frontiers | Frontiers in Behavioral Neuroscience

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fnbeh.2022.962494

OPEN ACCESS

EDITED BY
Dorota Frydecka,
Wroclaw Medical University, Poland

REVIEWED BY
Zafeirios Fountas,
Huawei, United Kingdom
Karl Friston,
University College London,
United Kingdom

*CORRESPONDENCE
Dimitrije Marković
dimitrije.markovic@tu-dresden.de

SPECIALTY SECTION
This article was submitted to
Learning and Memory,
a section of the journal
Frontiers in Behavioral Neuroscience

RECEIVED 06 June 2022
ACCEPTED 26 September 2022
PUBLISHED 17 October 2022

CITATION
Marković D, Reiter AMF and Kiebel SJ
(2022) Revealing human sensitivity to a
latent temporal structure of changes.
Front. Behav. Neurosci. 16:962494.
doi: 10.3389/fnbeh.2022.962494

Revealing human sensitivity to a latent temporal structure of changes

Dimitrije Marković[1]*, Andrea M. F. Reiter[1,2,3] and Stefan J. Kiebel[1,4]

[1]Department of Psychology, Technische Universität Dresden, Dresden, Germany, [2]Department of Child and Adolescence Psychiatry, Psychosomatics and Psychotherapy, Centre of Mental Health, University Hospital Würzburg, Würzburg, Germany, [3]German Center of Prevention Research on Mental Health, Julius-Maximilians Universität Würzburg, Würzburg, Germany, [4]Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Dresden, Germany

Precisely timed behavior and accurate time perception plays a critical role in our everyday lives, as our wellbeing and even survival can depend on well-timed decisions. Although the temporal structure of the world around us is essential for human decision making, we know surprisingly little about how representation of temporal structure of our everyday environment impacts decision making. How does the representation of temporal structure affect our ability to generate well-timed decisions? Here we address this question by using a well-established dynamic probabilistic learning task. Using computational modeling, we found that human subjects' beliefs about temporal structure are reflected in their choices to either exploit their current knowledge or to explore novel options. The model-based analysis illustrates a large within-group and within-subject heterogeneity. To explain these results, we propose a normative model for how temporal structure is used in decision making, based on the semi-Markov formalism in the active inference framework. We discuss potential key applications of the presented approach to the fields of cognitive phenotyping and computational psychiatry.

KEYWORDS

decision making, temporal structure, Bayesian inference, active inference, reversal learning

1. Introduction

The passage of time is a fundamental aspect of human experience. Our behavior is tightly coupled to our estimate of the elapsed time and the expectations about the time remaining to fulfill short or long-term goals. We are highly sensitive to the temporal structure of our everyday environment and capable of forming precise beliefs about the duration of various events (e.g., a theater play, traffic lights, waiting in a queue). In practice, temporal structure is typically latent (e.g., not reflected in external clocks) and we seem to rely on an internalized timing mechanism, such as various implicit clocking mechanisms (Buhusi and Meck, 2005). This enables us to provide temporal context and an order to events, and to form beliefs about the underlying temporal structure (Eichenbaum, 2014). It has been proposed that these temporal beliefs are used to make predictions and to adapt our behavior successfully to ever-changing conditions (Griffiths and Tenenbaum, 2011). Therefore, understanding how we learn and represent the temporal structure of our every day environment (Kiebel et al., 2008) and use these representations for making decisions (Marković et al., 2019) is essential for understanding human adaptive behavior (Purcell and Kiani, 2016).

Neuronal and behavioral mechanisms of time perception have been studied in humans and animals, traditionally using interval timing tasks (Meck, 1996; Eagleman, 2008). The key insights of these experiments are that humans and animals integrate the experience of between event duration, in a given context, to form beliefs about possible future duration they might experience. They use these beliefs when estimating or reproducing a newly experienced interval (Jazayeri and Shadlen, 2010); in line with a Bayesian account of decision-making (Shi et al., 2013). However, it is still an open question how we integrate time perception and beliefs about durations into everyday decision making. Recently, distinct but interlinked research fields have illustrated the importance of temporal representations for cognition and decision making in sequential and dynamic tasks (McGuire and Kable, 2012; Eichenbaum, 2014; Vilà-Balló et al., 2017; Nobre and Van Ede, 2018). The sequential neuronal activity in the hippocampus has been suggested to represent elapsed time (Friston and Buzsáki, 2016; Buzsáki and Llinás, 2017; Eichenbaum, 2017), which have led to the postulate of time cells in the hippocampus (Itskov et al., 2011; Eichenbaum, 2014; MacDonald et al., 2014) critical for memory and decision-making. For example, in the research on temporal aspects of attention it has been demonstrated that temporal expectations guide allocation of attentional resources in time (Nobre and Van Ede, 2018). Similarly, inter-temporal choices or one's willingness to wait for higher reward is strongly influenced by temporal expectations (McGuire and Kable, 2012).

Motivated by the rich literature on temporal representations in the brain, here we focus on the question of how humans form complex temporal representation of their environment. We test how such temporal representations support decisions about whether to explore or to exploit in anticipation of a change in the environment. We introduce a novel computational model of behavior that describes learning of a latent temporal structure of a dynamic task environment in the context of sequential decision making. The computational model is applicable to any task that can be cast as a dynamic multi-armed bandit problem (Gupta et al., 2011) with semi-Markovian changes or switches in the underlying latent states (Janssen and Limnios, 1999). Here we specifically apply the model to describe learning in a sequential (probabilistic) reversal learning task (Costa et al., 2015; Reiter et al., 2016, 2017; Vilà-Balló et al., 2017). We do so by manipulating temporal contexts in this task: Subjects encountered semi-regular intervals between contingency reversals in one environment. Their behavior was contrasted with behavior in another environment where intervals between contingency reversals were irregular.

The proposed behavioral model was based on three components: (i) a set of templates representing possible latent temporal structure of reversals using an implicit representation of between reversal duration (Yu, 2015), (ii) the update of beliefs about states and temporal templates derived *via* approximate inference (Yu and Kobayashi, 2003; Parr et al., 2019), and (iii) the action selection, that is the planning process, cast as active inference (Friston et al., 2017; Markovic et al., 2021). Together these components allow us to define an efficient and approximate active learning and choice algorithm of latent temporal structures based on variational inference (Blei et al., 2017). Here we extend on our previous investigation of human behavior in temporally structured dynamic environments (Marković et al., 2019). In this work, we demonstrated that a computational model which infers a between-event duration, can be used to reveal subjects' beliefs about the latent temporal structure in a dynamic learning task. However, a question that has remained open is how humans acquire temporal structure in the first place. Understanding the learning of temporal structure is critical for revealing between-individual variability in temporal expectations and capturing the evolution of temporal representations within individuals. Critically, with the extended model we present here, we are indeed able to capture the learning of temporal representation and address the non-stationarity of subjects' temporal representation during the course of the experiment.

Our aim is to address the following questions: (i) Are subjects a priori biased toward expecting regular or irregular temporal structure? (ii) Are subjects able to learn latent temporal structure without explicit instructions? (iii) How does the quality of temporal representation impact their performance? Using simulations we can illustrate the interaction of accurate representation of temporal structure and behavior, mainly performance on the task and the engagement with exploratory behavior. Using model-based analysis, that is, by estimating the prior beliefs—under a semi-Markovian generative model—that best explain observed choice behavior, we demonstrate high diversity between subjects both in their prior beliefs about temporal structure, and their ability to adapt their beliefs to different latent temporal structure. Crucially, we link the quality of temporal representation to subjects' performance both in terms of group-level performance and within-subject variability of their performance during the task.

In what follows we will first briefly describe the experimental task, provide the overall summary of behavioral characteristics, introduce the behavioral model, and finally show results of the model-based analysis of behavior. The formal details of the approach are described in Section 4.

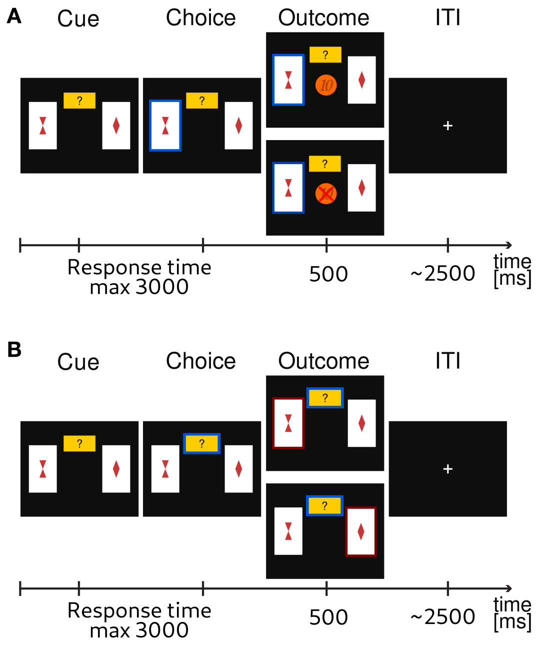


FIGURE 1
Exemplary trial sequence of the probabilistic reversal learning task. Subjects were instructed that one card always had a higher probability of a monetary reward. They were instructed to choose the card that they thought would lead to a monetary gain with higher probability, or, alternatively, choose to explore (small yellow rectangle with question mark). The latter would provide them with a correct information about what option would have had a higher probability of reward. **(A)** If participants had chosen one of the cards, the corresponding card was highlighted and feedback was displayed. The feedback consisted of either the visual display of a 10 Euro cents coin in the center of the screen for a gain outcome, or a crossed 10 Euro cents for a loss outcome. **(B)** If the participant had chosen the explore option, the card with the currently highest reward probability was highlighted (either the left or the right card).

2. Results

A typical probabilistic reversal learning task asks subjects to make a binary choice between two options, e.g., A and B, where each option is associated with a probability of receiving a reward or punishment. For example, initially choosing A returns a reward with a high probability $p_H = 0.8$ and choosing B returns a reward with low probability $p_L = 0.2$. Importantly, after several trials the reward contingencies reverse, i.e., switch, such that choosing B returns the reward with high probability p_H. However, subjects are not informed about the reversal and they have to infer that a change occurred from the feedback they receive in order to adapt their behavior. From the point of view of participants, a reversal can be difficult to detect as outcomes are probabilistic. This means that if someone observes a loss after a sequence of gains, e.g., when choosing the option A, this could be caused either by: (i) a true reversal, where now option B is rewarded with the probability p_H or (ii) by an unlucky outcome of an otherwise correct choice. To obtain a more direct information about the subjective uncertainty of participants about the correct choice (i.e., choosing the option with high reward probability, p_H) on any given trial, we extended the standard design with an additional third exploratory option. This new option does not result in monetary gain or loss but provides information about the correct choice on a current trial. A high uncertainty about the best choice (current context) can be easily resolved by selecting the epistemic option. We will label all choices of the exploratory options as exploratory, and all other choices as exploitative (note that the outcomes of exploitative options also provide some information about the current context). A trial sequence of the experimental task is shown in Figure 1.

To investigate subjects' ability to learn latent temporal structure we defined two experimental conditions (manipulated in a between-subject design), one with irregular reversals and another with regular reversals (see Figure 2). In the condition with irregular reversals, the moments of reversals are not predictable and between-reversal intervals are drawn from a geometric distribution (Figure 2A). In the condition with regular reversals, the moments of reversal are predictable, and they occur at semi-regular intervals, drawn from a negative binomial distribution (Figure 2B). Subject were randomly assigned to one of the two possible conditions, as illustrated in Figure 2. In the first condition, subjects experience irregular reversal statistics for 800 trials, after which the reversals occur at semi regular intervals for the last 200. In the second condition, subjects experience semi-regular reversal statistics for 800 trials, and then the irregular reversal statistics during the last 200 trials. Note that when changing the temporal statistics we copied the time series of reversals from the initial 200 trials of the different condition. The motivation for using parts of the trajectories from one condition in another condition comes from the process we use to define the moments of reversals in both conditions. We aimed to tailor both experimental conditions in a way that maximizes the behavioral differences between subjects entertaining different underlying beliefs about latent statistics of reversals. Such optimization results in improved model selection and parameter estimates as distinct latent beliefs result in more pronounced behavioral differences.

Therefore, we have generated a large number (10^5) of trajectories of length $T = 800$ for each condition and kept the one for which we found the maximal performance difference between agents with a correct representation of latent reversal statistics and an agent with a representation from the opposite condition. As we kept only single trajectory of reversals for each condition, we have fixed the moments of reversal for each subject group (depending on the condition, reversals occur always on the same trials). Furthermore, the same choice by different subjects exposed to the same condition leads to the same outcome on any given trial (the outcome statistics were generated only once for each condition and trial, and then replayed to all subjects depending on their choices and the condition they were assigned to). Hence, we removed the noise in behavioral responses which would be induced by unique experiences of each subject in the experiment, were we to generate moments of reversal and response-outcomes on-the-fly for each subject.

2.1. Analysis of choice data

We will first describe the behavioral characteristics of the two groups of subjects exposed to the two different experimental conditions. The two behavioral measures of interests here are the *performance* (odds of being correct, i.e., odds of choosing the option with the higher reward probability) and *probing* (odds of exploring, i.e., odds of choosing the exploratory option). We describe all the behavioral measures in detail in Section 4.5.

Subjects ($N = 74$) were pseudo-randomly assigned to one of the two experimental conditions, where $n_r = 41$ participants were assigned to the condition with regular reversals, and $n_i = 33$ to the condition with irregular reversals. Note that some subjects rarely engaged with exploratory option. Out of 50 subjects who where exposed to the variant of the experiment with exploratory option (24 subjects performed a standard version of the task without exploratory option, see Section 4.3 for more details), 5 subjects never engaged with the exploratory option. In Figure 3, we provide a summary of average behavioral measures for individual subjects. We do not find any significant performance differences between the two regularity conditions (see Figure 3A). However, for the subset of subjects which interacted with the exploratory option (45 subjects) we find that the performance is positively correlated with probing (Pearson correlation coefficient for all data points $r = 0.6$, with $p < 10^{-4}$; for the regular condition $r = 0.73, p < 0.0001$, and for the irregular condition $r = 0.52, p < 0.02$; see Figure 3B). Interestingly, neither of the two behavioral measures (when plotted as a within subject average over the course of experiment), reveals obvious between-condition differences. However, when comparing the temporal profile of these measures over the course of experiment (see Supplementary Figure 1), one notices large variability both between subjects but also within a subject over the course of experiment; suggesting ongoing learning of the task structure. In what follows we will classify the heterogeneity of behavioral responses using a model-based analysis.

2.2. Behavioral model

The behavioral model will allow us to investigate the process of learning of the latent temporal structure in different

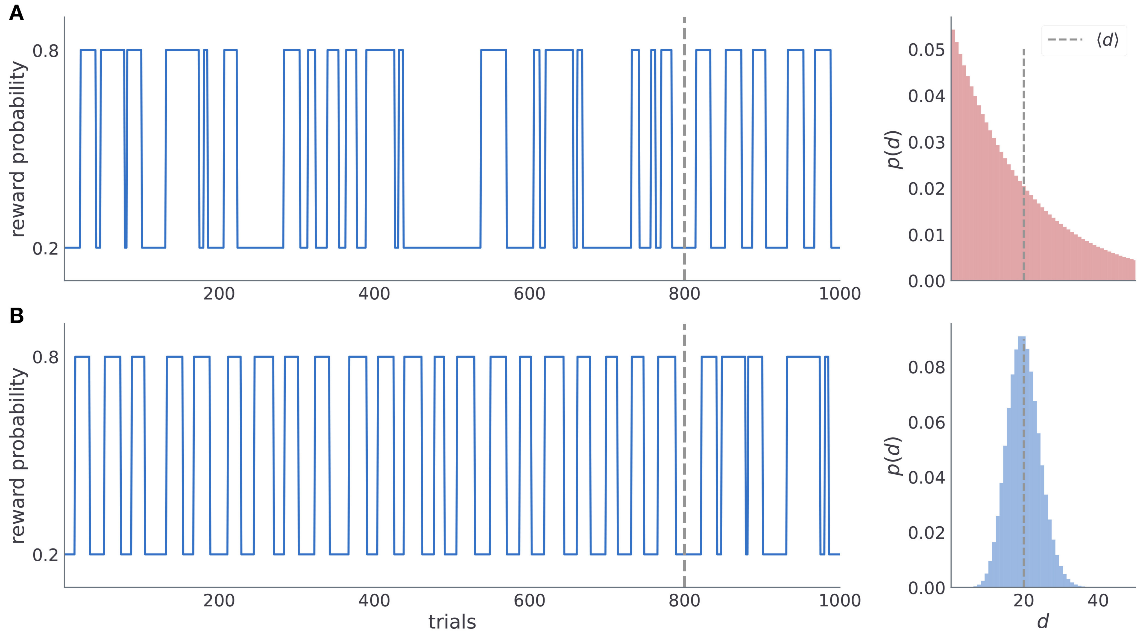


FIGURE 2
Time series of reward probabilities. **(A)** Condition with irregular reversals, and **(B)** condition with (semi-)regular reversals. The reward probability of the high-probability stimuli at any time step was set to $p_H = 0.8$ and the low-probability stimuli to $p_L = 0.2$. Dashed vertical lines shows the moment of change of the latent temporal structure: (i) from irregular to semi-regular statistics in the irregular condition, and (ii) from the semi-regular to irregular statistics in the regular condition. Figures on the right illustrate the generative distribution of the between-reversal intervals d for each condition. Note that the mean between reversal duration $\langle d \rangle$ is identical in both conditions.

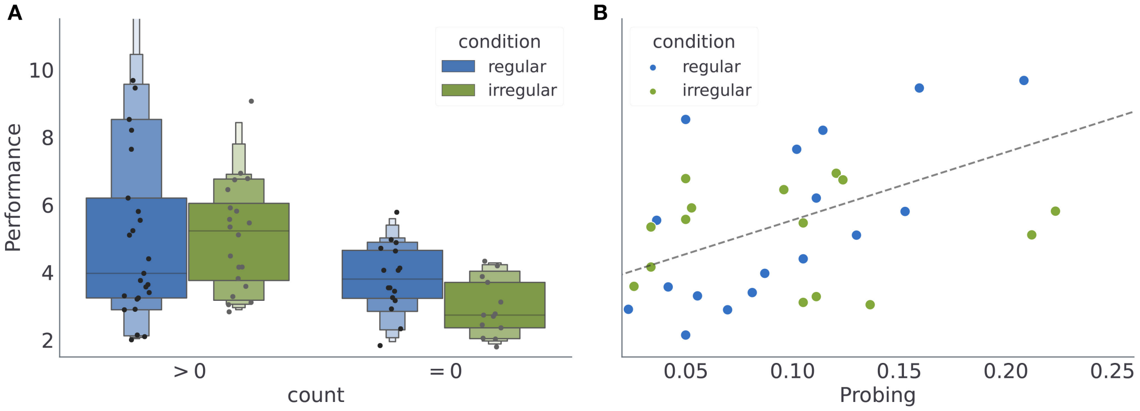


FIGURE 3
Averages of behavioral summary measures. **(A)** Distribution of the mean performance of subjects with low and high number of exploratory choices (see Section 4). **(B)** Dependence of the mean performance on the mean probing, where we excluded participants without exploratory choices (count = 0). Note that when computing mean performance and mean probing for each participants, we have excluded the first 400 (initial responses during which the subjects might have still been adjusting to the task) and last 200 (responses after the change in the reversal statistics) responses of each participant, see Section 4.7 for the motivation for the cutoff.

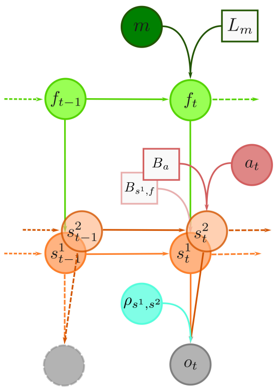
experimental conditions, reveal subjects' preferences to engage with an exploratory option (collect information), and subjects' motivation to collect rewards. We achieve this by fitting free model parameters to behavioral responses of each subject (see Section 4 for more details). Our aim with the model based analysis is to quantify beliefs about temporal structure of reversals and understand how the belief updating influences subjects' behavior.

We conceptualized the behavioral model as an active inference agent (Friston et al., 2015, 2016) with hidden semi-Markov models (Yu, 2010), which are capable of representing and inferring latent temporal structure. In active inference, besides defining perception and learning as a Bayesian inference process, action selection is also cast as an inference problem aimed at minimizing the expected surprise about future outcomes, that is, the expected free energy (Smith et al., 2022; see also Equation 18). Through its dependence on the expected free energy, the action selection has an implicit dual imperative (see possible factorization of the expected free energy in Equation 18): The expected free energy combines intrinsic and extrinsic value of a choice, where intrinsic value corresponds to the expected information gain, and the extrinsic value to the expected reward of different choices. The implicit information gain or uncertainty reduction pertains to beliefs about the task's dynamical structure and choice-outcome mappings (e.g., Schwartenbeck et al., 2013; Kaplan and Friston, 2018). Therefore, selecting actions that minimize the expected free energy dissolves the exploration-exploitation trade-off, as every action is driven both by expected value and the expected information gain. This is a critical feature of active inference models which allows us to account for exploratory choices (see Figure 1).

We express the agent's generative model of task dynamics in terms of hidden semi-Markov models (HSMM) (Yu, 2015; Marković et al., 2019). The HSMM framework extends a standard hidden Markov model with an implicit (or explicit) representation of durations between consecutive state changes. HSMM have found numerous applications in the analysis of non-stationary time series in machine learning (Duong et al., 2005; Gales and Young, 2008), and in neuroimaging (Borst and Anderson, 2015; Shappell et al., 2019). HSMM have also been used in decision making for temporal structuring of behavioral policies (Bradtke and Duff, 1994) or in temporal difference learning as a model of dopamine activity when the timing between action and reward is varied between experimental trials (Daw et al., 2002).

Here, we use the semi-Markov representation of task dynamics within the behavioral models to define an agent that can learn latent temporal structure, form beliefs about moments of change, and anticipate state changes. We implemented the learning of the hidden temporal structure of reversals as a variational inference scheme, where we assume that the agent entertains a hierarchical representation of the reversal learning task, with a finite set of models of possible temporal structure of the dynamic environment. In other words, we assume that human brain entertains a set (possibly a very large set) of temporal templates. In Figure 4, we show the graphical representation of the generative model of behavior, which is described detail in Section 4.6. Here we will briefly introduce the relevant parametrization of the behavioral model, which are critical for understanding the model comparison results presented in the next subsection.

Each temporal template m corresponds to a pair of parameters $m = (\mu, \nu)$ that define the frequency of reversals



$$p(m) = Cat\left(\boldsymbol{M}\right), \quad m = (\mu, \nu), \quad \mu \in \{5, \ldots, 45\}, \nu \in \{1, \ldots, 10\}$$

$$p\left(f_t | f_{t-1}, m\right) = Cat\left(\boldsymbol{L}_{m,f}\right), \quad f_t \in \{1, \ldots, \nu + 1\}$$

$$p(a_t) = s_{max}\left(\gamma G_{a_t}\left[\boldsymbol{P}_o\right]\right), \quad P(o_t) = Cat\left(\boldsymbol{P}_o\right)$$

$$p\left(s^2_t | s^2_{t-1}, a_t\right) = Cat\left(\boldsymbol{B}_a\right)$$

$$p\left(s^1_t | s^1_{t-1}, f_t\right) = Cat\left(\boldsymbol{B}_{s^1,f}\right)$$

$$p\left(\boldsymbol{\rho}_{s^1,s^2}\right) = Dir\left(\boldsymbol{\alpha}_{s^1,s^2}\right)$$

$$p\left(o_t | s^1_t, s^2_t\right) = Cat\left(\boldsymbol{\rho}_{s^1,s^2}\right)$$

FIGURE 4
Graphical representation of the generative model and model summary. At the top of the hierarchy is the temporal template variable *m*. The total number of temporal templates is finite, e.g., $m \in \{1, \ldots, m_{max}\}$, and each template *m* provides an implicit representation of a prior probability distribution over between-reversal intervals *d*, parameterized with a pair $m = (\mu, \nu)$, where μ expresses mean between reversal interval, and ν plays a role of a precision parameter, that defines regularity of between-reversal intervals. The implicit representation of temporal structure is encoded with probability transition matrices $L_{m,f}$ of latent phases *f*. The number of latent phases depends on precision ν. The reversal can occur only when the end phase is reached ($f = \nu + 1$). Therefore, the phase variable *f* controls the transitions probability $\boldsymbol{B}_{s^1,f}$ between latent states of the task denoted as random variable $s^1 \in \{1, 2\}$. At every trial *t* the subject makes a choice a_t hence decides on which option (s^2) to select which results in an outcome o_t. The choices are deterministic, meaning that the corresponding transition probability corresponds to identity matrix, that is, $p\left(s^1_t | s^1_{t-1}, a_t\right) = p\left(s^1_t | a_t\right) = \delta_{s^1_t, a_t}$, hence $\boldsymbol{B}_a = I_3$. Finally, the choice-outcome contingencies are treated as latent variables $\boldsymbol{\rho}_{s^1,s^2}$ which have to be learned over the course of the experiment. We use a vague Dirichlet prior over choice-outcome contingencies. Inverting the generative model of outcomes using variational inference defines the inference and learning component of the behavioral model. In turn, marginal beliefs about latent states s^1_t, s^2_t and parameters $\boldsymbol{\rho}_{s^1,s^2}$ are used to define action selection, that is compute the choice likelihoods using the expected free energy (Equation 18).

μ and the regularity of reversals ν (the higher the value the more regular the changes are). In Figure 12, we illustrate three of these templates, which differ in their regularity parameter ν, but all have the same frequency parameter μ. It is important to note that when $\nu = 1$ (the lowest value) the temporal templates correspond to the hidden Markov model (HMM) representation. HMM representation implies that the moments of reversals are unpredictable, or maximally irregular. Here we use the HMM representation as a reference point for determining whether participants were able to learn latent temporal structure of reversals, and whether they a priori expected predictable moments of reversal.

When simulating behavior and fitting the model to participants' choices, we use a prior probability $p\left(m\right)$ over temporal templates m to restrict otherwise rich set of all possible temporal templates $m = \left(\mu, \nu\right)$, that span all combinations of $\mu \in \{5, \ldots, 45\}$ and $\nu \in \{1, \ldots, 10\}$. Hence, template prior $p\left(m\right)$ reflects prior expectations of an agent at the beginning of the experiment about the possible temporal structure of the task dynamics. Therefore, to capture a wide range of prior beliefs we require a flexible prior $p(m)$ that can reflect subjects with different prior expectations about temporal structure. Posterior estimates of the most likely parameterizations of the temporal prior, allows us to infer from the behavioral data if participants' beliefs are a priori precise and biased toward expecting irregular reversals, or are imprecise and accommodate a wide range of possible latent temporal structures. In the model, we use the following prior over temporal templates:

$$\begin{aligned} p\left(m | \nu_{max}\right) &= p\left(\mu, \nu | \nu_{max}\right) \\ &= p(\mu) p\left(\nu | \nu_{max}\right) \\ p(\mu) &= \frac{1}{40} \\ p\left(\nu | \nu_{max}\right) &= \begin{cases} \frac{1}{\nu_{max}} & \text{for } 1 \leq \nu \leq \nu_{max} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

where $\nu_{max} \in \{1, \ldots, 10\}$. Note that the prior regularity parameter ν_{max} reflects Bayesian prior expectations about the maximal precision of between-reversal intervals. In other words, ν_{max} captures the agent's expectations about the maximal regularity of reversals, and hence their predictability. Thus, with this parameterization we assume that subjects, at the beginning of the experiment, have uniform beliefs about a possible mean duration between reversal interval, but might differ in their propensity to represent high or low regularity of between-reversal intervals. For example, some subjects could hold precise beliefs that reversals were not under their control and were therefore inherently unpredictable (corresponding to $\nu_{max} = 1$). Such a subject would fail to learn—or accumulate evidence for—the regularity of reversals in the regular condition. Conversely, some participants may have imprecise prior beliefs about regularity ($\nu_{max} > 1$); enabling them to learn that reversals were regular, thus predictable, in the appropriate condition.

The beliefs about temporal templates, influence the beliefs about the reversal probability on any given trial (i.e., how likely is that a reversal occurs in the next trial), and consequently modulate beliefs about the latent state of the task (i.e., which card is associated with high reward probability) and corresponding outcome probabilities. In turn, the beliefs about the latent state influence the choices. As mentioned above, choices are defined as the minimizers of the expected free energy (surprise about future outcomes), typically denoted by G. Given the expected free energy $G_a\left[\boldsymbol{P}_o, \nu_{max}, t\right]$ of action a on trial t we define the choice likelihood as

$$a_t \sim p(a) \propto e^{\gamma G_a\left[\boldsymbol{P}_o, \nu_{max}, t\right]}. \tag{2}$$

Here, the parameter γ denotes choice precision, the vector of probabilities $\boldsymbol{P_o} = \left(p_-, p_+, \frac{1}{2}p_c, \frac{1}{2}p_c\right)$ denotes prior preferences over possible outcomes, that is, losses (−), gains (+), and cues (c). In active inference (Friston et al., 2017) prior preference parameter $\boldsymbol{P_o}$ defines a preference of the agent to observe rewards and collect information (engage with the exploratory option). This balance is at the core of active inference and rests upon choosing actions that minimize expected free energy (see Section 4). In turn, expected free energy can be decomposed into epistemic value (i.e., expected information gain) and extrinsic value (i.e., expected preferences or reward). The relative contribution of epistemic and extrinsic value depends upon the precision of preferences over outcomes. In other words, if subjects do not care which of the four outcomes they encounter, then they will behave in a purely exploratory fashion. Conversely, if they have precise or strong preferences, extrinsic value will dominate. In our setup, the precision of preferences rests on two differences; namely the difference between reward and loss, and the difference between collecting rewards or information. Interestingly, a prior preference for collecting information has, itself, epistemic affordance (or at least has greater epistemic value than collecting rewards). This kind of prior preference emerges during the formation of epistemic habits. In the terminology of reinforcement learning, the logarithm of prior preferences $\ln \boldsymbol{P_o}$ assigns a subjective value to possible outcomes, and the expectation of log-preferences defines the expected value of different actions (see Equation 18).

Importantly, we use Equation (2) in two different ways: (i) as a mapping from beliefs into actions which we used to simulate behavioral choices, and (ii) as a choice likelihood which we use for inverting the model when fitting the model to subjects' choices to derive the posterior estimates of free model parameters $(\gamma, p_-, p_+, \nu_{max})$, individually for each subject. Details of the model inversion procedure are described in Section 4.7.

2.2.1. Simulating the behavioral effect of prior expectations over temporal templates

By simulating the model's behavior given different values of temporal regularity parameter ν_{max}, we aimed to demonstrate that the agent can acquire a correct representation of the latent temporal structure in different experimental conditions, and that ν_{max} influences the dynamics of both performance and probing. Importantly, different values of ν_{max} should lead to sufficiently distinct behavior, if we hope to accurately associate subjects' behavior with underlying model parameterization.

The temporal regularity parameter ν_{max} is the key parameter in the model to understand how learning about temporal structure comes about. As ν_{max} constrains the maximal temporal regularity the agent expects in the task, it is a measure of subjects' sensitivity to the latent temporal structure. Importantly, we find that varying ν_{max} results in simulated behavior with distinct behavioral patterns in our two experimental conditions as shown in Figure 5. As we increase ν_{max} the behavioral performance increases, in both conditions. In contrast, as we increase ν_{max} the probing decreases, as the agent is more certain about the moment of reversal and requires information provided by exploratory option less often. Note that different values of ν_{max} induce stronger differences in both performance and probing in the regular condition, compared to the irregular condition. Practically, this means that we can infer ν_{max} from behavioral data with higher precision in regular than in irregular condition. We validate the classification accuracy of ν_{max} based on posterior estimates given simulated data in the form of confusion matrix as shown in Supplementary Figure 2. Note that even in the ideal case when behavior is generated exactly from the behavioral model, classification accuracy with regard to ν_{max} is substantially lower in irregular compared to irregular condition. We will clarify the impact of low classification accuracy in the next subsection when discussing the results of model-based analysis.

2.2.2. Demonstrating the learnability of latent temporal structure

As a next step we will illustrate that the agent with the highest value of temporal prior ($\nu_{max} = 10$)—that is, the agent with the most adaptable beliefs about the latent temporal structure—is capable of accurately inferring the correct temporal template m, and that the rate at which agent learns correct representations of the temporal structure depends on the given temporal context. Hence, we expect that human subjects, with similar prior expectations about temporal structure, should also be capable of learning the correct statistics. In Figure 6, we show posterior beliefs over temporal templates in the form of marginal posterior beliefs about the mean μ and the regularity ν at each time step of the experiment. We see that the agent quickly learns the correct mean between-reversal duration (already after 200 trials the highest posterior probability is close to $\mu = 19$), but it takes longer (more than 400 trials) to form precise beliefs about the level of temporal regularity. In contrast, in the irregular condition, learning the correct mean between-reversal-interval (fixed to $\mu = 19$ in both conditions) takes more time and is less precise, but the posterior estimates over the precision parameter (ν) converge faster to the correct values (already after 200 trials). Note that having the correct representation of both mean and precision parameters is more important in the regular condition as one can achieve higher improvements in the performance compared to the irregular condition, as we demonstrated previously in Marković et al. (2019).

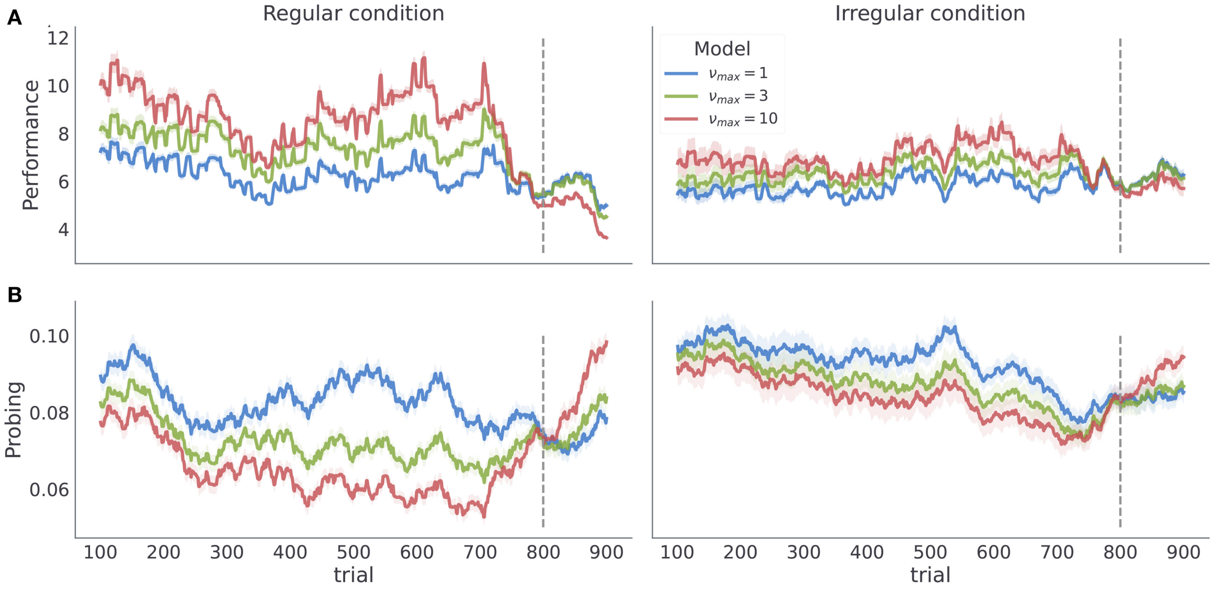


FIGURE 5
Model dependent dynamics of behavioral measures for varying ν_{max}. Each line corresponds to an average over $n = 50$ simulated trajectories with $\gamma = 5$, and $P_o = (0.1, 0.6, 0.15, 0.15)$. **(A)** Performance estimated as odds of generating a correct choice within a 200 trials long time window centered at trial index. **(B)** Probing, computed as odds of selecting the exploratory option within the 200 trials long time window. The shaded colored areas around the trajectories correspond to the 95% confidence interval.

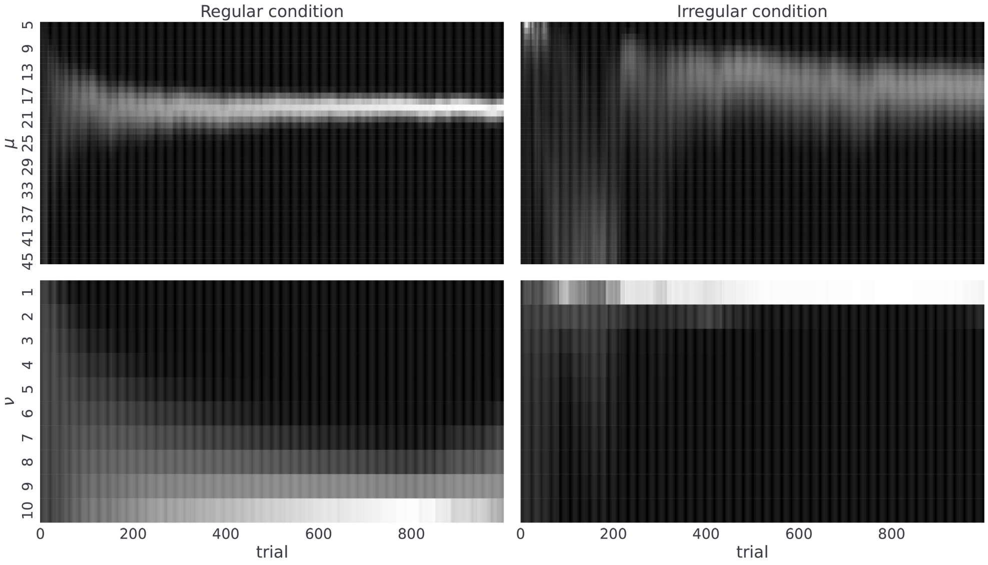


FIGURE 6
Posterior beliefs about temporal templates. Posterior beliefs of a single agent in the regular **(left)** and the irregular condition **(right)**. Posterior beliefs $q_t(m) = q_t(\mu, \nu)$ at each trial t over templates m are marginalized over precision parameter ν obtaining $q_t(\mu)$ **(top)** and mean parameter μ obtaining $q_t(\nu)$ **(bottom)**. The posterior beliefs are estimates obtained from a single run of the agent in both experimental conditions where we fixed the temporal prior parameter to $\nu_{max} = 10$, choice precision to $\gamma = 5$, and the preference vector to $P_o = (0.1, 0.6, 0.15, 0.15)$. The lighter the color the higher is the corresponding posterior probability for that parameter value.

2.2.3. Simulating the behavioral effect of prior preferences over outcomes

As mentioned above, the prior preference over outcomes $\boldsymbol{P}_o$ parameterize agents' motivation to collect rewards (generate correct choices) and collect information (engage with the exploratory option). Therefore, it is important to understand how prior preferences interact with performance and probing. We show that the more an agent engages with the exploratory options (i.e., the higher its preference for choice cues), the better its representation of latent temporal structure, and consequently the higher agent's performance. This is because selecting exploratory options maximally reduces the uncertainty about the latent state (which option has higher reward probability) which in turn allows an agent to learn a more accurate representation of the latent task dynamics. We visualize these dependencies in Figure 7, where we show what impact changing p_+ and p_- have on performance, probing, and the quality of temporal representation after 800 trials. In the Supplementary Figure 3 we show the same dependencies but with respect to changing p_+ and p_c, hopefully helping the reader to build an intuition about interactions between prior preference parameter and behavior. Note that in both figures we only consider cases in which $p_+ \geq p_-$ as this reflects higher prior preference for gains than for losses in the agent, which we expect to hold for all subjects.

2.3. Model-based analysis of subjects' choices

By estimating the prior beliefs—under a semi-Markovian generative model—that best explain observed choice behavior, we next ask whether human subjects can learn latent temporal regularities in the reversal learning tasks? An individual's capacity to learn correct temporal regularity corresponds to their behavior being associated with a less precise prior over temporal templates (Equation 1), that is, larger ν_{max}. An agent with imprecise prior over temporal templates is able to learn an accurate representation of a distribution of between-reversal-intervals, and to form expectations about the moment of reversals (see Figures 6, 7) in both conditions. Thus, we anticipated that between-subject variability in performance and probing would be reflected in different posterior estimates of the most likely ν_{max} value associated with the behavior of individual subjects.

Therefore, we first classify subjects based on the maximum a-posteriori estimate over possible values of $\nu_{max} \in \{1, \dots, 10\}$, as shown in Figure 8. For each subject we compute a posterior probability over ν_{max} and assign the subject the value of the temporal prior ν_{max} corresponding to the value with the highest exceedance probability (see Section 4.7). Using this procedure we find that 11 out of 41 subjects in the regular condition, and 1 out of 33 subjects in the irregular condition are assigned to the group with temporal prior $\nu_{max} > 1$. For the subjects in the regular condition this result suggests that about a quarter of subjects learned to anticipate reversals to a certain extent. As our aim is not to identify precisely participants' temporal prior, but simply to distinguish between subjects that learn temporal regularities ($\nu_{max} > 1$) from those that do not ($\nu_{max} = 1$), limiting the analysis to binary classification leads to the following classification accuracy in simulated data: (i) in the regular condition $\nu_{max} = 1, ACC = 1$, and $\nu_{max} > 1, ACC = 1$, (ii) in the irregular condition $\nu_{max} = 1, ACC = 1.0$ and $\nu_{max} > 1, ACC = 0.9$. Note that in regular condition we have around 10% chance of misclassifying a subject that actually has a less precise prior over temporal templates ($\nu_{max} > 1$).

The posterior estimates of model parameters shown in Figure 8 show that the majority of participants were assigned to the model class corresponding to the simplest HMM representation ($\nu_{max} = 1$) which assumes maximal irregularity. However, in the regular condition we also find a number of participants (27%) that exhibit more flexible priors, allowing us to form two subject groups. Importantly, when we plot the time course of both performance and probing, as shown in Figure 9, we find a trajectory of behavioral measures over the course of the experiment similar to what we see in simulated data. Namely, that the performance is higher and the probing reaches lower values in the group of participants associated with larger ν_{max} (compare with Figure 5—regular condition). We excluded the irregular condition from the visualization as we did not find sufficient number of subjects with associated with $nu_{max} > 1$. The behavioral trajectories of individual participants are shown in Supplementary Figure 1.

These findings show a good correspondence between simulated behavior for different parameterizations of the model ($\nu_{max} = 1$ vs $\nu_{max} > 1$ in Figure 5), and the participants' behavior associated with different model classes (Figure 9). There are two possible explanations for this: (i) the model inversion accurately captures the participants behavior and between-participant sensitivity to temporal regularities of the task, (ii) the group differences come from other free model parameters and do not correspond to differences in sensitivity to temporal structure. To exclude the second option we show in Figure 10 the mean of the posterior estimates of free model parameters γ, p_- and p_+. Note that in both experimental conditions we see a lack of separation between free model parameters associated with each model class.

There are a couple of interesting observations to be made from the posterior expectations of the free model parameters. First, we find in most participants rather large posterior estimates of choice precision γ, close to $\gamma = 5$ (see Figures 10B,D), suggesting that choice stochasticity is rather low in most participants. Low choice stochasticity means that choices are well aligned with the choice likelihoods encoded in terms of expected free energy (Equation 18). In

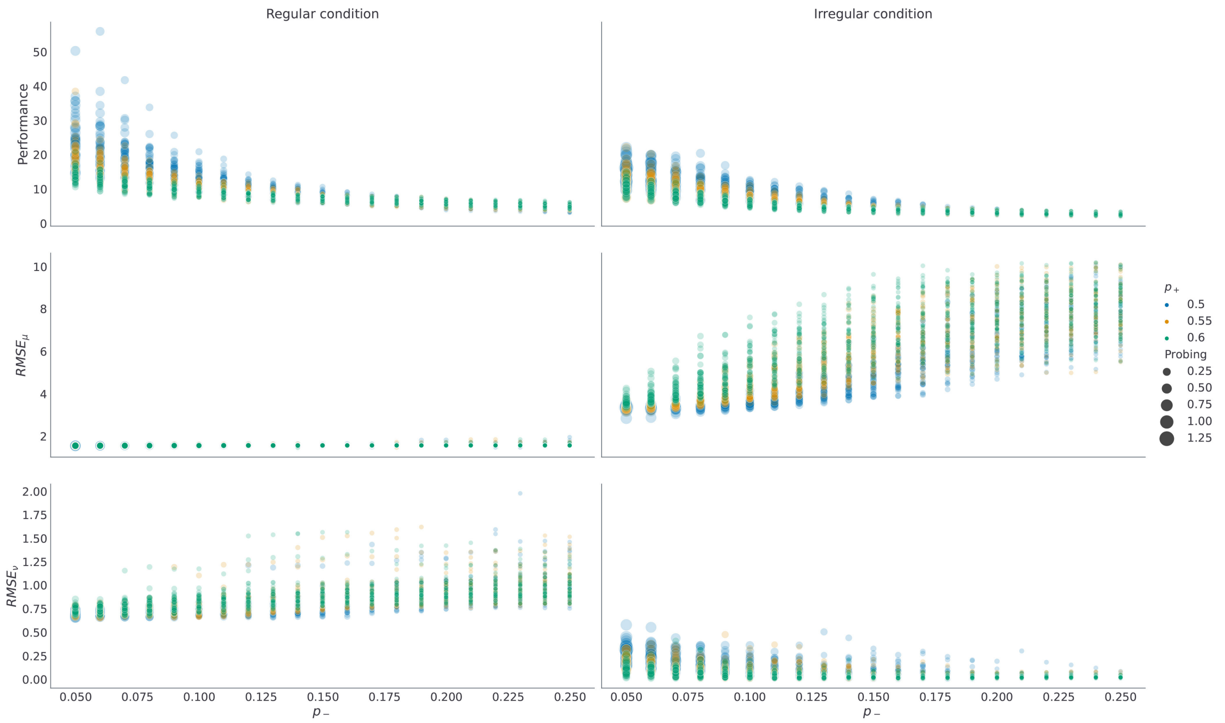


FIGURE 7
Dependence of performance, probing, and the quality of temporal representations on prior preferences over outcomes. Each dot in the plot corresponds to a single run with fixed prior preferences $P_o = \left(p_-, p_+, \frac{1}{2}p_c, \frac{1}{2}p_c\right)$, with temporal prior set to $\nu_{max} = 10$, and with decision precision set to $\gamma = 5$. For each possible pair ($(p_-, p_+) \in \left\{0.05, 0.06, \ldots, 0.25\right\} \bigotimes \left\{0.5, 0.55, 0.6\right\}$) we have repeated $n = 50$ simulations in each condition. Here $RMSE_{\mu,\nu}$ stands for the root mean square error of corresponding parameters μ, ν that define temporal template. The RMSE is computed using posterior probabilities $q_t(\mu, \nu)$ obtained at trial $t = 800$. The performance and the probing are computed as averages over responses from trial $t = 400$ until trial $t = 800$. Note that probing is increasing as we reduce p_- and keep p_+ fixed (circles of the same color), and as we reduce p_+ and keep p_- fixed, as the larger the sum $(p_+ + p_-)$ is, the lower is the preference for choice cues p_c, and hence the tendency of the agent to engage with the epistemic option. Higher probing (larger circle size) results in higher performance in both conditions (top row). Similarly, both $RMSE_\mu$ and $RMSE_\nu$ are lower for larger exploration odds, with the exception of $RMSE_\nu$ in irregular condition. Note that forming an accurate temporal representation is especially important in the regular condition, where forming correct anticipatory beliefs can substantially improve behavioral adaptation and simplify the problem of balancing between exploratory and exploitative choices. In contrast, in the irregular condition, having a precise representation of temporal structure does not impact performance substantially, and the agent performs better when engaging with the epistemic option more often.

other words, the chosen option is the option that minimizes expected free energy and the model is rather accurate in predicting behavioral responses. Second, the posterior estimates of outcome preference parameters p_-, and p_+ split subjects in two distinct groups, which correspond to their preference for receiving informative cues when selecting exploratory option. The 29 subjects who never engaged with the exploratory option have a higher preference for losses than for informative cues, hence $p_- \geq p_c$. We marked with the dashed gray line (Figures 10A,C) the limiting case of $p_- = p_c = \frac{1-p_+}{2}$, which separates the subjects which did not interact with the exploratory option (above the dashed line) and subjects that were relying on exploratory option to reduce their belief uncertainty (below the dashed line). Similarly, participants who prefer informative cues over gains would have prior preferences over cues in the region $p_c \geq p_+$. The dotted gray line (Figures 10A,C) marks the limiting case of $p_+ = p_c = \frac{1-p_-}{2}$. Note that only one subject in the irregular condition, and several subjects in the regular condition fall along this line.

3. Discussion

Sequential activity of neuronal assemblies is one of principled neuronal operations that support higher level cognitive functions (Eichenbaum, 2014; Buzsáki and Llinás, 2017) and allow humans to form complex spatio-temporal representation of our every day environment (Frölich et al., 2021). Akin to grid cells known to support representation of both spatial and non-spatial task states (Fu et al., 2021), time cells have been linked to temporal representation of state sequences critical for memory and decision-making (Eichenbaum, 2014). Importantly, in spite of these fruitful experimental findings we have

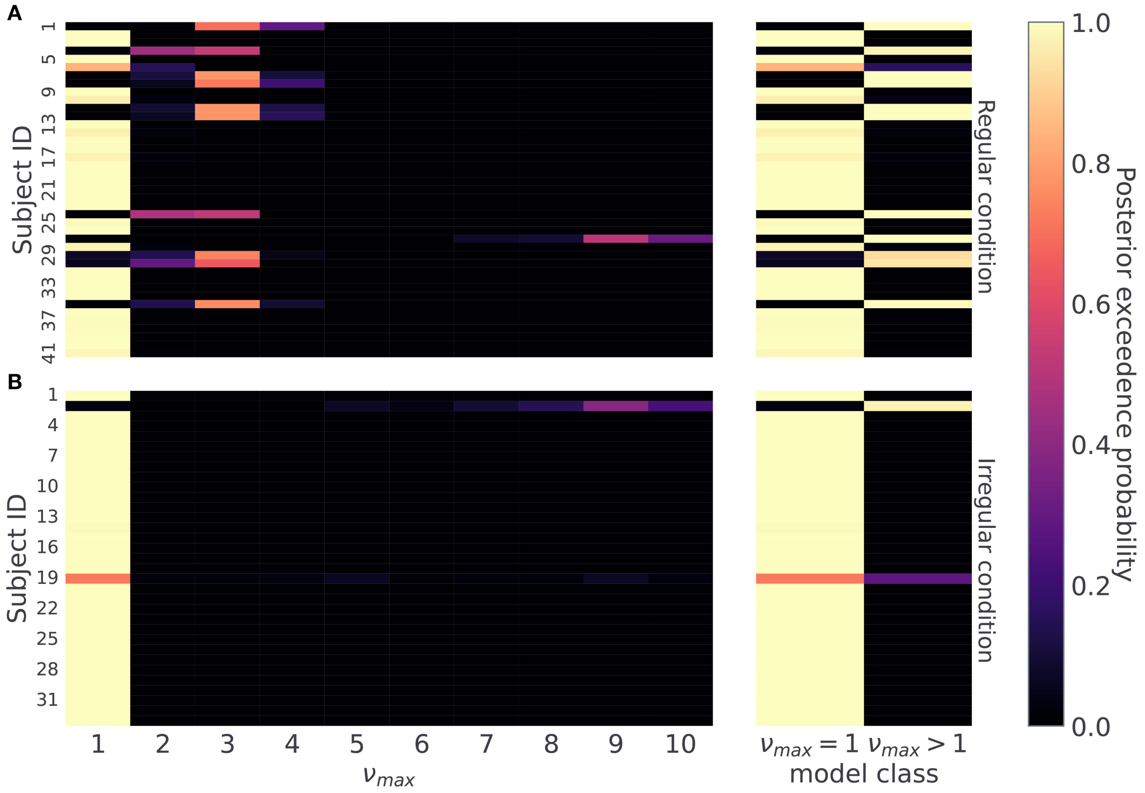


FIGURE 8

Posterior probability over temporal prior ν_{max}. Posterior probability of possible ν_{max} values for each subject, reflecting a subject's flexibility to learn latent temporal structure: **(A)** regular condition, and **(B)** irregular condition. On the right hand side, we combine posterior estimates into two classes, one for the limiting case $\nu_{max} = 1$, and another for all other options $\nu_{max} > 1$. This split differentiates subjects not sensitive to temporal regularities from the ones who a priori expected a regular temporal structure of reversals. Note that lighter colors correspond to higher posterior probability.

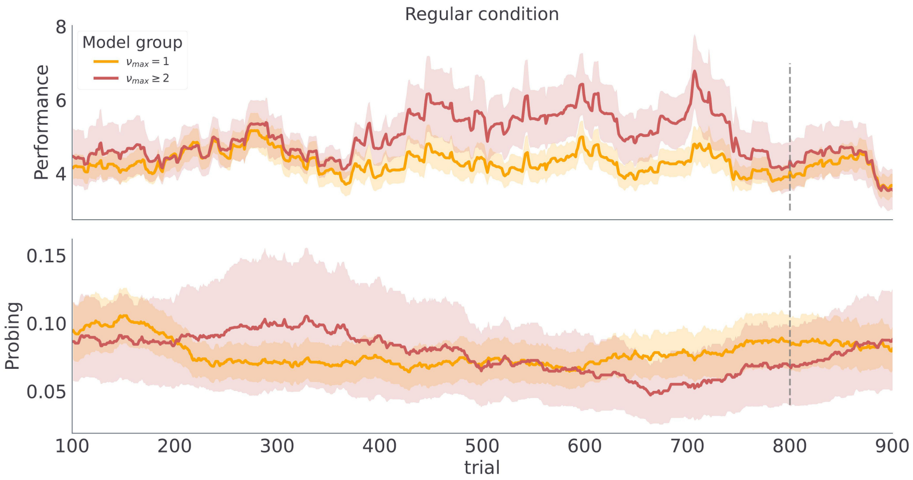


FIGURE 9

Category based mean estimate of behavioral measures. Each line corresponds to a model class average over behavioral trajectories of subjects assigned to that model class. Note the similarity of the trajectory profiles to the simulated trajectories in Figure 5 regular condition. The shaded colored areas around the trajectories correspond to the 95% confidence interval.

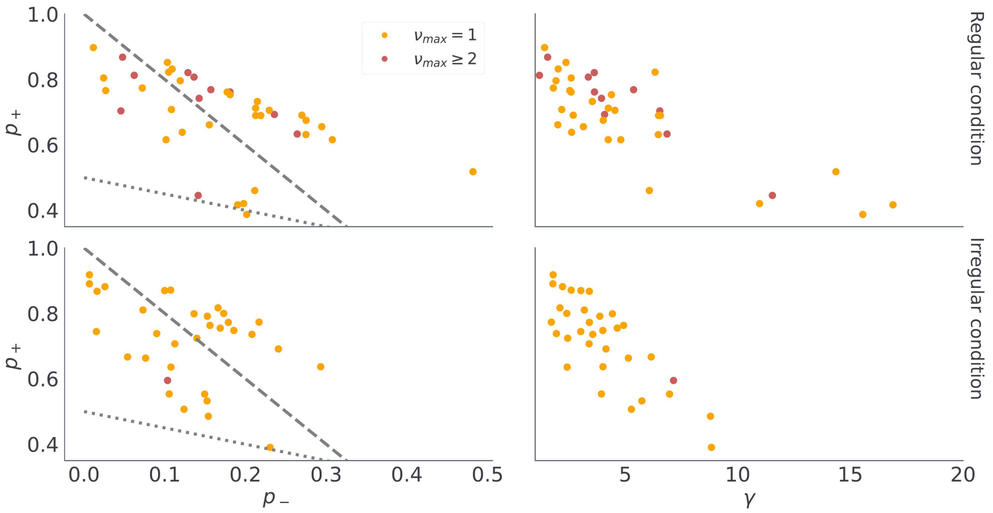


FIGURE 10
Posterior median of continuous model parameters. Each dot corresponds to the median of the $n = 1,000$ samples from the posterior distribution of γ, p_-, and p_+ for each participant. The dashed gray lines denotes equality between preferences for losses and epistemic cues, that is, when $p_- = \frac{p_c}{2}$ then $p_- = (1 - p_+)/2$. In turn, the doted grey line indicates equality between preferences for gains and epistemic cues, that is, when $p_+ = \frac{p_c}{2}$ then $p_- = 1 - 3p_+$. Note that we use the same color coding as in the previous figure to denote classification of participants into different model classes.

no clear computational understanding of how humans learn temporal structure in the service of successfully behavioral adaptation.

Here we introduced a novel computational model of behavior capable of learning latent temporal structure of a probabilistic reversal learning task with multiple reversals (Costa et al., 2015; Reiter et al., 2016, 2017; Vilà-Balló et al., 2017). The computational model combines hidden semi-Markov framework for representing latent temporal structure (Yu, 2015) and active inference for resolving exploration-exploitation trade-off (Friston et al., 2015, 2016). Crucially, the model can be used for investigating decision making in changing environments in any behavioral task that can be cast as a dynamic multi-armed bandit problem (Gupta et al., 2011; Markovic et al., 2021); of which the reversal learning tasks is a special case corresponding to a specific type of two-armed bandit problem.

The probabilistic reversal learning task, which we utilized to demonstrate flexibility of proposed model, is one of the most established paradigms for investigating human behavior in changing environments and quantifying cognitive disorders. We used model-based analysis of behavioral data to infer temporal expectations of subjects exposed to one of the two task variants: (i) with regular intervals between reversals, (ii) with irregular intervals between reversals. Notably, being able to form expectations about the moment of reversal is critical for achieving high performance in the probabilistic reversal learning task, which we illustrate using simulations. We demonstrated that participants behavior is highly heterogeneous reflecting the differences in participants expectations about temporal regularities. Crucially, the participants expectations about temporal regularities influence their ability to correctly learn latent temporal structure (especially relevant in the condition with regular between reversal intervals), and is reflected in their performance throughout the experiment.

We have extended the standard reversal learning task and incorporated an explicit exploratory option in addition to the two standard options whose choice results in monetary gain or loss. This exploratory option informs the participant about the currently correct choice. The additional behavioral response provides us with more direct access to the individual uncertainty about a correct choice and improves model selection. Interestingly, in addition to participants' diversity of temporal representation, we find stark differences in their preferences to engage with the exploratory option, suggesting individual differences for the value of information (Niv and Chan, 2011) and utilized strategies for resolving the exploration-exploitation trade-off. Critically, their epistemic preferences are not obviously correlated with the quality of the learned temporal structure, as in both groups participants show heterogeneous prior expectations about temporal regularities limiting the available temporal templates, hence the accuracy of temporal representations. However, the willingness to engage with the epistemic options does influence participants' performance, where higher engagement results in better performance. Therefore, these joint findings reveal distinct components of the computational mechanisms that underlie adaptive behavior in dynamic environments.

To recapitulate, we have effectively shown that it is possible to explain a subject's choice behavior in terms of their prior beliefs about temporal regularity, that is, a set of temporal templates they entertain, and other contingencies that characterize the (generic) paradigm at hand. This is potentially important because this kind of phenotyping could be deployed in a neurodevelopmental or psychiatric context to summarize any subject in terms of a small number of interpretable priors. Theoretically, this sort of phenotyping provides a sufficient description of a subject *via* the complete class theorem. The complete class theorem says that for any given pair of reward functions and behaviors there exists some priors that render the behavior Bayes optimal (Wald, 1947; Brown, 1981). To be Bayes optimal is to conform to the belief updates and action selection described by active inference. This means that there is always a set of prior beliefs that provide a sufficient account of any subject specific behavior.

Having said this, we have only explored a small subset of possible sets of temporal templates. We could apply the same technology (i.e., model inversion) to ask more general questions. For example, if some subjects a priori exclude from the templates the possibility of irregular reversals. There are other priors we could have explored that place various constraints on belief updating or divergences from particular prior beliefs. These might be interestingly related to notions of motivation, cognitive effort and resources in cognitive science (Pezzulo et al., 2018); however, this would require a specification of motivation, resources and effort in terms of belief updating, which is an outstanding challenge. Overall, we expected that, as we humans are exposed to predictable changes in our everyday environment, that there should be profound evidence that subjects utilize a higher order (semi-Markovian) model. The fact that we do not see that in the model selection results (in the regular condition the majority of subjects' behavior can be associated with the simplest Markovian assumption) suggests that a better experimental paradigm than the currently used reversal learning task is required. This paradigm should be more engaging and intuitively linked to distinct latent temporal regularities. A notable limitation of the current experimental paradigm is that it is not obvious to subjects that anticipating reversals can improve performance, or that potential performance improvement is sufficiently large to justify added effort required to keep track of higher-order statistics.

To accurately predict future it is critical not only to know that change might be coming but also when the change will occur. To anticipate the changes in our-everyday environments and adapt our behavior accordingly, it is critical to accurately estimate and represent elapsed time between relevant events. Although the presented model abstracts elapsed time as a hierarchically structured counting process, it is straightforward to model events duration in physical time, by using continues representation of the phase-type distribution. This way the underlying model corresponds to continues time semi-Markov processes (Hongler and Salama, 1996) where state transitions follow the master equations, allowing one to capture decision making in real-time. Notably, an implicit assumption we make here is that a simple counting process can represent elapsed time at multiple time scales. In fact, various experimental findings suggest that the brain employs counting mechanisms, represented over multiple timescales, and integrates those representations when generating behavior (Baldassano et al., 2017; Fountas et al., 2022). Similarly, a range of experimental findings has linked timing of events and hence forecasting the future to underlying Bayesian inference mechanisms (Jazayeri and Shadlen, 2010; Griffiths and Tenenbaum, 2011). Most recently, Maheu et al. (2022) has linked sequence learning and prediction in human subjects to an underlying hierarchical Bayesian inference model with distinct hypothesis spaces for statistics and rules corresponding to a set of deterministic temporal templates. The authors conclude that the hierarchical Bayesian inference mechanism underlies human ability to process sequence, similar to hierarchical semi-Markov framework proposed here.

Furthermore, in recent years, various neuroimaging studies have linked different neuro-cognitive domains, such as attention and working memory, to specific spatio-temporal expectations about underlying dynamics of the environment (Nobre and Van Ede, 2018). Interestingly, the human ability to estimate and reproduce elapsed time was also previously linked to reward discounting and intertemporal choice behavior (Ray and Bossaerts, 2011; Retz Lucci, 2013; Bermudez and Schultz, 2014). For example, McGuire and Kable (2015) demonstrated that "impulsivity" (reluctance to wait for a better reward), depends on the hidden statistics of delays—between an initial bad offer and a later but more valuable offer—which human participants experienced. Using a similar "limited offer" game (with a constant latent temporal statistics) and active inference representation of behavior (Schwartenbeck et al., 2015) have linked the dopaminergic midbrain activity with expected certainty about desired outcomes. In Mikhael and Gershman (2019), the authors have linked time perception and dopaminergic neuronal activity, demonstrating the role of value-based prediction errors in time representation. Furthermore, time perception and timed behavior have been linked to all major neuromodulatory systems (Meck, 1996) either directly using neuropharmacological manipulations (Crockett and Fehr, 2014) or indirectly using neurological disorders (Story et al., 2016) and aging research (Read and Read, 2004).

Together these findings provide important evidence for the role of temporal expectations in goal-directed decision making and let one speculate whether a range of aberrant behaviors might be related to an erroneous representation of the temporal structure of the task. Importantly, the computational behavioral model that we introduced here can emulate the learning of temporal structure, hence can become a potent tool linking aberrant behavior found in cognitive disorders to erroneous prior beliefs about the rules that govern the dynamics of the environment, as suggested by the active inference account of human behavior (Friston et al., 2015, 2016, 2017).

To conclude, the results presented here provide novel insights into computational mechanism underlying the human ability to learn hidden temporal structure of the environment and the computational principles they utilize for making decisions based on temporal representations. The fact that we find behavioral heterogeneity in a population of healthy young adults suggests a potential use of the proposed design and behavioral model for cognitive phenotyping and for revealing causes of aberrant behavior in clinical populations.

4. Methods and materials

4.1. Code availability statement

All code for reproducing the figures and running data analysis and simulation algorithms is available at https://github.com/dimarkov/pybefit.

4.2. Experiment

4.2.1. Probabilistic reversal learning

In the experimental task subjects were deciding between two cards shown on a screen, each showing a different stimulus (a geometric shape, e.g., rectangle, triangle, or a question mark) as shown in Figure 1. The reward probabilities associated with the two choice options were anti-correlated on all trials: whenever reward probability of choice A was high ($p_H = 0.8$) the reward probability of choice B was low ($p_L = 0.2$), and vice versa. Note that $p_H = 1 - p_L$ on all trials. The location of each stimulus on the screen (right or left side) was kept fixed over trials. After each choice the stimulus was highlighted and depicted for 1.5s minus the reaction time. The feedback in the form of a gain or a loss was shown for 0.5s. Similarly, the feedback after an exploratory choice was also shown for 0.5s. If no response occurred during the decision window of 3s, the message “too slow” was presented, and no outcome was delivered.

All subjects underwent a training session during which they had the opportunity to learn the statistics of the rewards associated with high p_H and low p_L reward probability choices. The set of stimuli used in the training phase differed from the one used during the testing phase. Subjects were instructed that they could either win or lose 10 cents on each trial, and that they will be paid the total amount of money they gained during the testing phase at the end of the experiment. Each subject performed 40 training trials with a single reversal after the 20th trial. Before the start of the testing phase subjects were told that the reward probabilities might change at regular intervals (in both conditions) over the course of the experiment. No other information about reversals or the correlation of choices and outcomes was provided. Thus, the subjects had no explicitly instructed knowledge about the anti-correlated reward probabilities or between-reversal-intervals before the experiment.

Note that, out of $n = 74$ participants $n_p = 24$ were exposed to the variant of the reversal learning task without epistemic option. This group of subjects belongs to an initial pilot study that used the standard two-choice task design. In the pilot study 14 subjects were assigned to the regular condition and 10 to the irregular condition. We decided to include the subjects from the pilot into the analysis, as we noticed that almost 30% of subjects, in the post pilot group, choose not to interact at all with the exploratory option, even when that was a possibility. We will not explore this finding here in more detail, but we can exclude their misunderstanding of the task as a potential confound, as we provided a detailed instructions and training before they performed the task (see Section 4.3 for more details).

4.3. Behavioral measures

To quantify behavior we have used two summary measures: (i) performance, defined as odds of making a correct choice, and (ii) probing, defined as odds of making an exploratory choice.

The process of computing *performance* is illustrated in Figure 11. We first label subjects’ responses as either correct or incorrect, depending on whether a card with higher reward probability was selected or not (see Figure 11A). Then we compute a probability of making a correct choice within a 201 trial window, centered at the current trial number t (see Figure 11B). Finally, for each trial we compute performance as odds of being correct (see Figure 11C).

The *probing* is computed in similar manner to performance, with the only difference that we label choices as either exploratory or exploitative depending on whether subjects have chosen the exploratory option (middle card in Figure 1), or not. Probing is defined as the odds of selecting the exploratory option within a 200 trials time window.

4.4. Behavioral model

To introduce the generative model of task dynamics, and subsequently derive the behavioral model *via* model inversion

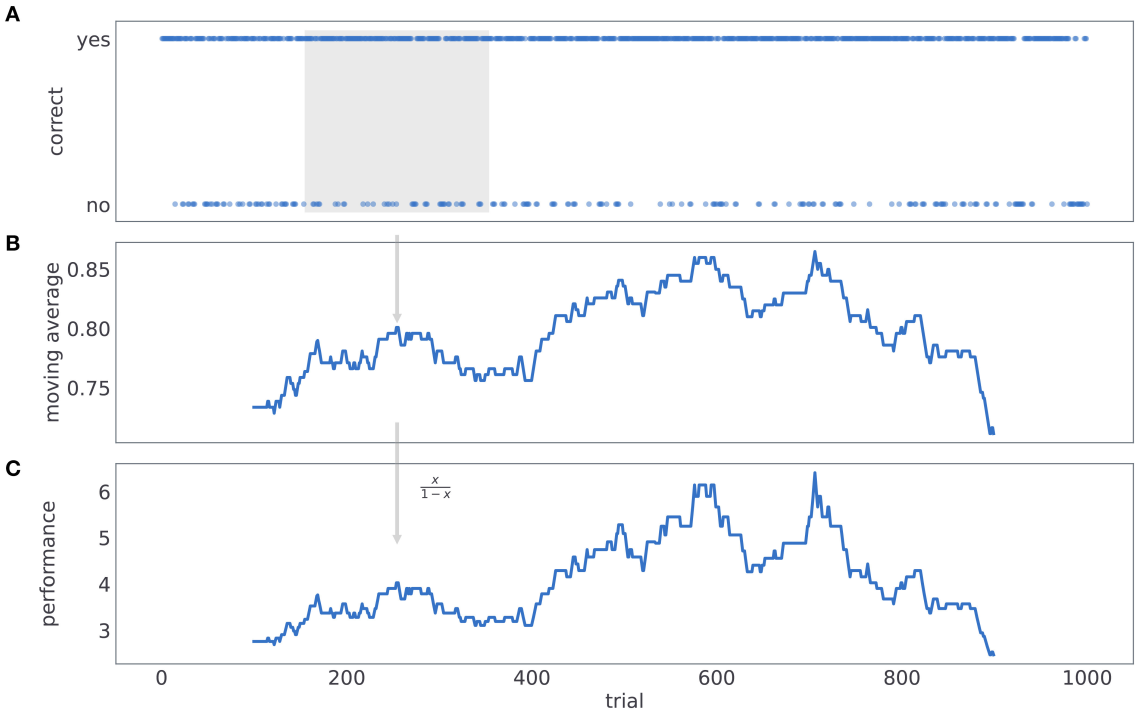


FIGURE 11
Computing behavioral performance. The process of computing *performance*. **(A)** We label subjects' responses as either correct or incorrect, depending on whether a card with higher reward probability was selected or not. **(B)** We compute a probability of making a correct choice within a 200 trial window, centered at the current trial number *t*. **(C)** For each trial we compute performance as odds of being correct.

methods, we will consider the following features of the task. At any trial the task environment is in one of the two possible states, defined as the configuration of reward contingencies. For example, state one corresponds to stimulus A being associated with a high reward probability p_H, and state two to stimulus B being associated with a low reward probability p_L. Subjects do not know in advance how likely rewards and losses are when making a correct choice compared to making an incorrect choice, and this is something they have to learn during the course of experiment. In other words, we also treat reward probabilities (p_H and p_L) as latent variables. Between trials the state can change, i.e., when a reversal occurs but only after a certain minimum number of trials has elapsed since the last state change. Depending on the experimental condition the between reversal duration will either be semi-regular (occurring every 20 trials with small variability) or irregular (occurring every 20 trials, but with maximal variability)

The explicit representation of state duration d enables us to associate changes in state transition probabilities with the current trial and the moment of the last change. The dependence of state transition probability on the number of trials since the last change corresponds to the formalism of hidden semi-Markov models (HSMM; Murphy, 2002; Yu, 2010), which allows mapping complex dynamics of non-stationary time series to a hierarchical, time aware, hidden Markov model. However, using an explicit representation of context duration is inefficient, as it requires an enormous state space representation. Here, we will instead adopt a phase-type representation of duration distribution (Varmazyar et al., 2019) which substitutes duration variable $d \in \{1, \ldots, \infty\}$ with a phase variable $f \in \left\{1, \ldots, f_{max}\right\}$, allowing for a finite state representation of an infinite duration state space.

In what follows we will define the components of the generative model (observation likelihood, the dynamics of latent variables, and the parameterization of the dynamics) and derive the corresponding update rules for latent variables and state, hence enabling the learning of different temporal contexts during the experiment. The graphical representation of the generative model is shown in Figure 4.

Practically we introduce four latent states, to describe the task on any trial:

- First, the configuration of reward contingencies can be in one of the two possible states. Hence, $s^1_t \in \{1, 2\}$ which describes which card is associated with high reward probability and which with low reward probability.
- Second, choosing one of the options on a given trial corresponds to setting the task in one of the three possible

choice states $s_t^2 \in \{1, 2, 3\}$ (chosen left card, chosen middle card—exploratory option, and chosen right card) corresponding to the chosen option. The choice of the option is deterministic and this state is always known with certainty after the choice is made.

- Third, current phase $f_t \in \{1, \ldots, \nu + 1\}$ of the task dynamics. The phase latent variable controls transitions of latent state s_t^1, where the change of state is only possible if the end phase ($f_t = \nu + 1$) is active on the current trial. Note that the larger the number of phases is (parameter $\nu \in 1, \ldots$) the more regular is the occurrence of reversals. We have limited here the number of phases by setting $\nu = 10$, as this is sufficiently large for accurate representation of reversal dynamics in regular condition.
- Fourth, temporal template m. Latent temporal template defines the frequency of reversals, μ (mean between-reversal duration) and the number of latent phases ν, that is the regularity of reversals.

4.4.1. Observation likelihood

The observation likelihood links latent states (s_t^1, and s_t^2) with probabilities of observing different possible outcomes in those states.

In the temporal reversal learning task there are four possible outcomes: (1) loss of 10 Eurocents, (2) gain of 10 Eurocents, (3) the correct card is left card, or (4) the correct card is the right card. Therefore, we define the observation likelihood as a categorical distribution

$$p\left(o_t | \boldsymbol{\rho}, s_t^1, s_t^2\right) = \prod_{i=1}^{4} \rho_{s_t^1, s_t^2, i}^{\delta_{o_t, i}} \tag{3}$$

where i denotes the outcome type, $o_t \in \{1, \ldots, 4\}$. The probabilities of different outcomes are parameterized *via* $\rho_{s_t^1, s_t^2, i}$, where each state tuple (s_t^1, s_t^2) corresponds to a unique probability of observing any of four possible outcomes. We define prior beliefs about outcome probabilities in the form of a product of Dirichlet distributions

$$p\left(\boldsymbol{\rho}\right) = \prod_{s^1=1}^{4} \prod_{s^2=1}^{3} Dir\left(\boldsymbol{\rho}_{s^1, s^2} | \boldsymbol{\alpha}_{s^1, s^2}^0\right). \tag{4}$$

We set the parameters of Dirichlet priors to the following values:

$$\boldsymbol{\alpha}_{s^1=1, s^2}^0 \equiv \begin{array}{c|ccc} o_t \backslash s^2 & 1 & 2 & 3 \\ \hline 1 & 6 & 1 & 32 \\ 2 & 32 & 1 & 6 \\ 3 & 1 & 1000 & 1 \\ 4 & 1 & 1 & 1 \end{array}, \quad \boldsymbol{\alpha}_{s^1=2, s^2}^0 \equiv \begin{array}{c|ccc} o_t \backslash s^2 & 1 & 2 & 3 \\ \hline 1 & 32 & 1 & 6 \\ 2 & 6 & 1 & 32 \\ 3 & 1 & 1 & 1 \\ 4 & 1 & 1000 & 1 \end{array} \tag{5}$$

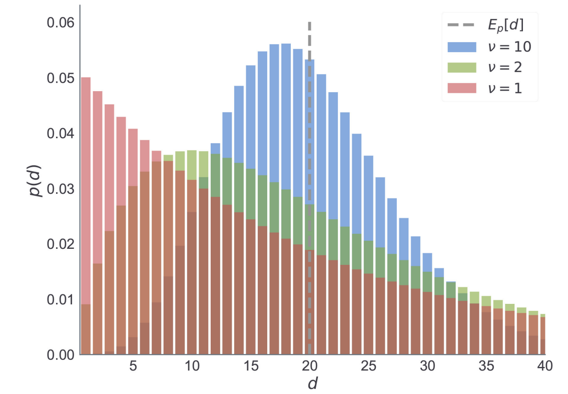


FIGURE 12
Negative binomial distribution. We illustrate here the changes in the negative binomial distribution as a function of shape parameter ν which is inversely proportional to the variance of between reversal durations. Note that for higher values of ν the distribution peaks around its expected value (dashed line). As the variance increases (green) the mode shifts toward zero. The limiting case of the negative binomial distribution in the form of geometric distribution (red) corresponds to $\nu = 1$. For all three cases we fixed the mean duration to the same value.

The above configuration for the parameterization of prior Dirichlet probabilities reflects an assumption that the participants have formed during training an initial—vague beliefs—about reward probabilities associated with different actions in different states. We assume that participants are highly certain that selecting the epistemic option does not return gain or loss (high value of $\boldsymbol{\alpha}_{s^1, s^2=2}$ for the corresponding outcome in both states). Furthermore, we assume that participants have formed good expectations gain/loss probabilities ($\langle p_H \rangle = \frac{32}{40} = 0.8$, and $\langle p_L \rangle = \frac{6}{40} = 0.15$), but that they are still uncertain about the exact values. Weak priors about outcome probabilities allow for ongoing adaptation of beliefs during the course of experiment.

4.4.2. Hidden state dynamics

To formalize the presence of sequential reversals, we define the phase dependent state transition probability as follows

$$p\left(s_t^1 | s_{t-1}^1, f_{t-1}\right) = \begin{cases} I_2, & \text{if } f_{t-1} = \nu + 1, \\ J_2 - I_2, & \text{if } f_{t-1} \leq \nu, \end{cases} \tag{6}$$

where I_2 denotes the 2 × 2 identity matrix and J_2 denotes the 2 × 2 all-ones matrix. The above relations describe a simple deterministic process for which the current state s_t^1 remains unchanged as long as the phase variable f_{t-1} remains below the end phase, $\nu + 1$. The transition between states occurs with certainty (e.g., if $s_{t-1}^1 = 1$ then $s_t^1 = 2$) once the end phase is reached, that is, when $f_{t-1} = \nu + 1$.

Although it is possible to condition state changes on a duration variable d, as demonstrated in Marković et al. (2019), such an explicit representation is inefficient as it requires large state spaces (Vaseghi, 1995; Yu and Kobayashi, 2003). Here we adopt the discrete phase-type (DPH) representation of duration distribution (Varmazyar et al., 2019). The DPH representation defines transitions between phase variables f_t and the following parameterization of phase transition probabilities corresponds to the DPH representation of the negative binomial distribution

$$p(f_t|f_{t-1}, m) = \begin{cases} \delta_m, & \text{if } f_{t-1} \leq \nu, \text{and } f_t = f_{t-1} + 1 \\ 1 - \delta_m, & \text{if } f_{t-1} \leq \nu, \text{and } f_t = f_{t-1} \\ \pi^m_{f_t}, & \text{if } f_{t-1} = \nu + 1, \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\pi^m_i = \binom{\nu}{i-1} (1 - \delta_m)^{\nu - i - 1} \delta_m^{i-1}$ for $i < \nu + 1$, and $\pi^m_{\nu+1} = 1 - \sum_{i=1}^{\nu} \pi^m_i$.

The corresponding negative binomial distribution of between-reversal duration can be expressed as follows

$$p_m(d) = \binom{d + \nu - 2}{d - 1} (1 - \delta_m)^{d-1} \delta_m^{\nu}; \quad d \in \{1, 2, \ldots\} \tag{8}$$

where the expected duration corresponds to

$$E_{p_m}[d] = \frac{\nu (1 - \delta_m)}{\delta_m} + 1 = \mu + 1; \quad \delta_m = \frac{\nu}{\mu + \nu}, \tag{9}$$

and variance, hence uncertainty about duration regularity, to

$$Var_{p_m}[d] = \mu + \frac{\mu^2}{\nu}. \tag{10}$$

Note that the parameter ν of the negative binomial distribution, acts as a precision parameter. We illustrate this in Figure 12.

The choice of prior beliefs about the between-reversal interval d in the form of a negative binomial distribution has interesting consequences on the dynamics of the marginal probability that a reversal will occur at some future point τ

$$\begin{aligned} \delta_m[\tau] &= p(s^1_{t+\tau} = 2|s^1_{t-1} = 2, f_{t-1} = \nu + 1, m) \\ &= \sum_{f_t, \ldots, f_{t+\tau}} \sum_{s^1_t, \ldots, s^1_{t+\tau-1}} p\left(s^1_t, f_t | s^1_{t-1} = 2, f_{t-1} = \nu + 1, m\right) \\ &\prod_{k=t+1}^{t+\tau} p\left(s^1_k, f_k | s^1_{k-1}, f_{k-1}, m\right) \end{aligned} \tag{11}$$

In Figure 13, we show the dependence of the future reversal probability $\delta_m[\tau]$ on the precision parameter ν, given a fixed mean duration $E_{p_m}[d] = 20$. Note that for $\nu = 1$ we get a constant transition probability, which corresponds to the expectations of change probabilities found in hidden Markov models. In contrast, for larger values of ν one obtains a trial-dependent, effective transition probability with values alternating between low and high probabilities in a periodic manner. This temporal dependence of the transition probability will affect the inference process. The agent will become insensitive to subsequent reversals occurring a few trials after the previous reversal, and highly sensitive to reversals occurring twenty to thirty trials after the previous reversal.

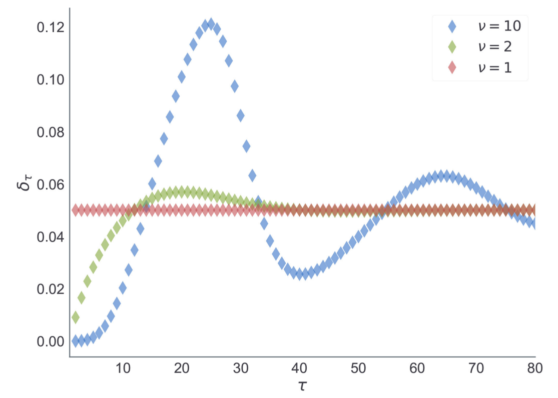


FIGURE 13
Expected transition probability at future trial τ. Estimate of the transition probability $\delta_m[\tau]$ (Equation 11), at a future trial τ conditioned upon a reversal at t and known initial state s^1_t. Each curve corresponds to estimates of the transition probability obtained from prior beliefs $p_m(d)$ shown in Figure 12.

Finally, the choice states s^2_t are fully dependent on the current choice $a_t \in \{1, 2, 3\}$, and we express the state transition probability as

$$p\left(s^2_t | s^2_{t-1}, a_t\right) = p\left(s^2_t | a_t\right) = \delta_{s^2_t, a_t}. \tag{12}$$

In practice this means that the agent is always certain about the choice it made and how that choice impacted the state of the task. Therefore, the posterior estimate over s^2_t can be trivially expressed as

$$q(s^2_t | a_t) = \delta_{s^2_t, a_t}.$$

4.4.3. Active inference

In active inference, agents form posterior beliefs both about latent states of the environment and about their own actions. In other words, both perception and action selection are cast as inference problems (Attias, 2003; Botvinick and Toussaint, 2012). Practically, we will use variational inference for defining update rules for beliefs (Blei et al., 2017; Friston et al., 2017). In what follows we will first introduce perception as minimization of the variational free energy (upper bound on log-marginal likelihood) with respect to posterior beliefs over latent states, and after that introduce action selection as minimization of the expected free energy (Smith et al., 2022), that is, expected surprisal about future outcomes.

We write the generative model of outcomes o_t on trial t as

$$\tilde{p}\left(o_t, s_t^1, s_t^2, f_t, m, \boldsymbol{\rho}\right) = p\left(o_t | s_t^1, s_t^2, \boldsymbol{\rho}\right) \tilde{p}_t\left(s_t^1 | f_t\right) p\left(s_t^2 | a_t\right) \\ \tilde{p}_t\left(f_t | m\right) \tilde{p}_t\left(m\right) \tilde{p}_t\left(\boldsymbol{\rho}\right), \tag{13}$$

where we use $\tilde{p}_t(\cdot)$ to denote prior beliefs conditioned on a sequence of past outcomes, $o_{1:t-1} = (o_1, \ldots, o_{t-1})$ and choices $a_{1:t-1} = (a_1, \ldots, a_{t-1})$. Given a choice a_t^* and an observed outcome o_t at trial t, the approximate posterior belief $q_t(x)$ over latent states $x = (s_t^1, s_t^2, f_t, \boldsymbol{\rho}, m)$ is obtained in two steps:

- We first compute the marginal likelihood with respect to $\tilde{p}_t\left(\boldsymbol{\rho}\right)$, and obtain the exact marginal posterior over discrete states using the Bayes rule

$$q_t\left(s_t^1, s_t^2, f_t, m\right) = \frac{\tilde{p}_t\left(o_t, s_t^1, s_t^2, f_t, m\right)}{\tilde{p}_t\left(o_t\right)}. \tag{14}$$

- Given the marginal posterior $q_t\left(s_t^1, s_t^2\right) = \sum_{f_t, m} q_t\left(s_t^1, s_t^2, f_t, m\right)$ we compute the posterior over outcome probabilities using the variational message passing update

$$q_t\left(\boldsymbol{\rho}\right) \propto \tilde{p}_t\left(\boldsymbol{\rho}\right) e^{\sum_{s_t^1, s_t^2} q\left(s_t^1, s_t^2\right) \ln p\left(o_t | s_t^1, s_t^2, \boldsymbol{\rho}\right)}. \tag{15}$$

As we initially defined the prior over outcome probabilities in the form of a Dirichlet distribution with parameters $\boldsymbol{\alpha^0}$, we can express the posterior estimate on every trial in the same functional form. Hence,

$$q_t\left(\boldsymbol{\rho}\right) = \prod_{s^1} \prod_{s^2} Dir\left(\boldsymbol{\rho}_{s^1, s^2} | \boldsymbol{\alpha}_{s^1, s^2}^t\right) \tag{16}$$

where

$$\begin{aligned} \alpha_{s^1, s^2 = a_t^*, i}^t &= \delta_{o_t, i} \cdot q\left(s_t^1 = s^1\right) + \alpha_{s^1, s^2 = a_t^*, i}^{t-1} \\ \alpha_{s^1, s^2 \neq a_t^*, i}^t &= \alpha_{s^1, s^2 \neq a_t^*, i}^{t-1} \end{aligned}, \tag{17}$$

and $\tilde{p}_t\left(\boldsymbol{\rho}\right) = q_{t-1}\left(\boldsymbol{\rho}\right)$. The above belief updating scheme corresponds to the variational surprise minimization learning algorithm (Liakoni et al., 2021; Markovic et al., 2021) adapted to the categorical likelihood and the Dirichlet prior.

4.4.4. Action selection

In active inference, decision strategies (behavioral policies) are chosen based on a single optimization principle: minimizing expected surprisal about observed and future outcomes, that is, the expected free energy (Schwartenbeck et al., 2019; Smith et al., 2022). Here, we will express the expected free energy of a choice a on trial t as

$$\begin{aligned} G_a = & \underbrace{D_{KL}\left(\tilde{p}_t(o_t|a) || P(o_t)\right)}_{\text{Risk}} + \underbrace{E_{\tilde{p}_t\left(s_t^1\right) \tilde{p}_t(\boldsymbol{\rho})}\left[H\left[o_t | \boldsymbol{\rho}, s_t^1, s_t^2 = a\right]\right]}_{\text{Ambiguity}} \\ \approx & - \underbrace{E_{\tilde{p}_t(o_t|a)}\left[\ln P(o_t)\right]}_{\text{Extrinsic value}} \\ & - \underbrace{E_{\tilde{p}_t(o_t|a)}\left[D_{KL}\left(q_t\left(s_t^1, s_t^2, f_t | o_t, a\right) || \tilde{p}_t\left(s_t^1, s_t^2, f_t | a\right)\right)\right]}_{\text{Epistemic value}} \\ & - \underbrace{E_{\tilde{p}_t(o_t|a)}\left[D_{KL}\left(q_t\left(\boldsymbol{\rho} | o_t, a\right) || \tilde{p}_t\left(\boldsymbol{\rho}\right)\right) + D_{KL}\left(q_t\left(m | o_t, a\right) || \tilde{p}_t\left(m\right)\right)\right]}_{\text{Novelty}} \end{aligned} \tag{18}$$

where $P(o_t)$ denotes prior preferences over outcomes, $H\left[o_t | \boldsymbol{\rho}, s_t^1, s_t^2\right]$ the entropy of outcome likelihood $p\left(o_t | \boldsymbol{\rho}, s_t^1, s_t^2\right)$, and $D_{KL}(p||q)$, stands for the Kullback-Leibler divergence between two probability densities : p and q. Note that action selection based on minimization of expected free energy would have an implicit dual imperative (see the different factorizations in Equation 18): On one hand, the expected free energy combines ambiguity and risk. On the other hand, it consists of information gain (epistemic value + novelty) and extrinsic value. Therefore, selecting actions that minimize the expected free energy dissolves the exploration-exploitation trade-off, as every action contains both expected value and information gain. This is a critical feature of action selection which allows us to account for epistemic choices as used in our experimental paradigm (see Figure 1).

At any trial t choice a_t is sampled from choice beliefs $p(a_t)$ (cf. planning as inference Attias, 2003; Botvinick and Toussaint, 2012) defined as

$$a_t \sim p\left(a_t | \gamma, \mathbf{P}_o, \nu_{max}\right) \propto e^{-\gamma G_a[\mathbf{P}_o, \nu_{max}, t]}, \tag{19}$$

where parameter γ corresponds to choice precision, which we will attribute to empirical choice behavior of participants. Therefore, for describing participants' behavior we assume that the action selection process is corrupted by external sources of noise; e.g., mental processes irrelevant for the task at hand. In our simulations we will fix γ to a reasonably large value, to achieve approximate free energy minimization as the following relation will be satisfied

$$a_t \approx argmin_a G_a, \text{ when } \gamma \gg 1. \tag{20}$$

Notably, here we consider the simplest form of active inference in which expected free energy is computed from a one-step-ahead prediction. This is a standard simplification for environments in which actions cannot interfere with the state transitions, as is the case in typical dynamic multi-armed bandit problems (Markovic et al., 2021).

To express the expected free energy, $G(a_t)$, in terms of beliefs about arm-specific reward probabilities, we will first constrain the prior preference to the following categorical distribution

$$P(o_t) = \prod_{o_t} [P_o]^{\delta_{o,o_t}}, \quad P_o = \left(p_-, p_+, \frac{1}{2}p_c, \frac{1}{2}p_c\right) \tag{21}$$

In active inference, prior preferences determine whether a particular outcome is attractive, that is, rewarding. Here we assume that all agents prefer gains ($o_t = 2$) over losses ($o_t = 1$). Hence, we constrain parameter values such that $p_+ > p_-$ holds always. The ratio $\frac{p_+}{p_c} = \lambda$ determines the balance between epistemic and pragmatic imperatives. When prior preferences for gains are very precise, corresponding to large λ, the agent becomes risk sensitive and will tend to forgo exploration if the risk is high (see Equation 18). Conversely, a low lambda corresponds to an agent which is less sensitive to risk and will engage in exploratory, epistemic behavior, until it has familiarized itself with the environment.

Given the following expressions for the marginal predictive likelihood,

$$\begin{gathered}\tilde{p}_t\left(o_t|a\right) = \sum_{s_t^1, s_t^2} \int p\left(o_t|\boldsymbol{\rho}, s_t^1, s_t^2\right) \tilde{p}_t\left(s_t^1\right) p\left(s_t^2|a\right) \tilde{p}_t\left(\boldsymbol{\rho}\right) \mathrm{d}\boldsymbol{\rho} \\ \tilde{p}_t\left(o_t|a\right) = \sum_{s=1}^{2} \tilde{p}_t\left(s_t^1 = s\right) \prod_{o=1}^{4} \left[\mu_{s,a,o}^{t-1}\right]^{\delta_{o_t,o}} \\ \mu_{s^1,s^2,o}^{t-1} = \frac{\alpha_{s^1,s^2,o}^{t-1}}{\sum_i \alpha_{s^1,s^2,i}^{t-1}}, \quad \bar{\mu}_{s^2,o}^{t-1} = \sum_{s^1} \tilde{p}_t\left(s_t^1 = s^1\right) \mu_{s^1,s^2,o}^{t-1}\end{gathered} \tag{22}$$

we get the following expressions for the expected free energy

$$\begin{gathered}G_t(a) = \sum_o \bar{\mu}_{a,o}^{t-1} \ln \frac{\bar{\mu}_{a,o}^{t-1}}{P_o} - \sum_{s^1} \tilde{p}\left(s_t^1 = s^1\right) \\ \sum_o \mu_{s^1,a,o}^{t-1} \left(\psi\left(\alpha_{s^1,a,o}^{t-1} + 1\right) - \psi\left(1 + \sum_j \alpha_{s^1,a,j}^{t-1}\right)\right)\end{gathered} \tag{23}$$

Above we have used the following relation

$$\int \mathrm{d}\boldsymbol{x} Dir\left(\boldsymbol{x}|\boldsymbol{\alpha}\right) x_i \ln x_i = \frac{\alpha_i}{\sum_j \alpha_j} \left(\psi\left(\alpha_i + 1\right) - \psi\left(1 + \sum_j \alpha_j\right)\right), \tag{24}$$

for computing ambiguity term in Equation (18).

4.5. Model inversion

To estimate subject-specific priors we effectively identified prior beliefs (i.e., ν_{max}, γ, and $\boldsymbol{P}_o$) that rendered the observed choices the most likely under active inference (i.e., under ideal Bayesian assumptions and the complete class theorem). In other words, for any given $\left(\nu_{max}, \gamma, \boldsymbol{P}_o\right)$, we can simulate belief updating — given subject specific outcomes to evaluate the expected free energy. The expected free energy then specifies the probability of choices at each trial. These probabilities can be used to assess the likelihood of any observed choice sequence of nth subject, conditioned upon a particular set of priors $[p\left(\nu_{max}, \gamma, \boldsymbol{P}_o\right)]$. One can then explore the space of priors (i.e., model parameters) to evaluate the marginal likelihood or model evidence for different combinations of priors.

In more detail, given a sequence of subjects' responses $\boldsymbol{A}^n = \left(a_1^n, \ldots, a_T^n\right)$, where n denotes subject index and $T = 1,000$ denotes the total number of trials, the response likelihood is defined as

$$P\left(\boldsymbol{A}^n|\gamma, \boldsymbol{P}_o, \nu_{max}\right) = \prod_{t=400}^{T} p\left(a_t = a_t^n|\gamma^n, \boldsymbol{P}_o^n, \nu_{max}^n\right). \tag{25}$$

Note that for estimating the posterior over model parameters $(\gamma, \boldsymbol{P}_o, \nu_{max})$ we ignore the first 400 responses from the likelihood. We expect that during these first trials, subjects are still getting used to the task, and potentially use additional strategies for representing the task and making choices. As we do not model all possible task representations, exclusion of initial trails reduces the noise in model comparison. Importantly, we do use the entire set of responses for computing belief trajectories of the active inference agents, that is, we expose the agent to the complete sequence of individual responses and the corresponding outcomes.

We define the prior over model parameters $\left(\nu_{max}^n, \gamma^n, \boldsymbol{P}_o^n\right)$ for the nth subject as follows:

$$p\left(\gamma^n, \boldsymbol{P}_o^n, \nu_{max}^n\right) = p\left(\gamma^n\right) p\left(\boldsymbol{P}_o\right) p(\nu_{max}^n), \tag{26}$$

where for a prior over choice precision parameter γ we use an inverse gamma distribution, thus

$$p\left(\gamma^n\right) \sim \Gamma^{-1}\left(2, 2\right), \tag{27}$$

and for the prior over prior preferences $\boldsymbol{P}_o$ we use a Dirichlet distribution, such that

$$\begin{gathered}p_i^n \sim Dir\left(\boldsymbol{p}|\boldsymbol{\beta}\right), \quad \beta_i = 1, i \in \{1, 2, 3\}, \\ \boldsymbol{P}_o^n = \left(\frac{p_1^n}{2}, \frac{p_1^n}{2} + p_2^n, \frac{p_3^n}{2}, \frac{p_3^n}{2}\right).\end{gathered} \tag{28}$$

With the above parameterization of prior preferences $\boldsymbol{P}_o$ we constrain the prior to reflect our expectations that all subjects will have higher preferences for gains than for losses, and that they will have equivalent preference associated with informative cues, that is, epistemic choices. Finally, we define a prior over the temporal precision parameter ν_{max} as a categorical distribution

$$\nu_{max}^n \sim Cat\left(\boldsymbol{r}^n\right) \tag{29}$$

where $\boldsymbol{r}^n = (r^n_1, \ldots, r^n_{10})$ denotes prior probability over possible ν_{max} values. Here we adopt the approach known as random effect Bayesian model selection (Stephan et al., 2009; Rigoux et al., 2014) which treats models (i.e., different ν_{max} values) as random effects that could differ between subjects and conditions, with an unknown population distribution. Hence, we introduce a condition specific hyper-priors over model probabilities in the form of a Dirichlet distribution

$$\begin{aligned} \tau &\sim C^+(0,1) \\ \boldsymbol{r}_1 &\sim Dir\left(\boldsymbol{r}_1|\boldsymbol{\alpha}_0/\tau\right) \\ \boldsymbol{r}_2 &\sim Dir\left(\boldsymbol{r}_2|\boldsymbol{\alpha}_0/\tau\right) \end{aligned} \tag{30}$$

where $\boldsymbol{r}_1$ corresponds to the condition with regular reversals and $\boldsymbol{r}_2$ to the condition with the irregular reversals. Finally, τ plays a role of a shrinkage parameter, that sets a non-zero probability to a configuration where all models have equal frequency in the population (in the limit $\tau \rightarrow 0$ we get $\boldsymbol{r}_1 = \boldsymbol{r}_2 = \frac{1}{10}$). The subject specific prior probability $\boldsymbol{r}^n$ corresponds to one of the two priors, based on the condition the subject was exposed to; hence, $\boldsymbol{r}^n \in \{\boldsymbol{r}_1, \boldsymbol{r}_2\}$.

To implement the above hierarchical generative model of subjects responses we used the probabilistic programming library Numpyro (Phan et al., 2019). Numpyro library provides an interface to multiple state-of-the-art inference schemes. For drawing samples from the posterior we have used Numpyro's implementation of the No-U-Turn sampler (NUTS) (Hoffman et al., 2014). NUTS is an self-tuning version of the Hamiltonian Monte Carlo, a popular Markov Chain Monte Carlo algorithm for avoiding random walks and sensitivity to between-parameter correlations. The limitation of NUTS is that it can only draw samples from continues random variables. Therefore, for implementation purposes we have to marginalize the generative model with respect to ν_{max}.

The marginalization results in the following marginal generative model:

$$\begin{aligned} \tau &\sim C^+(0,1) \\ \boldsymbol{r}_1 &\sim Dir\left(\boldsymbol{r}_1|\boldsymbol{\alpha}_0/\tau\right), \quad \alpha_{0,\nu} = 1, \nu \in \{1,\ldots,10\}, \\ \boldsymbol{r}_2 &\sim Dir\left(\boldsymbol{r}_2|\boldsymbol{\alpha}_0/\tau\right), \quad \alpha_{0,\nu} = 1, \nu \in \{1,\ldots,10\}, \\ \boldsymbol{r}^n &= f(\boldsymbol{r}_1, \boldsymbol{r}_2, n) \\ \boldsymbol{P}^n_o &\sim p\left(\boldsymbol{P}^n_o|\boldsymbol{\beta}\right), \quad \beta_i = 1, i \in \{1,2,3\}, \\ \gamma^n &\sim \Gamma^{-1}(2,2), \\ \boldsymbol{A}^n &\sim \sum_\nu r^n_\nu \prod_{t=400}^{1000} p\left(a_t|\gamma^n, \boldsymbol{P}^n_o, \nu_{max} = \nu\right). \end{aligned} \tag{31}$$

With the mixture model above we can unify the parameter estimation with the model comparison (selection). Given a sample from the posterior

$$\boldsymbol{r}^s_1, \boldsymbol{r}^s_2, \boldsymbol{P}^{n,s}_o, \gamma^{n,s} \sim p\left(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{P}^{1:N}_o, \gamma^{1:N}|\boldsymbol{A}^{1:N}\right) \tag{32}$$

we can obtain a sample from the marginal posterior probability over ν_{max} for the nth subject as

$$p^s\left(\nu^n_{max} = \nu|\boldsymbol{A}^{1:N}\right) = \frac{p\left(\boldsymbol{A}^n|\gamma^{n,s}, \boldsymbol{P}^{n,s}_o, \nu^n_{max} = \nu\right) r^{n,s}_\nu}{\sum_i p\left(\boldsymbol{A}^n|\gamma^{n,s}, \boldsymbol{P}^{n,s}_o, \nu^n_{max} = i\right) r^{n,s}_i}. \tag{33}$$

To classify subjects' behavior in terms of adaptability of temporal representations we use the exceedance probability (Rigoux et al., 2014) of the marginal posterior defined as

$$\begin{aligned} i^{n,s} &= argmax_\nu p^s\left(\nu^n_{max} = \nu|\boldsymbol{A}^{1:N}\right), \\ X^n_i &= \frac{1}{S}\sum_{s=1}^{S} \delta_{i,i^{n,s}}, \end{aligned} \tag{34}$$

thus, obtaining the probability that the ith model has the highest marginal posterior probability for the nth subject. The value X^n_i is plotted in Figure 8. Finally, the most likely precision parameter ν^n_{max} of the nth subject corresponds to $\nu^n_{max} = argmax_i X^n_i$ which we than used for classification as illustrated in Figures 9, 10.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/h526v/.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethical Board of Technical University Dresden. The patients/participants provided their written informed consent to participate in this study.

Author contributions

DM, AR, and SK contributed to conception and design of the study and wrote the sections of the manuscript. DM and AR collected the data. DM developed the model, performed the data analysis, and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The study was supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/3—Project A09 awarded to SK and SFB 940/3—Project B7 awarded to AR. SK acknowledges further support by DFG TRR 265/1 (Project ID 402170461, B09) and Germany's Excellence

Strategy—EXC 2050/1 (Project ID 390696704)—Cluster of Excellence Centre for Tactile Internet with Human-in-the-Loop (CeTI) of Technische Universität Dresden. AR acknowledges further support by the German Research Foundation (DFG RE 4449/1-1, RTG 2660-B2) and by a 2020 BBRF NAR-SAD Young Investigator Grant from the Brain and Behavior Research Foundation. This study was partially funded by funding opportunities for young scientists (Anschubfinanzierung) from the Department of Psychology of the Technische Universität Dresden.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbeh.2022.962494/full#supplementary-material

References

Attias, H. (2003). "Planning by probabilistic inference," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, eds C. M. Bishop and J. B. Frey (PMLR), 9–16. Available online at: http://proceedings.mlr.press/r4/attias03a/attias03a.pdf

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron* 95, 709–721. doi: 10.1016/j.neuron.2017.06.041

Bermudez, M. A., and Schultz, W. (2014). Timing in reward and decision processes. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20120468. doi: 10.1098/rstb.2012.0468

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Borst, J. P., and Anderson, J. R. (2015). The discovery of processing stages: analyzing EEG data with hidden semi-Markov models. *NeuroImage* 108, 60–73. doi: 10.1016/j.neuroimage.2014.12.029

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Bradtke, S., and Duff, M. (1994). "Reinforcement learning methods for continuous-time Markov decision problems," in *Advances in Neural Information Processing Systems*, eds G. Tesauro, D. Touretzky, and T. Leen (MIT Press). Available online at: https://proceedings.neurips.cc/paper/1994/file/07871915a8107172b3b5dc15a6574ad3-Paper.pdf

Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *Ann. Stat.* 9, 1289–1300. doi: 10.1214/aos/1176345645

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765. doi: 10.1038/nrn1764

Buzsáki, G., and Llinás, R. (2017). Space and time in the brain. *Science* 358, 482–485. doi: 10.1126/science.aan8869

Costa, V. D., Tran, V. L., Turchi, J., and Averbeck, B. B. (2015). Reversal learning and dopamine: a Bayesian perspective. *J. Neurosci.* 35, 2407–2416. doi: 10.1523/JNEUROSCI.1989-14.2015

Crockett, M. J., and Fehr, E. (2014). "Pharmacology of economic and social decision making," in *Neuroeconomics*, 2nd Edn, eds P. W. Glimcher and E. Fehr (San Diego, CA: Academic Press), 259–279. doi: 10.1016/B978-0-12-416008-8.00014-0

Daw, N., Courville, A. C., and Touretzky, D. (2002). "Timing and partial observability in the dopamine system," in *Advances in Neural Information Processing Systems*, eds S. Becker, S. Thrun, and K. Obermayer (MIT Press). Available online at: https://proceedings.neurips.cc/paper/2002/file/13111c20aee51aeb480ecbd988cd8cc9-Paper.pdf

Duong, T. V., Bui, H. H., Phung, D. Q., and Venkatesh, S. (2005). "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1*, 838–845. IEEE. doi: 10.1109/CVPR.2005.61

Eagleman, D. M. (2008). Human time perception and its illusions. *Curr. Opin. Neurobiol.* 18, 131–136. doi: 10.1016/j.conb.2008.06.002

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* 15, 732–744. doi: 10.1038/nrn3827

Eichenbaum, H. (2017). On the integration of space, time, and memory. *Neuron* 95, 1007–1018. doi: 10.1016/j.neuron.2017.06.036

Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., and Roseboom, W. (2022). A predictive processing model of episodic memory and time perception. *Neural Comput.* 34, 1501–1544. doi: 10.1162/neco_a_01514

Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends Cogn. Sci.* 20, 500–511. doi: 10.1016/j.tics.2016.05.001

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Frölich, S., Marković, D., and Kiebel, S. J. (2021). Neuronal sequence models for Bayesian online inference. *Front. Artif. Intell.* 4:50. doi: 10.3389/frai.2021.530937

Fu, Z., Beam, D., Chung, J. M., Reed, C. M., Mamelak, A. N., Adolphs, R., et al. (2021). The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* 376:6953. doi: 10.1126/science.abm9922

Gales, M., and Young, S. (2008). The application of hidden markov models in speech recognition. *Foundation. Trend. Sign. Process.* 1, 195–304. doi: 10.1561/2000000004

Griffiths, T. L., and Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: people combine prior knowledge with observations when estimating duration and extent. *J. Exp. Psychol.* 140:725. doi: 10.1037/a0024899

Gupta, N., Granmo, O.-C., and Agrawala, A. (2011). "Thompson sampling for dynamic multi-armed bandits," in *2011 10th International Conference on Machine Learning and Applications and Workshops.* p. 484–489. doi: 10.1109/ICMLA.2011.144

Hoffman, M. D., and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1593–1623. doi: 10.48550/arXiv.1111.4246

Hongler, M., and Salama, Y. (1996). Semi-Markov processes with phase-type waiting times. *Zeitsch. Angew. Math. Mech.* 76, 461–462.

Itskov, V., Curto, C., Pastalkova, E., and Buzsáki, G. (2011). Cell assembly sequences arising from spike threshold adaptation keep track of time in the hippocampus. *J. Neurosci.* 31, 2828–2834. doi: 10.1523/JNEUROSCI.3773-10.2011

Janssen, J., and Limnios, N. (1999). *Semi-Markov Models and Applications*. New York, NY: Springer. p. 404. doi: 10.1007/978-1-4613-3288-6

Jazayeri, M., and Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nat. Neurosci.* 13, 1020–1026. doi: 10.1038/nn.2590

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.* 112, 323–343. doi: 10.1007/s00422-018-0753-2

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi: 10.1371/journal.pcbi.1000209

Liakoni, V., Modirshanechi, A., Gerstner, W., and Brea, J. (2021). Learning in volatile environments with the bayes factor surprise. *Neural Comput.* 33, 269–340. doi: 10.1162/neco_a_01352

MacDonald, C. J., Fortin, N. J., Sakata, S., and Meck, W. H. (2014). Retrospective and prospective views on the role of the hippocampus in interval timing and memory for elapsed time. *Timing Time Percept.* 2, 51–61. doi: 10.1163/22134468-00002020

Maheu, M., Meyniel, F., and Dehaene, S. (2022). Rational arbitration between statistics and rules in human sequence processing. *Nat. Hum. Behav.* 6, 1087–1103. doi: 10.1038/s41562-021-01259-6

Marković, D., Reiter, A. M., and Kiebel, S. J. (2019). Predicting change: approximate inference under explicit representation of temporal structure in changing environments. *PLoS Comput. Biol.* 15:e1006707. doi: 10.1371/journal.pcbi.1006707

Markovic, D., Stojic, H., Schwoebel, S., and Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. *arXiv preprint arXiv:2101.08699*. doi: 10.1016/j.neunet.2021.08.018

McGuire, J. T., and Kable, J. W. (2012). Decision makers calibrate behavioral persistence on the basis of time-interval experience. *Cognition* 124, 216–226. doi: 10.1016/j.cognition.2012.03.008

McGuire, J. T., and Kable, J. W. (2015). Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. *Nat. Neurosci.* 18, 760–766. doi: 10.1038/nn.3994

Meck, W. H. (1996). Neuropharmacology of timing and time perception. *Cogn. Brain Res.* 3, 227–242. doi: 10.1016/0926-6410(96)00009-2

Mikhael, J. G., and Gershman, S. J. (2019). Adapting the flow of time with dopamine. *J. Neurophysiol.* 121, 1748–1760. doi: 10.1152/jn.00817.2018

Murphy, K. P. (2002). *Hidden semi-Markov models (HSMMs), vol. 2*. Citeseer. Available online at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.2942&rep=rep1&type=pdf

Niv, Y., and Chan, S. (2011). On the value of information and other rewards. *Nat. Neurosci.* 14, 1095–1097. doi: 10.1038/nn.2918

Nobre, A. C., and Van Ede, F. (2018). Anticipated moments: temporal structure in attention. *Nat. Rev. Neurosci.* 19:34. doi: 10.1038/nrn.2017.141

Parr, T., Markovic, D., Kiebel, S. J., and Friston, K. J. (2019). Neuronal message passing using mean-field, Bethe, and marginal approximations. *Sci. Rep.* 9, 1–18. doi: 10.1038/s41598-018-38246-3

Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* 22, 294–306. doi: 10.1016/j.tics.2018.01.009

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in Numpyro. *arXiv preprint arXiv:1912.11554*. doi: 10.48550/arXiv.1912.11554

Purcell, B. A., and Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4531–E4540. doi: 10.1073/pnas.1524685113

Ray, D., and Bossaerts, P. (2011). Positive temporal dependence of the biological clock implies hyperbolic discounting. *Front. Neurosci.* 5:2. doi: 10.3389/fnins.2011.00002

Read, D., and Read, N. L. (2004). Time discounting over the lifespan. *Organ. Behav. Hum. Decis. Process.* 94, 22–32. doi: 10.1016/j.obhdp.2004.01.002

Reiter, A. M., Deserno, L., Kallert, T., Heinze, H.-J., Heinz, A., and Schlagenhauf, F. (2016). Behavioral and neural signatures of reduced updating of alternative options in alcohol-dependent patients during flexible decision-making. *J. Neurosci.* 36, 10935–10948. doi: 10.1523/JNEUROSCI.4322-15.2016

Reiter, A. M., Heinze, H.-J., Schlagenhauf, F., and Deserno, L. (2017). Impaired flexible reward-based decision-making in binge eating disorder: evidence from computational modeling and functional neuroimaging. *Neuropsychopharmacology* 42, 628–637. doi: 10.1038/npp.2016.95

Retz Lucci, C. (2013). Time, self, and intertemporal choice. *Front. Neurosci.* 7:40. doi: 10.3389/fnins.2013.00040

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies-revisited. *Neuroimage* 84, 971–985. doi: 10.1016/j.neuroimage.2013.08.065

Schwartenbeck, P., FitzGerald, T., Dolan, R., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4:710. doi: 10.3389/fpsyg.2013.00710

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., and Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cereb. Cortex* 25, 3434–3445. doi: 10.1093/cercor/bhu159

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., and Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife* 8:e41703. doi: 10.7554/eLife.41703

Shappell, H., Caffo, B. S., Pekar, J. J., and Lindquist, M. A. (2019). Improved state change estimation in dynamic functional connectivity using hidden semi-Markov models. *NeuroImage* 191, 243–257. doi: 10.1016/j.neuroimage.2019.02.013

Shi, Z., Church, R. M., and Meck, W. H. (2013). Bayesian optimization of time perception. *Trends Cogn. Sci.* 17, 556–564. doi: 10.1016/j.tics.2013.09.009

Smith, R., Friston, K. J., and Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *J. Math. Psychol.* 107:102632. doi: 10.1016/j.jmp.2021.102632

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage* 46, 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025

Story, G. W., Moutoussis, M., and Dolan, R. J. (2016). A computational analysis of aberrant delay discounting in psychiatric disorders. *Front. Psychol.* 6:1948. doi: 10.3389/fpsyg.2015.01948

Varmazyar, M., Akhavan-Tabatabaei, R., Salmasi, N., and Modarres, M. (2019). Classification and properties of acyclic discrete phase-type distributions based on geometric and shifted geometric distributions. *J. Indus. Eng. Int.* 15, 651–665. doi: 10.1007/s40092-018-0299-x

Vaseghi, S. (1995). State duration modelling in hidden Markov models. *Signal Process.* 41, 31–41. doi: 10.1016/0165-1684(94)00088-H

Vilà-Balló, A., Mas-Herrero, E., Ripollés, P., Simó, M., Miró, J., Cucurell, D., et al. (2017). Unraveling the role of the hippocampus in reversal learning. *J. Neurosci.* 37, 6686–6697. doi: 10.1523/JNEUROSCI.3212-16.2017

Wald, A. (1947). An essentially complete class of admissible decision functions. *Ann. Math. Stat.* 18, 549–555. doi: 10.1214/aoms/1177730345

Yu, S.-Z. (2010). Hidden Semi-Markov models. *Artif. Intell.* 174, 215–243. doi: 10.1016/j.artint.2009.11.011

Yu, S.-Z. (2015). *Hidden Semi-Markov Models: Theory, Algorithms and Applications*, 1st Edn. Elsevier Science Publishers B. V. doi: 10.1016/B978-0-12-802767-7.00002-4

Yu, S.-Z., and Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Process. Lett.* 10, 11–14. doi: 10.1109/LSP.2002.806705