UCSF
UC San Francisco Previously Published Works

Title

The choice-wide behavioral association study: data-driven identification of interpretable behavioral components

Permalink

https://escholarship.org/uc/item/3n76x3hr

Journal

Nature Communications, 17(1)

ISSN

2041-1723

Authors

Kastner, DB
Yokota, NO
Lee, CY
et al.

Publication Date

2026-12-01

DOI

10.1038/s41467-026-76681-3

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

nature communications

Article

https://doi.org/10.1038/s41467-026-76681-3

The choice-wide behavioral association study: data-driven identification of interpretable behavioral components

Received: 2 March 2026

Accepted: 4 August 2026

Published online: 13 August 2026

Check for updates

David B. Kastner[1] ✉, Nicole O. Yokota[1], Christina Y. Lee[1], Cristofer Holobetz[1], Ali Chaudhry[1], Greer Williams[1], Jane N. Ton[1], Mihika Mehra[1], Viktor Kharazia[1], Dilworth Y. Parkinson[2], Joseph P. Romano[3,4] & Peter Dayan[5,6]

Behavior contains rich structure across many timescales. However, there is a dearth of methods to identify relevant components, especially over the longer periods required for learning and decision-making. Inspired by the goals and techniques of genome-wide association studies, we present a data-driven method—the choice-wide behavioral association study: CBAS—that systematically identifies such behavioral features. CBAS extracts sequences of actions or choices that either significantly differ between groups or correlate with a covariate of interest, using a powerful resampling-based method to correct for multiple comparisons. We illustrate CBAS through application to tasks performed by flies, rats, and humans, showing that it provides unexpected and interpretable information in each case.

Understanding how behavior differs between or among groups of humans or other animals is critical for generating and testing hypotheses about the functional role of genes, neural circuits, and regions of the brain. It is also essential for characterizing neurologic and psychiatric dysfunction[1]. However, behavior is highly complex, evolving over multiple timescales and exhibiting substantial path dependencies due to individual experience[2–6]. To deal with these complexities, it is increasingly popular to collect large datasets through automated behavioral paradigms, and to analyze them using computational and machine learning methods[7–16]. Approaches for analyzing such data range from top-down/model-dependent on one end of a spectrum to bottom-up/data-driven on the other end of the spectrum[17,18]. Methods that tend towards the model-dependent end of the spectrum are substantially more prevalent in the analysis of behavior; however, methods that tend toward the data-driven end of the spectrum dominate in other areas with rich and complex findings and afford the possibility of addressing some of the many limitations of top-down approaches.

In model-dependent analyses, behavioral data, such as choices in a decision-making task, are processed under the specific assumptions of a hypothesis or model. If the model or hypothesis is correct, this is highly efficient, since large volumes of data can be reduced to a handful of semantically meaningful parameters, such as learning rates or differential sensitivity to rewards or punishments. By finding specific parameters whose values differ significantly between groups, or correlate with facets of those groups, model-dependent analyses conflate identification with interpretation. That is both their strength and their weakness. If the model is correct, it allows for the difference or correlation to be directly interpreted in terms of the parameters concerned; however, if the model is incorrect or limited, confirmation bias[19] imposes an incorrect interpretation on the difference or correlation. That is, even when substantial effort is put into building multiple alternative models, and comparing them in a statistically rigorous manner, it remains possible that the best fitting model nevertheless fails to characterize the behavior properly. This limits the interpretability of the model or the inter-

[1]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA. [2]Advanced Light Source, Lawrence Berkeley National Labs, Berkeley, CA, USA. [3]Department of Statistics, Stanford University, Stanford, CA, USA. [4]Department of Economics, Stanford University, Stanford, CA, USA. [5]Max Planck Institute for Biological Cybernetics, Tübingen, Germany. [6]University of Tübingen, Tübingen, Germany. ✉e-mail: david.kastner2@ucsf.edu

rogation of hypotheses. In the worst cases, model and hypothesis-based analyses offer post-hoc explanations for any difference that can be found in a behavioral dataset[20].

Data-driven analyses start from the other end, characterizing behavior without relying on parametric assumptions, and making few assumptions about how the data are generated. Some of these approaches are unsupervised, for instance finding clusters in behavioral space and comparing them between groups[12,21,22]; others are more supervised, directly looking for discriminative differences between populations or individuals[7–10]. By not trying to shoehorn data into a limited set of parameters, these methods should be more sensitive; however, comparisons between groups pose severe statistical challenges, because of the complexity and dimensionality of behavioral datasets; and interpretation remains a challenge.

Modern machine learning methods[23–25] can sit between model-dependent and data-driven camps. They have the particular benefit of providing useful lower bounds for how well hypothesis-driven models should fit but often struggle with convincing interpretability and data efficiency.

Here, we present the choice-wide behavioral association study (CBAS), a data-driven analysis method designed to identify sequences of choices (or other discrete behavioral features) that differ between or among groups. CBAS has two components: (1) breaking down behavior into a comprehensive language that allows comparisons to be made between two groups or to assess correlation with a covariate of interest; and (2) using rigorous, resampling-based, corrections to account for the resulting large number of comparisons and to maintain statistical power despite statistical complexity in the data. CBAS breaks the conflation between detection and interpretation of model-dependent analyses because it is designed simply to identify differences or find correlations with a covariate. The presence (or absence) of that significance is interpretation-free. However, by design, the output of CBAS provides a rich and rigorous substrate for interpreting behavior, for testing the adequacy of models and hypotheses concerning the source of behavioral differences, and our method provides a general data-driven framework to advance our resulting understanding.

Results

Choice as a common discretization for behavior

The development of CBAS was motivated by the powerful data-driven approaches of genome-wide association studies (GWAS), whole exome sequencing (WES), and whole genome sequencing (WGS)[26]. Prior to the wide-scale implementation of these methods, candidate gene studies that attempted to associate candidate genes with phenotypes were often underpowered and failed to replicate[27–29]. It has been argued that many hypothesis-driven behavioral analyses suffer the same problems[30]. GWAS/WES/WGS famously solved the problem of statistical reliability by employing rigorous multiple-comparisons correction. These corrections depend on the discrete representation afforded by base-pairs[31,32]. CBAS follows an equivalent statistical path, and also requires a similarly appropriate, and potentially general, discretization, but for behavior. In many cases, choices provide just such a discretization (Fig. 1a).

We draw upon the analogy with the genomic methods (GWAS/WES/WGS) to explain the function and goals of CBAS. Genomic methods aim to identify variation in the genome (at the level of DNA base pairs or larger structural changes) that is associated with a phenotype of interest, such as disease. These variants can include single nucleotide polymorphisms (SNPs), insertions/deletions, or structural changes, and may occur in coding or noncoding regions. The nature of the variants identified guides understanding of the biological mechanisms underlying the phenotype or disease. In these terms, CBAS aims to identify variation in behavioral structure (at the level of sequences of discretized units such as choice) that is associated with a phenotype of interest such as a strain of fly or an index of compulsivity revealed by answers to questionnaires in humans. The structure of the sequences of choices guides understanding into the behavioral algorithms underlying the phenotype.

We apply CBAS to three tasks in three diverse species—flies, humans, and rats—to show the breadth of applicability of CBAS. For simplicity and clarity, we describe the motivation and development of CBAS using one of the tasks and then show its application to the other tasks below. The task that we will initially use to describe CBAS involves drosophila walking on a Y-maze (Fig. 1a; left). The flies are tracked as they spontaneously make the choice of going to either the left or right arm after leaving a current arm[31]. This left/right discretization of the task has enabled many conclusions to be drawn about the genetic nature of individual variability[31–36]. Here, we compare the choices of two outbred strains of fly, Cambridge-A (CA) and w1118, from a publicly available dataset[31]. The w1118 strain is the background strain for many transgenic flies[31]. Analyses in the original paper[31] provide a strong hypothesis for a potential source of the difference between the groups: handedness, or the preference for left or right turns, by the individual flies; however, there are many different ways in which that might manifest, and the rich dataset combined with CBAS allows multiple possibilities to be evaluated.

CBAS evaluates all sequences of choices up to a set length

Using choice as the basis of the comparison for the behavioral analysis requires an additional consideration beyond what is done for the genome with base-pairs. For GWAS/WES/WGS, an individual base pair can be a meaningful unit of information (although this is only a partial story[37,38]) that can be compared between subjects. However, isolated choices are rarely meaningful units by themselves, as the choices that precede and follow them can change their interpretation. CBAS does not just evaluate the occurrence of individual choices, but, rather, the occurrence of all sequences up to a certain, user-defined, length. In general, the longer the sequences, the more data and computational resources will be necessary, but the richer the aspects of behavior that can be encompassed.

For the fly spontaneous alternation behavior, we applied CBAS to all sequences of choices of turns up to a length of 10. The choice of 10 was made to allow for the possibility to identify rich structure that differentiates the groups. For each sequence of choices, all instances of that sequence up to a criterion of 250 choices are found in the dataset, and the rate of usage of that sequence in the two populations is used as the point of comparison at the basis of CBAS (Fig. 1b shows one particular example sequence).

Correcting for the many comparisons inherent in evaluating the occurrence of such sequences is complicated by the fact that their incidence can be highly correlated. Standard methods to correct statistically for multiple comparisons (e.g., Bonferroni, Holm, Benjamini-Hochberg, Benjamini-Yekutieli) are underpowered for correlated data. Instead, we use an approach based on resampling to correct for the multiple comparisons—called the Romano-Wolf method. This approach retains power in the face of correlations[39]. In practice, we set the significance threshold according to median control of the false discovery proportion at 5%. This control is comparable to 5% control of the false discovery rate, which provides mean control of the false discovery proportion[40].

Using the Romano-Wolf method for resampling-based multiple comparisons correction[39,40], CBAS identifies many sequences of choices that differ significantly between the two strains of flies in the spontaneous alternation task (1,605/2,046 sequences; Fig. 1c). We are also able to determine the sample size required to provide ~80% power to detect a difference at this level of statistical certainty, using subsampling of the data (Fig. S1).

https://doi.org/10.1038/s41467-026-76681-3

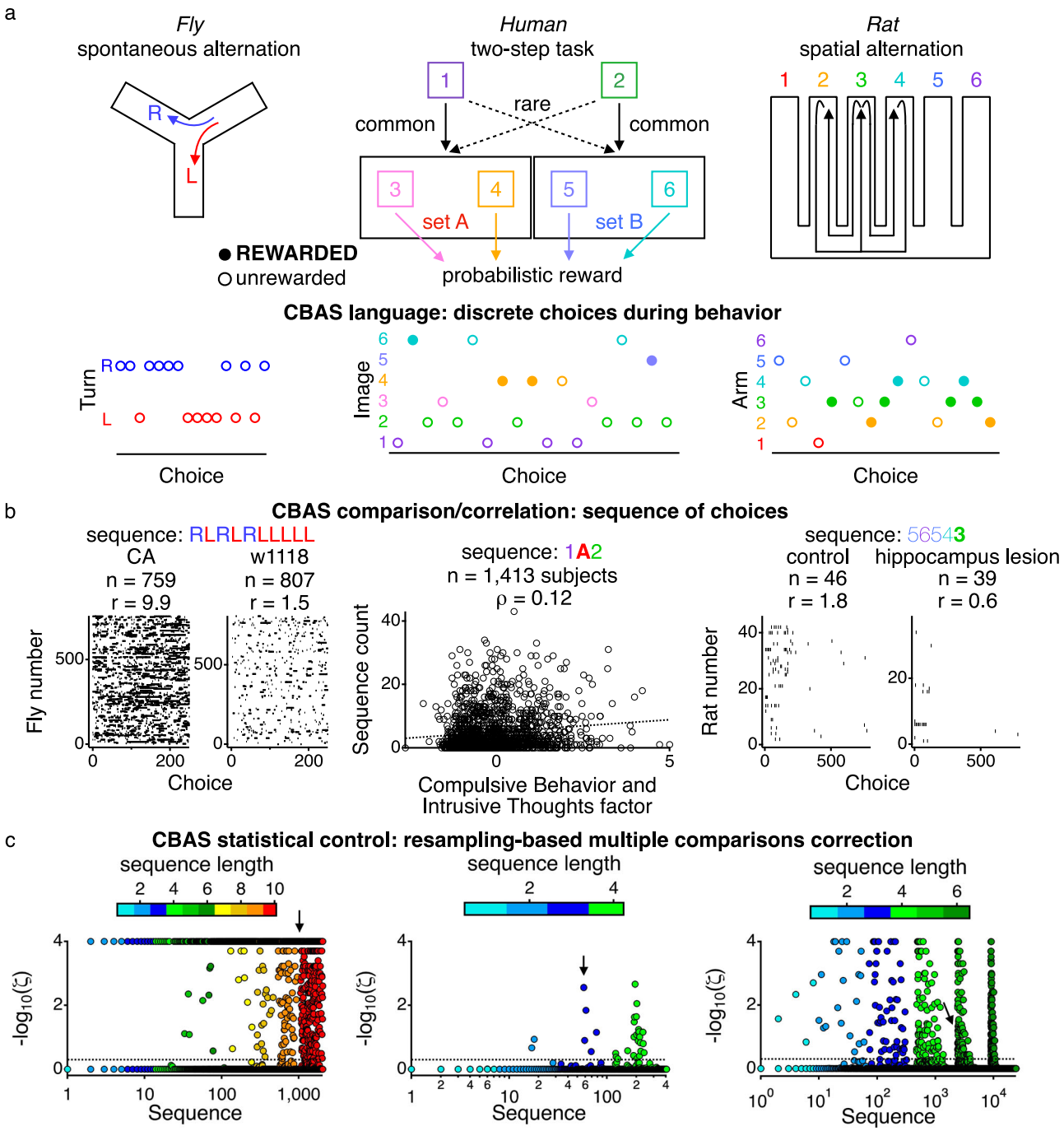


Fig. 1 | Choice-wide behavioral association study (CBAS) is applicable across diverse tasks and species. a CBAS represents behavior as sequences of discrete choices. Left: fly Y-maze showing left (red) and right (blue) turns (data from Buchanan et al.[31]). Middle: human two-step task showing example choice sequences (colors indicate image choices; data from Gillan et al.[43]). Right: rat spatial alternation task showing example arm choices (colors indicate arm identity). Rats are rewarded for alternating between outer arms (2 and 4) after every visit to the center arm (3; arrows). Flies receive no reward; filled circles and bold numbers indicate rewarded choices in the human and rat tasks. **b** CBAS evaluates all choice sequences up to a specified length and compares sequence usage between groups (flies and rats) or correlates sequence usage with a covariate (humans), while correcting for multiple comparisons. Left: example 10-turn sequence used more frequently by CA than w1118 flies (9.9 vs. 1.5 occurrences per animal; $\zeta = 1 \times 10^{-4}$). Middle: usage of sequence 1**A**2 is positively correlated with Compulsive Behavior and Intrusive Thoughts (CBIT) score ($r = 0.12$, $\zeta = 2.8 \times 10^{-3}$); dotted line shows the linear fit. Right: example arm-choice sequence 5654**3** occurs more frequently in control than hippocampal-lesioned rats (1.8 vs. 0.6 occurrences per animal; $\zeta = 0.047$). **c** Manhattan plots showing significance for all sequences evaluated in each dataset. Points are colored by sequence length and ordered by sequence frequency within each length. The horizontal dotted line indicates the significance threshold ($\gamma = 5\%$ median false discovery proportion control; $\zeta = 0.5$). The Romano-Wolf method for multiple comparisons correction was used to calculate the adjusted p-values (ζ) which were estimated using 10,000 resamples (making the maximum possible value for the y-axis be 4.00004); arrows indicate the example sequences shown in **b**.

https://doi.org/10.1038/s41467-026-76681-3

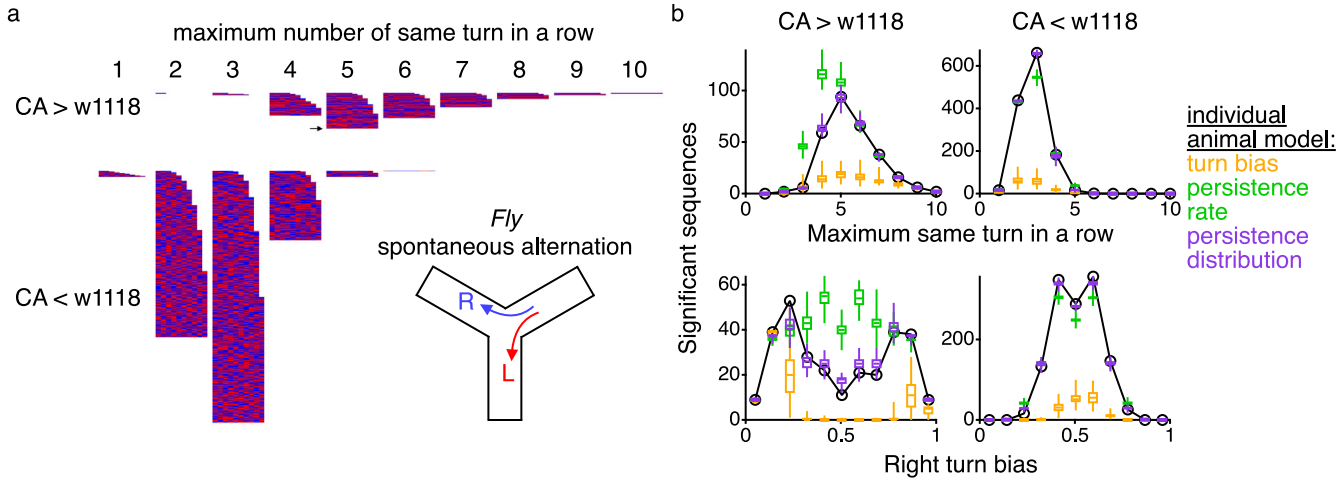


Fig. 2 | CBAS identifies interpretable differences between fly strains. CBAS was applied to spontaneous alternation behavior in flies as they made left and right turns on a Y-maze. **a** All sequences up to length 10 that differed significantly between the CA and w1118 strains are shown, with left turns in red and right turns in blue. Sequences used more frequently by CA are shown above, and those used more frequently by w1118 below. Columns are organized by persistence, defined as the maximum number of consecutive turns in the same direction within each sequence. The example sequence from Fig. 1b is indicated by an arrow. **b** Numbers of significant sequences identified by CBAS for the data (black) and three generative models of fly behavior (colored box plots). The turn bias model generates choices from each fly's overall left/right turning bias; the persistence rate model from each fly's overall tendency to continue turning in the same direction; and the persistence distribution model from each fly's full distribution of run lengths in the same direction. The first 250 turns are used for all models. Top panels group sequences by persistence (as in **a**); bottom panels group sequences by the fraction of right turns. Left and right columns quantify sequences occurring more frequently in the CA and w1118 strains, respectively. Model results are based on 100 simulated datasets. Boxes indicate the median and interquartile range; whiskers indicate the full range.

CBAS identifies interpretable differences for fly spontaneous alternation

CBAS identifies many sequences that significantly differ between the CA and w1118 fly lines. Upon inspection of all of the significantly different sequences, a clear, interpretable, contrast is apparent between those sequences that occur more in the CA line and those that occur more in the w1118 line (Fig. 2a). The CA line utilizes sequences with extended numbers of the same turn in a row, whereas the w1118 line utilizes sequences with more frequent changes in the direction of turning. Therefore, CBAS not only identifies that there is a difference between the two fly lines, but it also provides information to support an interpretation as to the nature of the difference.

The paper from which this open-source data comes from highlights the handedness of the individual flies, whereby individual flies show a propensity towards either right or left turns[31]. Such an observation creates a straightforward testable hypothesis about the nature of the difference between the two lines, i.e., that the lines differ in the extremes of their handedness bias, with individual CA flies hypothesized to being more extreme in their left/right turn bias than the w1118 lines. This hypothesis can be formalized by randomly generating virtual twin subjects by sampling turns according to the left/right turn biases of each individual fly. CBAS can compare this turn bias model of the lines, through generating an expectation for what its output would look like if the lines solely differed based on the left/right turn bias (Fig. 2b). The CBAS results from the data are inconsistent with this hypothesis (judged by the numbers of sequences that are significantly different in two transparent metrics).

Instead of modeling the choices of the individual flies as being generated by a left/right bias on each turn, we could instead generate the turns based on the bias of the individual flies to persist in their direction of turning (i.e., turning in the same direction as their previous turn). If we generate virtual twin subjects by sampling turn directions according to these one-step persistence biases of each individual fly—persistence rate—the resulting CBAS output is closer to the data than the left/right turn bias but still is inconsistent overall (Fig. 2b). However, if the choices of virtual twin flies are generated by using the distributions of persistence for each of the number of turns in a row from each individual fly—persistence distribution—then the CBAS of that model almost perfectly matches the CBAS output of the data according to our metrics (Fig. 2b). The virtual twins generated with the persistence distribution of the flies differ from the virtual twins generated with the persistence rate of the flies by utilizing the full distribution of their tendency to continue in the same direction for each possible number of turns instead of just the average persistence rate of each animal. This means using the full distribution instead of just the central tendency of this distribution. The better fit of the virtual twins using the full persistence distribution indicates that there is meaningful information for the behavior of the flies beyond just their central tendency to persist in turning in the same direction.

Graceful decay of CBAS output with decreasing group size

We exploited the very large numbers of flies in the spontaneous alternation dataset to understand how the output of CBAS behaves with different sample sizes. This determines our ability to make conclusions and derive interpretations from fewer subjects. We first evaluate the number of significant sequences identified from smaller group sizes as a fraction of the total number of significant sequences identified with the full dataset (Fig. 3a). As expected, with smaller group sizes, we find fewer overall significantly different sequences; however, across the range of group sizes evaluated, the median fraction of sequences was larger than the fraction of the population being used in the smaller group CBAS (Fig. 3a). This indicates that at smaller group sizes, the proportion of sequences identified by CBAS for these subjects grows faster than the proportion of subjects in the CBAS. This provides a rapid increase in the amount of information provided by CBAS as the sample size increases.

We next evaluate the fraction of the sequences identified by CBAS for the smaller sized groups that are not identified in the CBAS on the full dataset (which we might interpret as false positives). Since we control the median false discovery proportion at 5% and not the familywise error rate, we do not expect this value to be zero; however,

https://doi.org/10.1038/s41467-026-76681-3

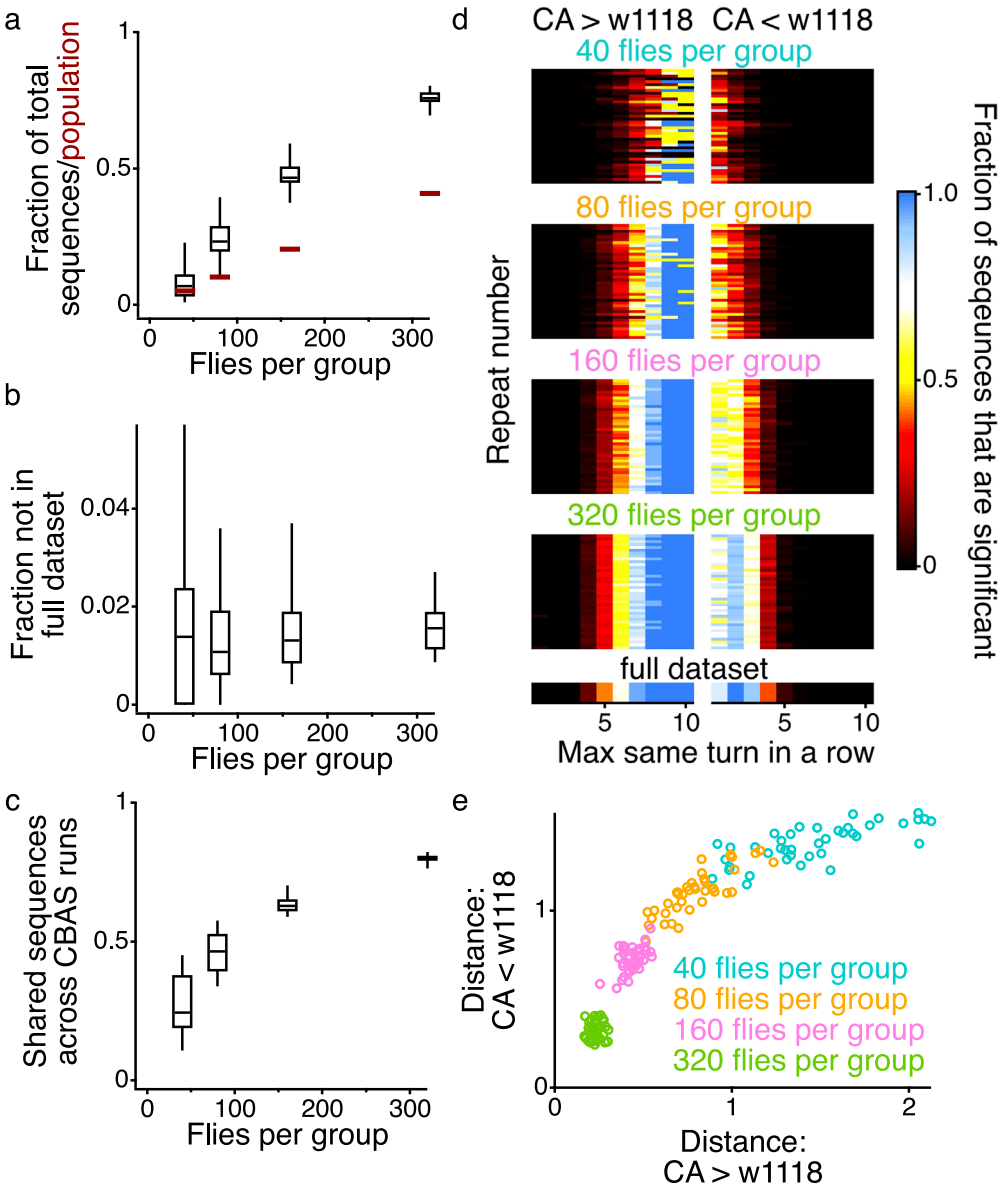


Fig. 3 | Scaling of CBAS with sample size. CBAS was applied to random subsamples of the fly dataset (40 repeats per sample size, sampled without replacement) to assess the effect of sample size on performance. **a** Fraction of significant sequences identified in each subsampled dataset that were also identified in the full dataset. Horizontal lines indicate the corresponding fraction of flies included in each subsample. **b** Fraction of significant sequences identified in each subsample that were not significant in the full dataset. **c** Overlap in significant sequences identified from paired, nonoverlapping subsamples of equal size. **d** Significant sequences grouped by persistence (maximum number of consecutive turns in the same direction). Values indicate the fraction of significant sequences within each category, facilitating comparison across sample sizes. The bottom row shows the full dataset. **e** Euclidean distance between each subsampled result in d and the corresponding result from the full dataset. Colors denote sample size. For **a**–**c**, boxes indicate the median and interquartile range, and whiskers indicate the full range across resampled datasets.

it was consistently small across all the group sizes evaluated (Fig. 3b). The medians across the different repeats of the same sample sizes across the different sample sizes are all less than 2%, and the values do not actually get grossly larger with the smaller sample sizes. This is a testament to, and a good illustration of, the control provided by the Romano-Wolf method for multiple comparisons correction.

We evaluated the similarity of the sequences identified by CBAS with nonoverlapping sets of subjects with the same group size (Fig. 3c). At smaller groups sizes (40 per group), this overlap could be quite low (median 24.4%), even though there was sufficient evidence to discriminate the strains (Fig. S1). This means that the specific sequences identified by CBAS can be quite variable from one experiment to the next, especially at smaller group sizes. Given that, we sought to understand if the structure of the sequences comported with the conclusion from the full dataset—that the CA line uses more of the same turns in a row and the w1118 line more frequent changes in direction of turn. Even though the overlap between CBAS on different groups might not be that high (Fig. 3c), the sequences identified still follow the same structure as the full dataset in regard to number of the same turns in a row (seen across the repeats for each group size in Fig. 3d). As the group size increases, the output of CBAS becomes more consistent, and more similar to the results from the full dataset (Fig. 3e), licensing more generalized conclusions.

We additionally took advantage of the large group size to evaluate the sensitivity of the conclusions to the choice of maximum sequence length evaluated by CBAS. Using only a maximum sequence length of 5

choices supports the same conclusion found with using a maximum sequence length of 10 choices (Fig. S2a). However, with a maximum sequence length of 3 or 2 choices it is no longer possible to separate between the persistence rate and persistence distribution models, even though such short sequence lengths still rule out the turn bias model (Fig. S2b).

CBAS can also correlate sequences of choices with a covariate of interest

The second task to which we apply CBAS is the two-step task, developed to test the interplay between model-based and model-free (reinforcement) learning algorithms[41] (Fig. 1a; middle). In this task, subjects choose between pairs of images at two different stages. The image choice at the first stage stochastically governs the pair of images from which the subject can choose at the second stage, and which enjoy drifting individual probabilities of delivering reward when picked. Model-dependent analyses of variants of this task (which make inferences based on the stochasticity in the transition between first and second stages) have led to conclusions about the algorithms used in different brains regions[41,42] and differences that underlie psychiatric symptoms[43], notably compulsivity. However, the interpretations are not without controversy[44], and the complexity of behavior makes it difficult to evaluate ways in which the model might be missing features in the data.

Data for the two-step task come from a large population of human subjects in the open-source dataset from Gillan et al.[43]. We applied CBAS to all sequences of up to 4 choices long over the course of the 200 trials−400 total choices−of the task. Following Akam et al.[42] we collapse image 3 and 4 into set A and image 5 and 6 into set B because the specific images within the set are not relevant for the critical decision at the first stage. Choices that are rewarded are treated as distinct from those that are not rewarded, as those choices can lead to different behaviors. That means that the sequence 2**A**1 is different from 2a1, where the upper-case emboldened letter indicates that a reward was received. We chose to limit CBAS to a maximum of only 4 choices long as that length of sequence captures the critical transition for this behavior between a reward at the end of the prior trial and the choice at the first stage of the next trial.

In building the two-step task dataset, the authors sought to examine how behavior varied along continuous trait dimensions. They did this by performing factor analysis on the answers to a series of psychiatric symptom questionnaires and showing that the way subjects performed the two-step task was correlated with one of the factors (factor 2) that was associated with compulsive behavior and intrusive thoughts (CBIT)[43]. Therefore, we extended CBAS so that it could also evaluate such correlations. We applied CBAS to the correlation between the sequence usage for each subject and their CBIT score (Fig. 1b), finding 31/408 sequences to be significant (Fig. 1c).

CBAS provides evidence for testing model-dependent analyses of the two-step task

To understand the output of CBAS for the two-step task, we review the expectation associated with the hypothesis-driven analysis of this experiment. The two-step task was designed to evaluate the interplay between model-based and model-free decision making[41] based on the stochastically rare transitions (from image 1 to set B; and image 2 to set A). Model-based decision making develops an understanding of the structure of the world (which is the model) and makes choices based on that understanding. Model-free decision making makes choices based on reinforced past successful actions, without developing an understanding of the structure of the world. A way to see the difference between these two decision-making schemes is when a learner chooses image 2 at the first step, makes an unlikely stochastic transition to set A, and gains a reward at this second step ('2**A**' in our language). A regular model-free learner will reinforce the choice of image 2 (along with the image chosen at A) and will therefore be more likely to choose image 2 on the next trial. By contrast, a model-based learner will be more likely to choose image 1 on the next opportunity, based on the observation that this will make the transition to set A more likely (because of the common/rare transition structure in the environment)[41].

We evaluated all of the sequences that CBAS found as either significantly positively or negatively correlated with CBIT (Fig. 4a). CBAS identified a positive correlation with many sequences involving being rewarded at **A** and then selecting image 2 or being rewarded at **B** and then selecting image 1. In particular, the structures 1**A**2 or 2**B**1 occur commonly in the positively correlated sequences. These behavioral motifs are inconsistent with either model-free or model-based learning, as the subjects get rewarded after choosing from the common side (A from 1 or B from 2) but then select the image that will rarely bring them back to the previously rewarded side (2 from A or 1 from B). The model-based decision maker would return to 1 as that is the side that would commonly go to set A where there was the reward. The model-free decision maker would also return to 1 as that was the side that was reinforced as being connected to the recent reward at A.

We next used CBAS to test whether the combined model-based/model free reinforcement learning (MB/MF RL) model used by Gillan et al.[43] which has parameters describing model-based and model-free decision making and the propensity to employ each, correctly captures the data. To do this, as with the fly data, we generated virtual twin choices for each subject based on fits of each subject to this model and ran CBAS on the model output (Fig. 4b). For the human data, the virtual twin for each subject had choices generated by the model using the best fit parameters for that subject. This led to a positive correlation between CBIT and the model inconsistent sequences with the choices 1**A**2 and 2**B**1, like the human data. However, there was the important difference in that CBAS identified a positive correlation with the unrewarded versions of those units (1a2 and 2b1) that was not present in the human data.

CBAS therefore identified an important failing in the MB/MF RL model. We can explain this by noting that this model has a parameter for "stickiness"−the propensity to continue to choose the same of the first images (1 or 2) in a row. Decreased stickiness would also comparatively favor sequences that change the first image, including 1**A**2, 1a2, 2**B**1, and 2b1. Indeed, in our fits, along with the negative correlation between the propensity to employ model-based choices and CBIT ($\rho = -0.08$; $p = 1.7 \times 10^{-3}$) reported by Gillan et al.[43], there was also a negative correlation between the stickiness parameter and CBIT ($\rho = -0.11$; $p = 1.6 \times 10^{-5}$).

As an alternative, we generated behavior based on just model-inconsistent choices. That is, we created a virtual twin of each subject s from the MB/MF RL model by hybridizing the rate with which that subject performed model-inconsistent choices with choices that were otherwise generated according to the fit parameters for a single, reference subject s*. The reference subject s*, was randomly chosen for each of the 50 repeats. That is, we used the MB/MF RL model based on s* to generate data, but every time the model produced 1**A** or 2**B** we probabilistically selected the next choice according to the overall model-inconsistent rate of subject s (but then continued to generate choices as if nothing else had changed). This ensures that the majority of choices came from a subject with a single CBIT value, and only the choices following 1**A** or 2**B** were manipulated (19% on average). The CBAS results from this model-variant were much closer to the CBAS output of the data in the metric reported (shown in the 'Model inconsistent' results in Fig. 4b). Furthermore, when we fit the output of the model-inconsistent variant of the model to the MB/MF RL model, we found that a positive correlation between model-inconsistent choices and CBIT alone can appear as a negative correlation between the model-based parameter of the model and CBIT (Fig. 4c). To understand this, take the case that the CBIT values for the subjects

https://doi.org/10.1038/s41467-026-76681-3

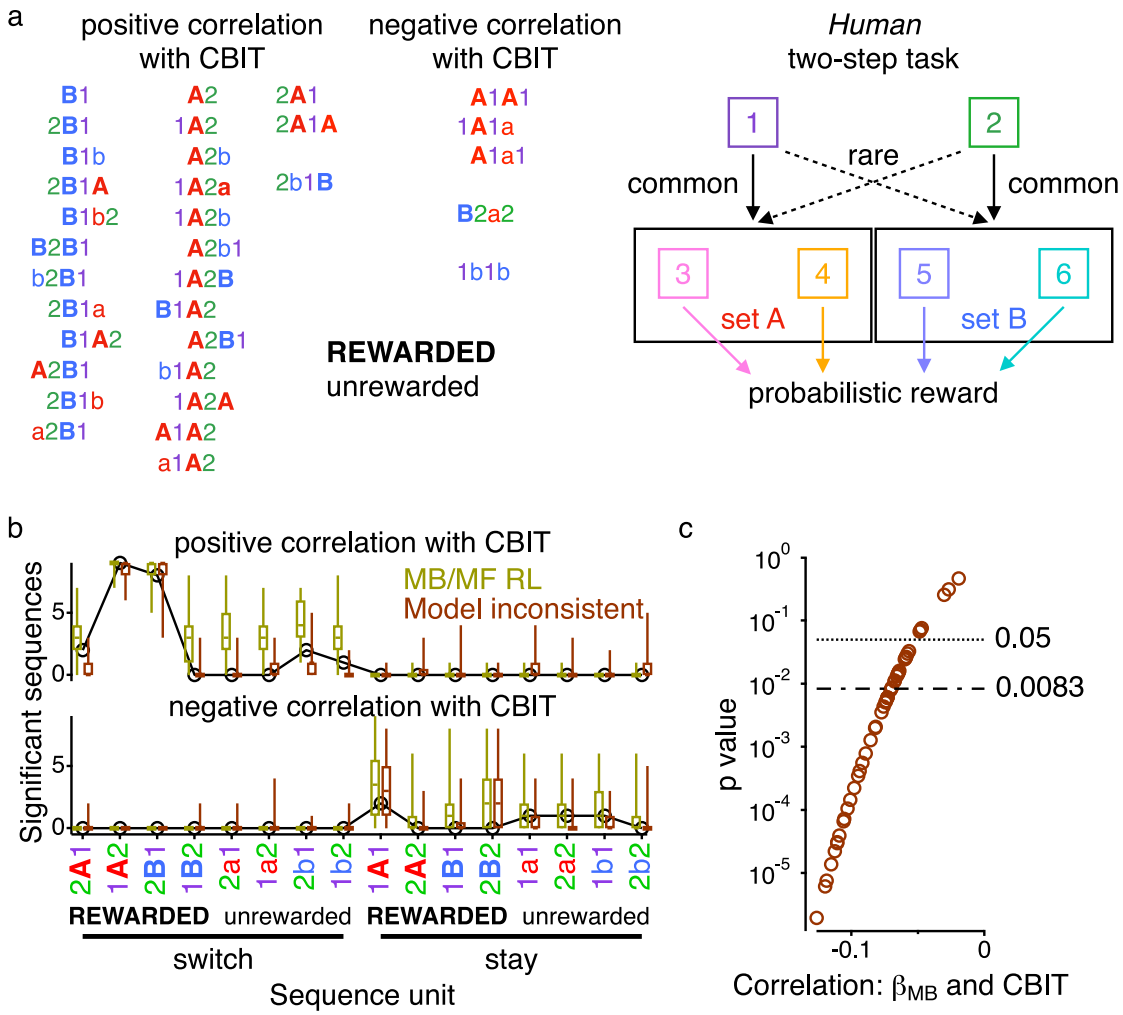


Fig. 4 | CBAS identifies behavioral features associated with compulsivity in the two-step task. CBAS was applied to human choice sequences (up to length 4) from the two-step task and correlated sequence usage with the compulsive behavior and intrusive thought (CBIT) factor score. **a** Sequences whose usage was positively (left) or negatively (right) correlated with CBIT. Colors indicate the stages of the image sets. Uppercase bold letters denote rewarded choices. Sequences positively correlated with CBIT predominantly contain the model-inconsistent motifs **A2** and **B1**. **b** Numbers of significant sequences identified by CBAS for the data (black) and two generative accounts (box plots). The model-based/model-free reinforcement learning (MB/MF RL) account generates choices using subject-specific fitted parameters. The model-inconsistent account selectively reproduces each subject's tendency to make model-inconsistent choices while generating all remaining choices from a common fitted model. Sequences are grouped by the motifs indicated on the x-axis. Boxes indicate the median and interquartile range across 100 simulated datasets; whiskers indicate the full range. **c** Correlation between the fitted model-based parameter of the MB/MF RL account and CBIT after fitting the model-inconsistent simulated datasets. Horizontal dotted lines indicate significance thresholds of $p = 0.05$ and the Bonferroni-corrected threshold ($p = 0.0083$). A total of 50 model repeats was fit, 86% and 56% had p values < 0.05 and 0.0083, respectively, despite no manipulation of model-based planning. For each simulated dataset, p values were calculated using two-tailed linear correlation testing between the model-based parameter and CBIT.

were positively correlated with their rate of making model inconsistent choices, but did not differ at all in their model-based planning. Then, if we fit those subjects with the MB/MF RL model and used the hypothesis-driven analysis typically applied to the two-step task to interpret the results, we would see what amounts to a spurious negative correlation between model-based planning and CBIT.

CBAS identifies a hippocampal dependent strategy for spatial alternation learning

The third, and final task we evaluated with CBAS is spatial alternation behavior in rats. For this task, to get reward, the rats must alternate between pairs of arms of a track whilst visiting a different arm of the track in between (Fig. 1a; right). Spatial alternation is a common behavioral paradigm for phenotyping and neurophysiology, and the discretization of the behavior into the arms chosen by the animals forms the basis of many of the conclusions from these studies[45–49]. However, our recent work calls into question the assumptions and hypotheses that motivate the standard analysis for spatial alternation behavior[50,51]. Such limitations could even be general across other tasks used to evaluate memory[52]. Furthermore, even though we have developed many variants of reinforcement learning agents to fit individual behavior, these agents learned in ways that were very evidently different from the animals[51], limiting their use for phenotyping. Here, we use CBAS to evaluate the role of the hippocampus in this task by comparing rats with lesions to their hippocampus to rats that underwent control surgery.

For the spatial alternation task, we applied CBAS to all sequences of arm choices up to 6 choices long over the first 800 arm choices of the rats. As with the two-step task, rewarded choices are treated differently from unrewarded choices. As with the fly spontaneous alternation dataset, CBAS compared the rate of usage of each sequence between the populations of hippocampally-lesioned and control rats (Fig. 1b; right). 409/24,342 sequences differed significantly (Fig. 1c). We opted for sequences with a maximum of 6 choices long because

https://doi.org/10.1038/s41467-026-76681-3

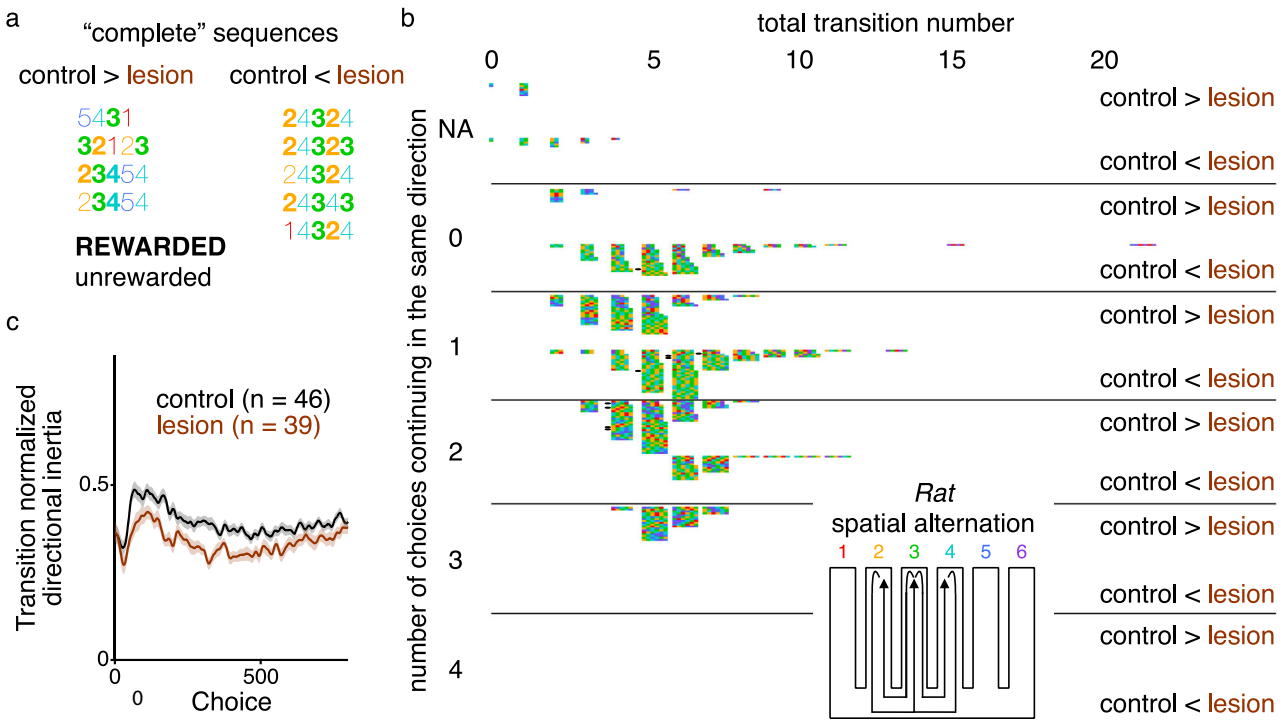


Fig. 5 | CBAS identifies a hippocampus-dependent behavioral strategy during spatial alternation. CBAS was applied to rats performing the spatial alternation task, using sequences of arm choices up to length 6. **a** "Complete" sequences occurring more frequently in control rats (left) or hippocampal-lesioned rats (right). A complete sequence is one whose longer, sufficiently frequent extensions differ in prevalence between groups in a consistent manner (Fig. S3). Bold numbers indicate rewarded choices. **b** All significant sequences grouped by directional inertia (number of choices continuing in the same direction; rows) and total number of transitions in the sequence (columns). Sequences are colored according to arm identity. Locations of "complete" sequences from (**a**) indicated by tic mark to the left of the sequence. For sequences 2 choices or less the directional inertia is not defined, and they are labeled as NA. **c** Transition-normalized directional inertia, a summary metric motivated by the CBAS results, plotted over the first 800 choices of the task. Control rats show significantly greater transition-normalized directional inertia than hippocampal-lesioned rats ($p = 4.0 \times 10^{-4}$, two-tailed permutation test), consistent with greater systematic movement through the track. Shaded regions indicate mean ± s.e.m. across animals.

sequences that long can contain all of the ways that rewards can be received for this task and provide rich structure for the many ways in which reward can be missed.

CBAS identified many sequences discriminating the fly strains. However, the dimensionality of the task is quite low, so it is fairly straightforward to structure all of the sequences to develop an interpretation (Fig. 2). By contrast, CBAS did not identify that many sequences in the two-step task that significantly correlated with CBIT, so the extra complexity and dimensionality of this task did not prevent us from directly evaluating all the sequences to develop an interpretation (Fig. 4). For the case of rat spatial alternation, CBAS identified a good number of sequences, and the task is complex and high dimensional, making it a challenge to directly develop an interpretation of the output of CBAS.

Therefore, we developed an analysis method that takes advantage of the structure of the sequences at the output of CBAS to help inform an interpretation of the spatial alternation behavior. CBAS evaluates all sequences up to a certain length, and so sequences with shared and overlapping structure might all be significant. We therefore defined the notion of a "complete" target sequence (Fig. S3) as one with the property that all sufficiently prevalent longer sequences that contain it are found by CBAS to differ significantly in the same way as the target (e.g., all are found to be greater in the control than the lesion rats). Effectively, a "complete" sequence has the property that no matter what happens before or after that sequence, the series of choices remain significantly different between the groups. Such sequences are a powerful way to simplify the complex behavior.

We find "complete" sequences that the control rats do more than the lesion rats and that the control rats do less than the lesion rats (Fig. 5a). Inspection of these sequences provides two axes to compare all of the significant sequences. The "complete" sequences that the control rats use more than the lesion rats show systematic exploration of the arms by largely continuing in the same direction and transitioning mostly to neighboring arms. These features are much less the case in the "complete" sequences that the control rats do less than the lesion rats. We used these two axes to compare all the significant sequences (Fig. 5b). Consistent with the finding for "complete" sequences, sequences that the control rats use more than the lesion rats tend towards higher directional inertia—the propensity to continue in the same direction—with smaller overall transition distance.

The hypothesis generated by the "complete" sequences and verified across all of the significant sequences, motivates a relatively straightforward metric by which we can evaluate all of the choices of the rats for comparison between the two groups. We quantified the transition normalized directional inertia in the control and lesion rats. This metric measures, for each choice, whether that choice continues in the same direction as the preceding choice and normalizes it by the number of arms transitioned. For example, choosing arm 3 after coming from 2 and 1 (sequence 1-2-3) has a transition normalized directional inertia of 1, choosing arm 4 after coming from 2 and 1 (sequence 1-2-4) has a transition normalized directional inertia of 0.5, and choosing arm 1 after coming from 2 and 1 (sequence 1-2-1) has a transition normalized directional inertia of 0. Lesion and control rats show a clear and significant difference in their transition normalized directional inertia (Fig. 5c).

https://doi.org/10.1038/s41467-026-76681-3

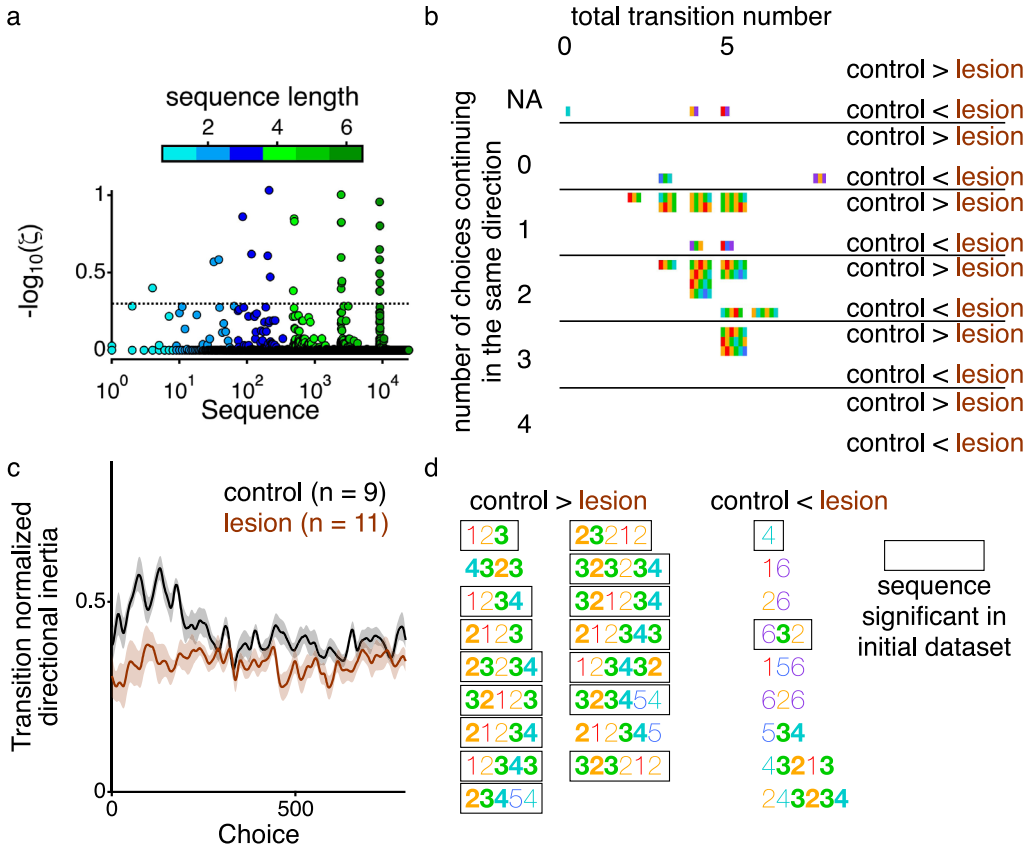


Fig. 6 | Replication of the CBAS-derived behavioral hypothesis in an independent cohort of rats. CBAS was applied to a separate small cohort of control and hippocampal-lesioned rats performing the spatial alternation task. **a** Manhattan plot showing the significance of all sequences compared between groups. Points are colored by sequence length and ordered by sequence frequency within each length, as in Fig. 1. **b** Significant sequences grouped by directional inertia and total transition distance, as in Fig. 5b. The replicated sequences show the same overall organization as the original dataset. **c** Transition-normalized directional inertia over the first 800 choices of the task. Control rats show significantly greater transition-normalized directional inertia than hippocampal-lesioned rats ($p = 2.2 \times 10^{-3}$, two-tailed permutation test), replicating the result from Fig. 5. Shaded regions indicate mean ± s.e.m. across animals. **d** Significant sequences occurring more frequently (left) or less frequently (right) in control rats relative to hippocampal-lesioned rats. Boxes indicate sequences that were also identified as significant in the original dataset.

Although "complete" sequences were not used to interpret the fly and human CBAS outputs, evaluating them would lead to identical conclusions (Fig. S4).

CBAS generated hypothesis replicates for spatial alternation

The fact that the transition normalized directional inertia is significantly different between the control and lesion rats, is partly a circular claim. We motivated the metric based on the sequences identified by CBAS as differing between the groups. It did not have to be the case that the metric would differ across all the choices of the animals (as opposed to just the significant sequences), but it should not be overly surprising that it does. Therefore, we sought to replicate our understanding in a separate dataset of hippocampal lesion and control rats. Importantly, the data from these new lesion and control rats were neither a part of the initial CBAS evaluation nor the comparison of the groups via the transition normalized directional inertia metric.

Even with a smaller sample size, CBAS identifies differences due to the hippocampal lesioning (Fig. 6a), consistent with the power calculation on the initial dataset (Fig. S1e&f). The sequences identified by CBAS separate along the two axes identified by the "complete" sequences from the initial dataset (Fig. 6b). This comportment of the structure of the sequences with a smaller sample size is consistent with what was found with subsampling the fly data (Fig. 3d). Consistent with the hypothesis generated from the CBAS-motivated understanding of the initial dataset, the groups significantly differ in their transition normalized directional inertia (Fig. 6c). Indeed, we even replicated specific sequences across the two datasets (Fig. 6d). Interestingly, a much larger fraction of the sequences that the control rats use more than the lesion rats replicate compared to the sequences the control rats use less than the lesion rats. This is consistent with the overall hypothesis that the hippocampus enables a learning strategy that systematically engages with the arms of the track through moving in the same direction and transitioning to nearby arms. Without a hippocampus that systematic behavior goes away, exposing less systematic choices, that inevitably will be less likely to replicate between samples.

Discussion

We have presented a general data-driven behavioral analysis method, CBAS, for identifying interpretable behavioral components that is grounded in sequences of choices made by subjects. There has been significant progress, in recent years, in tracking and analyzing, short timescale actions of subjects[7–16]; however, we lack methods for generally analyzing and interpreting sequences of these actions or choices and decisions of subjects that unfold over a longer period than moment to moment movements. CBAS provides just such a method and is applicable across a wide array of species and different behavioral paradigms (Fig. 1). It can be used to test models and hypotheses (Figs. 2, 4), and, for instance, using generalizable simple principles

about the relationship between sequences, it can generate hypotheses in complex behaviors where reliable computational understanding has yet to emerge (Figs. 5, 6). Through taking advantage of large-scale data collection and rigorous statistical methods, CBAS has the potential to transform our use of behavior in a comparable way to how GWAS/WES/WGS changed the paradigm for genomic studies.

Somewhat surprisingly, in all tasks, CBAS provides information and generates interpretations that are either inconsistent with, or unexpected from, the prior understanding of these data and tasks. For the flies, as mentioned above, the dataset came from a paper characterizing the existence of handedness in these flies, as manifested by extremes in individual preferences for left or right turns[31]. The output of CBAS on these data allowed for the testing of generative models to explain that handedness. The most straightforward model that one could postulate based on the hypothesis of handedness is not consistent with the data (Fig. 2b); whereas a model drawing upon the distribution of individual flies to persist in the same direction for a range of the same turn in a row was almost indistinguishable from the data according to our metrics. Therefore, CBAS generates testable hypotheses as to underlying computational structure that these flies draw upon as they perform this behavior.

For the two-step task, the CBAS results provide an alternative interpretation for the relationship between the task and individual CBIT levels. Previously, these data were interpreted to mean that individuals with more CBIT are less likely to draw upon model-based planning[43]. The CBAS results indicate that the relationship with CBIT rests more on the increase of model-inconsistent choices with CBIT (Fig. 4a, b), and that that increase alone could appear as a negative correlation with model-based planning, even if there was in reality not one (Fig. 4c). What an increase in model-inconsistent choices means in the context CBIT is something that should be explored in the future. Possible ways in which model-inconsistent choices positively correlate with CBIT include an increase in exploration in the context of the lower task success that has been shown to associate with OCD[53]. That lower task success could lead to failures in both model-based and model-free decision making, which then manifest as more model inconsistent choices. Another route would be an increase in a desire for symmetry that can be a prominent component of OCD[54]. That desire for symmetry could lead to an evening out of choices to both of the first images independent of the potential for success of those choices. These would then naturally be model inconsistent.

The result that the hippocampus is involved in spatial alternation behavior is not a new one. There have been prior reports of changes to spatial alternation behavior in the context of hippocampal lesioning[45] and manipulation[48,55]. However, all of the prior work occurred from the prevailing perspective of spatial alternation exclusively reflecting memory allocation, something we have recently shown to be inconsistent with the way in which the rats learn[50,51]. Guided by the results of CBAS, we find that the hippocampus is responsible for implementing a specific strategy (or set of strategies) whereby the animals systematically move across the track and transition between closer arms as they learn. This strategy is far more consistent with the hippocampus implementing a structuring of the task space[56] as described in recent models of hippocampal function[57,58] and reported in recent experimental results across species[59–61], than with the typical memory function assumed in the prior studies of spatial alternation behavior.

These three tasks were just the first ones to which we applied CBAS; they were not selected from a larger set of tasks and used for illustrative purposes. We strongly anticipate that there will be comparable benefits to applying CBAS and other data-driven methods to a broad array of datasets and behavioral tasks to provide a stronger foundation for our understanding of behavior. The ability of CBAS to identify differences or correlations in a dataset is directly built into its function. However, the ability to interpret the differences or correlations found remains nontrivial. We have provided examples of ways to extract interpretable information from the sequences output by CBAS and have developed a general analysis ("complete" sequences) to assist in that process. However, there remains room for future work to assist in going from the identification of a difference or correlation by CBAS to understanding and interpreting what those differences and correlations mean. Of course this is in no way a unique problem to CBAS. Genomics methods (GWAS/WES/WGS) similarly require secondary and follow up analyses.

Considerations for the application of CBAS to datasets

To apply CBAS to a dataset, there are some decisions that need to be made. At a fundamental level, CBAS requires a discretized description of the behavior. This was motivated by the common use of choices to describe behavior (Fig. 1a). Furthermore, there is evidence that, at a rather fundamental level, behavior occurs through discrete choices[12,62]. CBAS is not currently capable of working on continuous data. For some continuous data, this limitation can be overcome by imposing discretization or tokenization of the data. For example, if the behavior measures the continuous movement of an animal in space, as in an elevated plus maze[63], then one can discretize the location of the animal in units of time, or a string of discretized locations independent of the time spent in those locations. As another example, for acoustic startle and prepulse inhibition[64] the continuous startle measurement could be discretized into a range of startle magnitudes. For other types of continuous data, such artificial discretization might not be ideal. What CBAS then requires is a metric to determine when two continuous behavioral trajectories are sufficiently close to be counted as the same. In future work, it will be interesting to explore methods for creating and validating such metrics.

The discretization informs the language that is used to make up the sequences that form the basis of the statistical evaluation within CBAS. In deciding on the language there is a range of biases that can be imposed with different decisions. For example, with the two-step task, we combined the second image choice into sets A or B. This prevents CBAS from differentiating a difference in choices at that level. Based on the literature we deemed that bias to be justified, but it is a bias, nonetheless. It is critical to remember and consider the biases imposed by the language when interpreting the results of CBAS.

A next decision to make is the length of sequence up to which CBAS will evaluate. The longer the sequences, the more information CBAS can provide, but the more data and computational resources that will be necessary. If the language is large, i.e., there are many things that can happen in the behavior at any point in time, and the transition matrix between one choice and another is not sparse, then computational resources might not be feasible to calculate the statistics underlying CBAS, especially if longer sequences are evaluated. Additionally, it is necessary to specify a criterion for the total number of choices per animal or session that are decomposed to form the sequences that are compared. Although it is possible just to include all the data, this might introduce bias if the groups have different total numbers of choices overall. Conversely, one might try and study aspects of learning or the acquisition of competence by applying CBAS to restricted portions of the behavior. It would also be possible to use a discrete language that characterizes aspects of behavior that unfold over longer timescales (for instance, defining a discrete language using unsupervised clustering over windows of choices), or indeed to employ a multi-scale description.

Finally, CBAS either needs to compare groups or correlate with a covariate of interest. In this work, when comparing groups, we focused on groups of animals within the same species; however, CBAS could just as easily compare the same behavior across different species[65], enabling data-driven information for cross species insights. Currently, CBAS is not able to evaluate just a single group of animals by itself in an unsupervised manner without a covariate of interest. If there is only a single group of subjects with no experimentally defined covariate, but a regular clustering algorithm detects structure, then CBAS can be run on the output of that.

In sum, we offer CBAS as a data-driven analysis technique that extracts statistically rigorous differences in behavioral sequences between or among groups. We show that such sequence distinctions can be revealing about the underlying information processing functions that vary and can circumvent biases that plague model-dependent analysis methods. We illustrated the properties and promise of CBAS in three different datasets in three different species.

Methods

Romano-Wolf resampling based multiple comparisons correction

We follow the terminology and description laid out in Clarke et al.[66] to describe the Romano-Wolf multiple comparison correction. First, we describe the way the method corrects for the familywise error rate (FWER) and then explain how the procedure is extended to provide median control of the false discovery proportion. FWER control at a level of α means that across all comparisons there is a α percent chance of having at least one false positive rejection of a null hypothesis. The Romano-Wolf procedure provides FWER control through resampling the data[39]. We describe the case of testing a total of S hypotheses.

It is not generally known if the Romano-Wolf procedure controls for type III, or directional, errors. Type III errors are errors in the sign, or direction, of the conclusion. For example, if a statistical test provided information to reject the null hypothesis $\theta_1 = \theta_2$, and you then concluded that $\theta_1 > \theta_2$, when in fact $\theta_1 < \theta_2$. Therefore, instead of running a single two-tailed test for each sequence, we run two one-tailed tests for each sequence. As a consequence, the total number of hypotheses tested, S, is twice the total number of sequences being compared. For CBAS those hypotheses take one of two forms: (i) the rate of each sequence (r_s) is the same between two groups $\Delta r_s = 0$, or (ii) that there is no correlation (ρ_s) between each sequence and a covariate of interest, $\rho_s = 0$. For the one-tailed versions of each hypothesis, we ask if $r_{s_1} - r_{s_2} > 0$ and $r_{s_2} - r_{s_1} > 0$ for the comparitive CBAS, where r_{s_N} is the rate of sequence s for group N, or if $\rho_s > 0$ and $\rho_s < 0$ for the correlational CBAS.

The first step in the procedure is to create a studentized test statistic for each hypothesis. The studentization is different based on whether the CBAS is comparing two groups or calculating a correlation. In the case where two groups are being compared:

$$t_s = \frac{\Delta r_s}{\sigma_s} \tag{1}$$

where σ_s is the standard error of Δr_s, which we calculate by combining the standard error of the mean of the rate for each group using error propagation, i.e. $\sigma_s = \sqrt{{\sigma_{s_1}}^2 + {\sigma_{s_2}}^2}$, where σ_{s_N} is the standard error of the occurrence rate of sequence s for group N.

In the case where the correlation is being calculated the studentized test statistic is[67]:

$$t_s = \frac{\sqrt{n}\hat{\rho}_n}{\hat{\tau}_n} \tag{2}$$

where:

$$\hat{\rho}_s = \frac{\sum X_i Y_i - n\bar{X}_s\bar{Y}}{\sqrt{\sum (X_i - \bar{X}_n)^2 \sum (Y_i - \bar{Y})^2}} \tag{3}$$

$$\hat{t}_s = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_s)^2 (Y_i - \bar{Y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_s)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4}$$

For Eqs.(2), (3), and (4), n is the number of subjects for which the correlation is being calculated, X_i and Y_i are the values of the metrics being correlated (in our case, the sequence count and CBIT score for each individual respectively), $\bar{X}_s$ is the mean sequence count for the specific sequence being considered, and $\bar{Y}$ is the mean of the covariate of interest (CBIT score, which is the same for any sequence).

When two groups are being compared, we resample from the entire population with replacement (separately for each group) and build up a null distribution by bootstrapping M times. The test statistic from the m^{th} bootstrap sample for $m = 1, \ldots M$ is:

$$t_s^{*,m} = \frac{\Delta r_s^{*,m}}{\sigma_s^{*,m}} \tag{5}$$

where, $\Delta r_s^{*,m}$ is the difference in the rate of each sequence whilst resampling, with replacement, from the entire population, ignoring the group labels. $\sigma_s^{*,m}$ is the accompanying bootstrap standard error of the difference in means. The resampled group sizes are the same as the two groups of interest.

In the case where the correlation is being calculated, the test statistic based on the m^{th} bootstrap sample is:

$$t_s^{*,m} = \frac{\sqrt{n}\hat{\rho}_s^{*,m}}{\hat{\tau}_s^{*,m}} \tag{6}$$

where, $\hat{\rho}_s^{*,m}$ is the correlations of each sequence whilst permuting the relationship between the metrics being calculated across the entire population. $\hat{\tau}_s^{*,m}$ is the accompanying normalization for the permuted relationship. The permuted group size is the same as the original.

For the main CBAS calculations, we used a value of $M = 10,000$. When we run many repeats of different models or for power calculations, we use a value of $M = 1000$. Importantly, for each individual resampling, m, the same resampled set is used for all sequences.

The test statistics, and their accompanying estimators are ordered from largest to smallest values. This creates an $M \times S$ matrix where each column contains all the estimators of the test statistics. The first column contains the estimators from the largest test statistic, the second column contains the estimators from the second largest, etc.

To define the distribution for which each test statistic is compared, which then determines the adjusted p-value, the following algorithm is used. The first sequence considered is the one with the maximum test statistic, t_s. Its comparison distribution is defined as the maximum value within each row of the matrix of estimators of the test statistic:

$$\max(t^{*,m}) = \max\{t_1^{*,m}, \ldots, t_S^{*,m}\} \tag{7}$$

which provides a total of M values, $t^{*,m}$ (there is no longer an association with s, because these values can come from a resampling of any of the sequences). Using those M values, the adjusted p-value is calculated as follows:

$$p_s^{adj} = \frac{\#\{\max(t^{*,m}) \geq t_s\} + 1}{M + 1} \tag{8}$$

After calculating p_s^{adj} for the first sequence, the column with the test statistic estimators generated from the first sequence is removed from the matrix. This now leaves a matrix that is $M \times (S - 1)$. The above procedure is then used to calculate p_s^{adj} for the sequence with the second largest test statistic, and then its column of test statistic estimators is removed, etc until all p_s^{adj} have been calculated. Following the algorithm described in Clarke et al.[66], prior to calculating the p-value for the next sequence, we enforce monotonicity of the p-values

by resetting the p-value for each sequence:

$$p_s^{adj} = \max\left\{p_s^{adj}, p_{s-1}^{adj}\right\} \quad (9)$$

To control the false discovery proportion (i.e., control the number of false positives divided by the total number of hypotheses rejected), the idea of k-FWER is introduced[40]. For control of the FWER, k is equal to 1, and that leads to an α percent chance that there is at least 1 false positive among all hypotheses rejected (for FWER control $\alpha = 0.05$ is most commonly used). If k equals 2, then there is an α percent chance that there are at least 2 false positives among all hypotheses rejected. Therefore, to get control of the false discovery proportion we need to find the k that provides the proportion of interest given the number of hypotheses rejected. So, if we want a false discovery proportion, γ, of 0.05, we need $k \sim 0.05\times$ number of hypotheses rejected.

Romano and Wolf also derived an algorithm to do just that. The algorithm is as follows. Start with $k = 1$. Apply the k-FWER procedure, and note the total number of hypotheses rejected, N. If $N < \frac{k}{\gamma} - 1$, stop and you have identified k. Otherwise, increase k by 1, and repeat[40]. The way you determine k-FWER is in Eq. 7, instead of taking the maximum value in each row, you take the k^{th} largest value. Finally, to get median control of the false discovery proportion, $\alpha = 0.5$. This means that 50% of the time you will get a value greater than γ, and 50% of the time you will get a value less than γ, leading to median control of the false discovery proportion[40]. This is a similar decision to what is done when calculating the false discovery rate (FDR) with Benjamini-Hochberg or Benjamini-Yekutieli, except the false discovery rate controls the mean of the false discovery proportion instead of the median.

Lastly to avoid confusion around what the statistical control is we label the adjusted p-value with the k-FWER control (i.e., median control of the false discovery proportion) as ζ instead of p, comparable to the use of q with FDR control.

Power estimation for sample size calculation with CBAS

To estimate the statistical power of CBAS for a given sample size (Fig. S1) we resampled the dataset without replacement and ran CBAS to determine the number of significant sequences. For each sample size we performed 20 repeats. We also ran a CBAS comparing each group to itself, with 20 repeats for each group; or correlating the sequence counts with a randomly generated set of CBIT scores drawn from the same distribution as the actual CBIT scores (Fig. S1), with 40 repeats. Then for each sample size, the power is estimated by identifying the fraction of largest 20 number of significant sequences that came from the CBAS run on the resampled data (as opposed to the comparisons comparing groups to themselves or randomly generated CBIT).

Power estimation for sequence usage for "complete" sequences

Populations the same size as the datasets were bootstrapped (sampled with replacement). 1000 CBAS were run on different bootstrapped samples. The smallest sequence usage rate for which there was 80% power to detect a difference was defined as the smallest usage rate from the sequences that were significant on 80% of the CBAS runs.

Spontaneous alternation in flies

Data. Data was downloaded from: https://lab.debivort.org/neuronal-control-of-locomotor-handedness/. The first 250 turns for each fly were used. Flies that did not do at least 250 turn were excluded from the analysis. 466/1,225 CA and 565/1372 w1118 flies did not reach the 250-turn criterion.

Two-step task in humans

Data. Data was downloaded from: https://osf.io/usdgt/overview. All subjects in the dataset completed all 200 trials/400 choices. Therefore, all subjects were included in the analysis.

Table 1 | Stereotaxic coordinates of NMDA infusions to produce lesions within the hippocampal formation

AP (mm)	ML (mm)	DV (mm)
−2.8	±1.4	−3.0
−3.3	±2.4	−3.0
−4.1	±1.8	−2.8
−4.1	±3.4	−2.8
−4.8	±2.0	−2.8
−4.8	±4.2	−3.1
−5.5	±2.6	−3.0
−5.5	±3.6	−2.9
−5.5	±5.0	−3.5
−5.5	±5.0	−5.5
−6.2	±4.0	−3.4
−6.2	±5.4	−4.4

Adapted from Kim & Frank[45]. The coordinates are given for a Long Evans rat skull which is leveled so that bregma and lambda lie in the same horizontal plane. AP: anteroposterior, ML: mediolateral, DV: dorsoventral. The AP and ML coordinates are measured from bregma. DV coordinates are measured from the dural surface above the brain.

MB/MF RL model. The model was implemented following the description from Gillan et al.[43]. The model combines units for model-based and model-free control, as well as stickiness. For each trial, t, there are choices at two stages, $c_{1,t}$ and $c_{2,t}$. The choice at stage 1 leads stochastically to a stage 2 state, s_t. Reward, r_t, occurs following the second stage choice based on dynamic probabilities for each image at the second stage. $r_t = 1$ indicates reward and $r_t = -1$ indicates no reward. Stage 1 choices depend upon what is learned at stage 2; therefore, stage 2 learning is described first.

At stage 2, the model learns a value function, $Q_t^{\text{stage2}}(s, c)$, where s is the state. The reward received at each trial leads to updating Q according to: $Q_{t+1}^{\text{stage2}}(s_t, c_{2,t}) = (1-\alpha)Q_t^{\text{stage2}}(s_t, c_{2,t}) + r_t$, where α is the learning rate parameter. The probability to make a specific stage 2 choice is governed by a softmax: $P(c_{2,t} = c) \propto \exp(\beta^{\text{stage2}} Q_t^{\text{stage2}}(s_t, c))$, which is normalized over both options of c, and β^{stage2} is the inverse temperature parameter.

Stage 1 choices are determined by weighting the value predictions of the different components of the model about the stage 2 value of each stage 1 choice. Model-based values are dictated by the learned values of the stage 2 state, maximized over the two actions: $Q_t^{MB}(c_1) = \max_{c_2}(Q_t^{\text{stage2}}(s, c_2))$, where s is the stage 2 state predominantly produced by stage 1 choice c. Model-free values are dictated by two learning rules: $Q_{t+1}^{MF_0}(c_{1,t}) = (1-\alpha)Q_t^{MF_0}(c_{1,t}) + Q_t^{stage2}(s_t, c_{2,t})$ and $Q_{t+1}^{MF_1}(c_{1,t}) = (1-\alpha)Q_t^{MF_1}(c_{1,t}) + r_t$. Stage 1 choice probabilities are also governed by a softmax: $P(c_{1,t} = c) \propto \exp(\beta^{MB} Q_t^{MB}(c) + \beta^{MF_0} Q_t^{MF_0}(c) + \beta^{MF_1} Q_t^{MF_1}(c) + \beta^{\text{stick}} I(c = c_{1,t-1}))$, where the β values are the temperature parameters for each component, and $I(c = c_{1,t-1})$ is 1 for the choice that repeats the one made on the previous trial and 0 for the other choice.

Following each trial, the value estimates in all Q for all unchosen actions and unvisited states are decayed multiplicatively by $(1-\alpha)$.

Simulated annealing was used to fit the model, initializing with random parameters. At least 10 different fits were run for each set of parameters and the fit parameters leading to the largest log-likelihood was chosen. The data contained instances where a subject did not choose either of the images presented. Those choices were not included in the likelihood for the fitting. However, the lack of choice was coded as its own symbol and analyzed within CBAS for correlations with CBIT.

Spatial alternation in rats

Animals. All experiments on rats were conducted in accordance with University of California, San Francisco Institutional Animal Care and Use Committee and US National Institutes of Health guidelines. Rats were fed standard rat chow (LabDiet 5001). To motivate the rats to perform the task, reward was sweetened evaporated milk: 25 g of sugar per can (354 ml) of evaporated milk (Carnation). The rats were food restricted to up to 85% of their basal body weight. Outbred, long Evans rats were used. Rats were bred in-house with founder animals purchased from Charles Rivers and acquired from the Medical College of Wisconsin. Adult rats were aged ~4–6 months for the experiments.

Surgery. Anesthesia was induced and maintained using ketamine, xylazine, atropine, and isoflurane. The animals underwent surgery to either lesion the dorsal and intermediate hippocampus or as control for the lesioning. For the lesion rats, craniotomies were drilled bilaterally to allow for access to all 24 injection sites (Table 1) and the injection needle (33 gauge) was inserted through the dura. For each lesion site, the needle was lowered 0.1 mm below the target coordinates and then raised back to the target prior to injection. 120 or 110 nl of NMDA (20 mg/ml) was injected at each site at a rate of 0.1 μl/min, for males and females, respectively. The needle was left at the site for 2 min after finishing the injection. For the rats that underwent control surgeries, everything was the same as the lesion rats, except no NMDA was injected, and the needle did not dwell at the hypothetical injection site. All animals that underwent the lesion surgery received Diazepam 10 mg/kg intraperitoneally post-operatively.

Hippocampal lesioning was confirmed via micro-CT (Fig. S5) collected at the Advanced Light Source at Lawrence Berkeley National Lab. Brains were processed and imaged, and scans were processed as described in Kastner et al.[68]. CT images were aligned using manual registration in Slicer-3D (http://www.slicer.org)[69,70]. Male and female brains were separately registered.

Spatial alternation behavior. The automated behavior system for spatial alternation behavior was previously described[51]. There are different symbols on each arm of the track serving as proximal cues, and there are distal cues distinguishing the different walls of the room. Pneumatic pistons (Clippard) open and close the doors. Python scripts, run through Trodes (Spike Gadgets), control the logic of the automated system. The reward wells contain an infrared beam adjacent to the reward spigot. The automated system uses the breakage of that infrared beam to progress through the logic of the behavior. In addition to the infrared beam and the spigot to deliver the reward, each reward well has an associated white light LED.

Each cohort of rats is divided into groups of four (or three) animals. The same groups were maintained throughout the duration of the experiment. Within a group, a given rat is always placed in the same rest box, and the four rats of a group serially perform the behavior. The rats have multiple sessions on the track each day. Prior to beginning the spatial alternation task, the rats experience multiple days and sessions where they get rewarded at any arm that they visit (provided it is not an immediate repeat). During this period of the behavior, the duration of a session is defined by a fixed number of rewards, or a fixed amount of time on the track (15 min), whichever came first. During the alternation task the duration of a session was defined either by a fixed number of center arm visits and at least one subsequent visit to any other arm, or a fixed amount of time on the track (15 min), whichever came first.

The algorithm underlying the spatial alternation task is such that three arms on the track have the potential for reward: arms 2, 3, and 4 have the potential to be rewarded, and arms 1, 5, and 6 do not. Of those three potentially rewarded arms we refer to the middle of the three arms as the center arm (arm 3) and the other two arms as the outer arms (arms 2 and 4). Reward is delivered at the center arms if and only if: (1) the immediately preceding arm whose reward well infrared beam was broken was not the center arm. Reward was delivered at the outer two arms if and only if: (1) the immediately preceding arm whose reward well infrared beam was broken was the center arm, and (2) prior to breaking the infrared beam at the center arm, the most recently broken outer arm infrared beam was not the currently broken outer arm infrared beam. The one exception to the outer arm rules was at the beginning of a session, if no outer arm infrared beam was broken prior to the first infrared beam break at the center arm, then only the first condition had to be met.

For the running of the behavior, the infrared beam break determined an arm visit; however, the rats sometimes go down an arm, get very close to the reward wells, but do not break the infrared beam. Therefore, for all the analyses described for the rats, an arm choice is defined as when a rat gets close to a reward well. These times were extracted from a video recording of the behavior. This proximity-based definition of an arm visit added additional arm visits to those defined by the infrared beam breaks, and none of them could ever be rewarded, nor alter the logic of the underlying algorithm. However, because of the non-Markovian nature of the reward contingency, they could affect the rewards provided for subsequent choices. As an example see choices in Fig. 1a. Following being rewarded at arm 3, the animal visits arm 5, but does not break the infrared beam. The subsequent visit to arm 3 is not rewarded because the algorithm views it as an immediate repeat to the same arm, which are never rewarded. The subsequent visit to arm 2 is rewarded because from the perspective of the algorithm, which does not have access to the effectively unvisited arm 5, this is the correct choice.

For the initial dataset, a total of 46 rats that underwent control surgery (22 males, 24 females) and 45 rats that underwent hippocampal lesion surgery (23 males, 22 females) were run on the spatial alternation task. 6 of the rats that underwent hippocampal lesion surgery showed no evidence of lesioning on CT and were therefore not included in the analysis. For the replication dataset, a total of 9 rats that underwent control surgery (all males) and 11 rats that underwent hippocampal lesion surgery (all males) were run on the spatial alternation task. For the replication dataset, all of the rats that underwent hippocampal lesion surgery showed evidence of lesioning on CT. Rats were housed together prior to their undergoing surgery. They were single housed for post-operative recovery and through food restriction and the behavior. Experimenters were blind to the lesion/control status of the animals for the duration of the behavioral experiment and for processing/imaging of the animals for CT.

The first 800 choices during the spatial alternation behavior for each rat were used. Only 1 rats did not reach the criterion. That control, female, rat from the initial dataset, completed 618 choices, and was included in the analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Rat spatial alternation data generated for this work is available on Github at https://github.com/dbkastner/CBAS. Source data are provided with this paper.

Code availability

Code used to calculate CBAS is available on Github at https://github.com/dbkastner/CBAS. Code was written in IGOR Pro (Wavemetrics) version 8.0.

References

1. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).

2. Bialek, W. On the dimensionality of behavior. *Proc. Natl. Acad. Sci. USA* **119**, e2021860119 (2022).
3. Meister, M. Learning, fast and slow. *Curr. Opin. Neurobiol.* **75**, 102555 (2022).
4. Gomez-Marin, A. & Ghazanfar, A. A. The life of behavior. *Neuron* **104**, 25–36 (2019).
5. Trillmich, F., Günther, A., Müller, C., Reinhold, K. & Sachser, N. New perspectives in behavioural development: Adaptive shaping of behaviour over a lifetime? *Front Zool.* **12**, S1 (2015).
6. Dayan, P., Roiser, J. P. & Viding, E. The first steps on long marches: the costs of active observation. in *Rethinking Biopsychosocial Psychiatry*. (eds. Davies, W., Roache, R. & Savulescu, J.) (Rethinking Biopsychosocial Psychiatry., 2018).
7. Hession, L. E., Sabnis, G. S., Churchill, G. A. & Kumar, V. A machine-vision-based frailty index for mice. *Nat. Aging* **2**, 756–766 (2022).
8. Geuther, B. Q. et al. Action detection using a neural network elucidates the genetics of mouse grooming behavior. *eLife* **10**, e63207 (2021).
9. Geuther, B. et al. High-throughput visual assessment of sleep stages in mice using machine learning. *SLEEP* **45**, zsab260 (2021).
10. Sheppard, K. et al. Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation. *Cell Rep.* **38**, 110231 (2022).
11. Weinreb, C. et al. Keypoint-MoSeq: Parsing behavior by linking point tracking to pose dynamics. *Nat. Metods.* **21**, 1329–1339 (2024).
12. Wiltschko, A. B. et al. Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121–1135 (2015).
13. Pereira, T. D. et al. Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
14. Pereira, T. D. et al. SLEAP: A deep learning system for multi-animal pose tracking. *Nat. Methods* **19**, 486–495 (2022).
15. Marshall, J. D. et al. Continuous Whole-Body 3D Kinematic Recordings across the Rodent Behavioral Repertoire. *Neuron* **109**, 420–437.e8 (2020).
16. Dunn, T. W. et al. Geometric deep learning enables 3D kinematic profiling across species and environments. *Nat. Methods* **18**, 564–573 (2021).
17. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
18. Kumar, P., Dayan, P. & Wolfers, T. From complexity to precision—charting decision-making through normative modeling. *JAMA Psychiatry* **81**, 117–118 (2024).
19. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
20. Kerr, N. L. HARKing: Hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2**, 196–217 (1998).
21. Wiltschko, A. B. et al. Revealing the structure of pharmacobehavioral space through Motion Sequencing. *Nat. Neurosci.* **23**, 1433–1443 (2020).
22. Gschwind, T. et al. Hidden behavioral fingerprints in epilepsy. *Neuron* **111**, 1440–1452.e5 (2023).
23. Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P. & Balleine, B. W. Models that learn how humans learn: The case of decision-making and its disorders. *PLoS Comput. Biol.* **15**, e1006903 (2019).
24. Eckstein, M. K., Summerfield, C., Daw, N. D. & Miller, K. J. Predictive and interpretable: Combining artificial neural networks and classic cognitive models to understand human learning and decision making. *bioRxiv* 2023.05.17.541226 (2023).
25. Castro, P. S. et al. Discovering symbolic cognitive models from human and animal behavior. *bioRxiv* 2025.02.05.636732 (2025).
26. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59 (2021).
27. Farrell, M. S. et al. Evaluating historical candidate genes for schizophrenia. *Mol. Psychiatr.* **20**, 555–562 (2015).
28. Border, R. et al. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am. J. Psychiatry* **176**, 376–387 (2019).
29. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacol* **44**, 1518–1523 (2019).
30. Ioannidis, J. P. A. Why most published research findings are false. *Plos Med.* **2**, e124 (2005).
31. Buchanan, S. M., Kain, J. S. & Bivort, B. L. de. Neuronal control of locomotor handedness in Drosophila. *Proc. Natl. Acad. Sci.* **112**, 6700–6705 (2015).
32. Skutt-Kakaria, K., Reimers, P., Currier, T. A., Werkhoven, Z. & Bivort, B. L. de. A neural circuit basis for context-modulation of individual locomotor behavior. *bioRxiv* 797126 (2019).
33. Ayroles, J. F. et al. Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proc. Natl. Acad. Sci.* **112**, 6706–6711 (2015).
34. Werkhoven, Z. et al. The structure of behavioral variation within a genotype. *eLife* **10**, e64988 (2021).
35. Akhund-Zade, J., Ho, S., O'Leary, C. & Bivort, B. de. The effect of environmental enrichment on behavioral variability depends on genotype, behavior, and type of enrichment. *J. Exp. Biol.* **222**, jeb202234 (2019).
36. de Bivort, B. et al. Precise Quantification of Behavioral Individuality From 80 Million Decisions Across 183,000 Flies. *Front. Behav. Neurosci.* **16**, 836626 (2022).
37. Cui, T. et al. Gene–gene interaction detection with deep learning. *Commun. Biol.* **5**, 1238 (2022).
38. Fang, G. et al. Discovering genetic interactions bridging pathways in genome-wide association studies. *Nat. Commun.* **10**, 4274 (2019).
39. Romano, J. P. & Wolf, M. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *J. Am. Stat. Assoc.* **100**, 94–108 (2005).
40. Romano, J. P. & Wolf, M. Control of generalized error rates in multiple testing. *Ann. Stat.* **35**, 1378–1408 (2007).
41. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron* **69**, 1204–1215 (2011).
42. Akam, T. et al. The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron* **109**, 149–163.e7 (2020).
43. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).
44. da Silva, C. F. & Hare, T. A. Humans primarily use model-based inference in the two-stage task. *Nat. Hum. Behav.* **4**, 1053–1066 (2020).
45. Kim, S. M. & Frank, L. M. Hippocampal lesions impair rapid learning of a continuous spatial alternation task. *Plos One* **4**, e5494 (2009).
46. Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. *Neuron* **77**, 1163–1173 (2013).
47. Kay, K. et al. Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552–567.e25 (2020).
48. Fernández-Ruiz, A. et al. Long-duration hippocampal sharp wave ripples improve memory. *Science* **364**, 1082–1086 (2019).
49. Sigurdsson, T., Stark, K. L., Karayiorgou, M., Gogos, J. A. & Gordon, J. A. Impaired hippocampal–prefrontal synchrony in a genetic mouse model of schizophrenia. *Nature* **464**, 763–767 (2010).
50. Kastner, D. B., Gillespie, A. K., Dayan, P. & Frank, L. M. Memory alone does not account for the speed of learning of a simple spatial alternation task in rats. *J. Neurosci.* **40**, JN-RM–0972-20 (2020).
51. Kastner, D. B. et al. Spatial preferences account for inter-animal variability during the continual learning of a dynamic cognitive task. *Cell Rep.* **39**, 110708–110708 (2022).

52. Bennett, D. et al. Mouse Behavior on the Trial-Unique Nonmatching-to-Location (TUNL) Touchscreen Task Reflects a Mixture of Distinct Working Memory Codes and Response Biases. *J. Neurosci.* **43**, 5693–5709 (2023).
53. Scheunemann, J., Miegel, F., Moritz, S., Yassari, A. H. & Jelinek, L. Superstitious behaviors and beliefs in obsessive–compulsive disorder: An experimental investigation using the Illusory Contingency Task. *J. Obsessive-Compuls. Relat. Disord.* **49**, 101013 (2026).
54. Vellozo, A. P. et al. Symmetry dimension in obsessive–compulsive disorder: Prevalence, severity and clinical correlates. *J. Clin. Med.* **10**, 274 (2021).
55. Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454–1458 (2012).
56. Aggleton, J. P., Sanderson, D. J. & Pearce, J. M. Structural learning and the hippocampus. *Hippocampus* **17**, 723–734 (2007).
57. Whittington, J. C. R. et al. The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263.e23 (2020).
58. Benna, M. K. & Fusi, S. Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proc. Natl. Acad. Sci.* **118**, e2018422118 (2021).
59. Polti et al. Rapid encoding of task regularities in the human hippocampus guides sensorimotor timing. *eLife* **11**, e79027 (2022).
60. Zutshi, I. et al. Hippocampal neuronal activity is aligned with action plans. *Nature* **639**, 153–161 (2025).
61. Onih, A., Shen, X., Pentousi, L., Boboeva, V. & Akrami, A. The hippocampus enables abstract structure learning without reward. *bioRxiv* 2026.02.14.705916 (2026).
62. Markowitz, J. E. et al. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature* **614**, 108–117 (2023).
63. Walf, A. A. & Frye, C. A. The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. *Nat. Protoc.* **2**, 322–328 (2007).
64. Miller, E. A. et al. Robust and replicable measurement for prepulse inhibition of the acoustic startle response. *Mol. Psychiatr.* **26**, 1909–1927 (2021).
65. Wang, A., Williams, L., Ogunlade, O., Morales, J. & Kuchibhotla, K. Starting from zero: Task learning in completely naïve individuals. *PsyArXiv* 10.31234/osf.io/u5btm (2023).
66. Clarke, D., Romano, J. P. & Wolf, M. The Romano-Wolf multiple hypothesis correction in stata. *Stata J.* **20**, 12–843 (2020).
67. DiCiccio, C. J. & Romano, J. P. Robust permutation tests for correlation and regression coefficients. *J. Am. Stat. Assoc.* **112**, 1211–1220 (2017).
68. Kastner, D. B. et al. Scalable method for micro-CT analysis enables large scale quantitative characterization of brain lesions and implants. *Sci. Rep.* **10**, 20851 (2020).
69. Fedorov, A. et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson Imaging* **30**, 1323–1341 (2012).
70. Kikinis, R., Pieper, S. D. & Vosburgh, K. G. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. in *Intraoperative Imaging and Image-Guided Therapy* (ed. Jolesz, F.) Ch. 19, 277–289 (Springer NY, 2014).

Acknowledgements

We thank Claire Gillan and Benjamin de Bivort for making their data publicly available. We thank Thomas Akam, Nathanial Daw, Hannah Berg, and Adam Frank for very helpful discussions and comments on the manuscript. This research used resources of the Advanced Light Source, a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231.

Author contributions

D.B.K. and P.D. designed the study, D.B.K., J.P.R., and P.D. developed the analysis method, D.B.K., N.O.Y., C.Y.L., C.H., J.T., A.C., M.M., V.K. and G.W. collected the data, D.L.P. provided expertise and resources. D.B.K., N.O.Y., C.Y.L. and M.M. analyzed the data, D.B.K and P.D. wrote the manuscript, and D.B.K., J.P.R. and P.D. edited the manuscript.

Funding

This work was supported by grants from the Jane Coffin Childs Memorial Fund for Medical Research (D.B.K.), the UCSF Physician Scientist Scholars Program (D.B.K.), an NIH R25 (R25MH060482) (D.B.K), a Simons Foundation Autism Research Initiative grant (899599) (D.B.K.), an NIMH K08 Award (K08MH139983) (D.B.K.), the Max Planck Society (P.D.), and the Humboldt Foundation (P.D.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-76681-3.

Correspondence and requests for materials should be addressed to David B. Kastner.

Peer review information *Nature Communications* thanks Steve Chang, Rachel Hill, Yutaka Sakai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026