries for end to end GPU data science*. RAPIDS; 2018.

45. https://github.com/slundberg/shap (GitHub repository).

46. Tjandrasuwita M, Sun JJ, Kennedy A, Chaudhuri S, Yue Y: **Interpreting expert annotation differences in animal behavior**. *arXiv:210606114 [cs]* 2021.

47. * Sun JJ, Kennedy A, Zhan E, Anderson DJ, Yue Y, Perona P: *Task programming: learning data efficient behavior representations*. 2020. arXiv:2011.13917.

Sun et al. presents a tool (TREBA) that can be used to generate accurate and inherently explainable behavior classifiers in neuroscience through a limited set of annotations.

48. Scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation. [date unknown],

49. Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai D, Amde M, Owen S, et al.: MLlib: machine learning in Apache spark. [date unknown],

50. Chen T, Guestrin C: **XGBoost: a scalable tree boosting system**. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016, https://doi.org/10.1145/2939672.2939785.

51. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, *et al.*: **TensorFlow: large-scale machine learning on heterogeneous distributed systems**. *arXiv:160304467 [cs]* 2016.

53. Breiman L: **Random forests**. *Mach Learn* 2001, **45**:5–32.

54. Ribeiro MT, Singh S, Guestrin C: **"Why should I trust you?": explaining the predictions of any classifier**. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining - kdd '16*. ACM Press; 2016:1135–1144.

55. Verma S, Dickerson J, Hines K: *Counterfactual explanations for machine learning: a review*. 2020. arXiv:201010596 [cs, stat].

56. * Lipton ZC: **The mythos of model interpretability**. *arXiv: 160603490 [cs, stat]* 2017.
The paper discusses the fallacies of interpretable AI, and challenges the concept that post-hoc explainability is by default unfavorable relative to less complex 'inherently interpretable' models.

57. Nilsson SRO, Goodwin NL, Choong JJ, Hwang S, Wright HR, Norville Z, Tong X, Lin D, Bentzley BS, Eshel N, *et al.*: **Simple Behavioral Analysis (SimBA): an open source toolkit for computer classification of complex social behaviors in experimental animals**. *bioRxiv* 2020, https://doi.org/10.1101/2020.04.19.049452.

58. Kabra M, Robie AA, Rivera-Alba M, Branson S, Branson K: **JAABA: interactive machine learning for automatic annotation of animal behavior**. *Nat Methods* 2013, **10**:64–67.

59. Hsu AI, Yttri EA: **B-SOiD: an open source unsupervised algorithm for discovery of spontaneous behaviors**. *Nat Commun* 2021, **12**:5188.

60. Graving JM, Couzin ID: **VAE-SNE: a deep generative model for simultaneous dimensionality reduction and clustering**. *bioRxiv 2020.07.17.207993*, 2020.

61. Luxem K, Mocellin P, Fuhrmann F, Kürsch J, Remy S, Bauer P: **Identifying behavioral structure from deep variational embeddings of animal motion**. *bioRxiv 2020.05.14.095430*, 2020.

62. Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J: **Density-based clustering validation**. In *Proceedings of the 2014 SIAM international conference on data mining (SDM)*. Society for Industrial and Applied Mathematics; 2014:839–847.

63. Schweihoff F, Loshakov M, Pavlova I, Kück L, Ewell LA, Schwarz MK: **DeepLabStream enables closed-loop behavioral experiments using deep learning-based markerless, real-time posture detection**. *Commun Biol* 2021, **4**:130.

64. Kane GA, Lopes G, Saunders JL, Mathis A, Mathis MW: **Real-time, low-latency closed-loop feedback using markerless posture tracking**. *Elife* 2020, **9**, e61909.

65. ML Kit. Google developers [date unknown].

66. TensorFlow Lite | ML for mobile and edge devices. TensorFlow [date unknown].

67. Winters C, Gorssen W, Ossorio-Salazar VA, Nilsson S, Golden S, D'Hooge R: **Automated procedure to assess pup retrieval in laboratory mice**. *Sci Rep* 2022, **12**:1663.

68. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, Hof PR, Martone ME, Pols M, Tan S, *et al.*: **The Resource Identification Initiative: a cultural shift in publishing**. *J Comp Neurol* 2016 Jan 1, **524**:8–22, https://doi.org/10.1002/cne.23913. PMID: 26599696.

69. Chambers K, Collings A, Graf C, Kiermer V, Mellor DT, Macleod M, Swaminathan S, Sweet D, Vinson V: *Towards minimum reporting standards for life scientists*. 2019, https://doi.org/10.31222/osf.io/9sm4x.

70. * White SR, Amarante LM, Kravitz AV, Laubach M: **The future is open: open-source tools for behavioral neuroscience research**. *eNeuro* 2019, **6**. ENEURO.0223-19.2019.
This article highlights the advantages of producing software for behavioral neuroscience under open-source licenses, and details on the founding and utility of The OpenBehavior Project.

71. Bogue MA, Philip VM, Walton DO, Grubb SC, Dunn MH, Kolishovski G, Emerson J, Mukherjee G, Stearns T, He H, *et al.*: **Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data**. *Nucl Acid Res* 2020, **48**:D716–D723.

72. Geuther BQ, Peer A, He H, Sabnis G, Philip VM, Kumar V: **Action detection using a neural network elucidates the genetics of mouse grooming behavior**. *Elife* 2021, **10**, e63207.