ARTICLES
https://doi.org/10.1038/s41593-022-01085-7

nature neuroscience

Efficient coding of cognitive variables underlies dopamine response and choice behavior

Asma Motiwala[1,2] ✉, Sofia Soares[1,3], Bassam V. Atallah[1], Joseph J. Paton[1,4] ✉ and Christian K. Machens[1,4] ✉

Reward expectations based on internal knowledge of the external environment are a core component of adaptive behavior. However, internal knowledge may be inaccurate or incomplete due to errors in sensory measurements. Some features of the environment may also be encoded inaccurately to minimize representational costs associated with their processing. In this study, we investigated how reward expectations are affected by features of internal representations by studying behavior and dopaminergic activity while mice make time-based decisions. We show that several possible representations allow a reinforcement learning agent to model animals' overall performance during the task. However, only a small subset of highly compressed representations simultaneously reproduced the co-variability in animals' choice behavior and dopaminergic activity. Strikingly, these representations predict an unusual distribution of response times that closely match animals' behavior. These results inform how constraints of representational efficiency may be expressed in encoding representations of dynamic cognitive variables used for reward-based computations.

The theory of reinforcement learning (RL) provides a set of algorithms that may inform how animals learn to interact with their environment using reward feedback. A key component in many RL algorithms is a reward prediction error (RPE) that updates representations of value via temporal differences[1,2] (TDs). Correlates of TD RPEs have been found in the phasic activity of midbrain dopaminergic (DA) neurons[3–5], and electrical and optogenetic manipulations of these neurons can produce learning about the value of actions[6–8]. These data have provided compelling evidence that neural systems function similarly to TD RL algorithms. Indeed, a large body of research on DA signaling supports the hypothesis that reward-based decision-making in neural circuits is well-described by the framework of RL (see refs. [9,10] for reviews).

A key challenge in explaining DA activity in terms of RPEs is that RPEs depend on internal representations of the environment. However, these internal representations often cannot be directly characterized. More generally, understanding how animals construct internal representations to guide adaptive behavior is a key outstanding goal of cognitive and systems neuroscience. Within the RL framework, the nature of internal representations places constraints on both reward expectations and RPE signals that may be encoded in neural circuits[11–13]. Thus, examining activity of midbrain DA neurons in terms of RPEs during carefully controlled tasks can be a powerful means for describing the principles of internal representations that guide cognition and behavior[14–16].

As an example of such a task, we studied mice making decisions based on internal estimates of elapsed time. Previous work has shown that population activity in cortical and striatal circuits in time-based tasks shows rich time-varying activity[17–20]. We constructed RL agents that encode internal representations consistent with these observed patterns of activity and trained them on the interval discrimination task performed by mice. We examined how changing different aspects of the representations used by model agents can yield different predictions about RPEs and their relation to behavior on a trial-by-trial basis. We found that only agents with internal representations that were efficient in terms of representational cost at the expense of being inaccurate in encoding certain aspects of the task showed RPEs that match the profile of DA activity recorded in mice and its relation to behavior. Moreover, TD learning using this efficient representation predicted a pattern in procrastination of choices that closely matched animals' behavior.

Representational efficiency has been extensively studied in sensory systems where reconstruction accuracy of sensory variables is balanced with representation cost[21–24]. In many of these studies, representational resources are allocated based on the statistics of the sensory inputs. Other studies have shown that allocating representational resources based on behavioral salience can reproduce representations in some systems[25–27]. Based on these results, it has been proposed that constraints of representational efficiency are likely to affect animals' reward expectations as well[28–30]. However, it is unclear which redundancies in variables needed for reward-based computations are being exploited to achieve efficient representations and how such representations affect reward expectations. Our results provide empirical support for encoding schemes where representational efficiency is achieved such that only the overall number of rewards obtained is preserved (or RPEs are minimized), a finding with wide-reaching implications for the more general problem of understanding the neural mechanisms underlying cognition.

Results

We analyzed behavior and DA activity of mice performing a time interval discrimination task (Fig. 1a). On each trial, animals indicated whether the interval between two tones was longer or shorter than 1.5 seconds. Animals reported their decisions for 'long' or 'short' intervals in one of two choice ports. For the longest and shortest intervals, animals almost always chose the correct port, but, as

[1]Champalimaud Neuroscience Programme, Champalimaud Foundation, Lisbon, Portugal. [2]Present address: Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Present address: Harvard Medical School, Boston, MA, USA. [4]These authors contributed equally: Joseph Paton, Christian Machens. ✉e-mail: amotiwala@cmu.edu; joe.paton@neuro.fchampalimaud.org; christian.machens@neuro.fchampalimaud.org

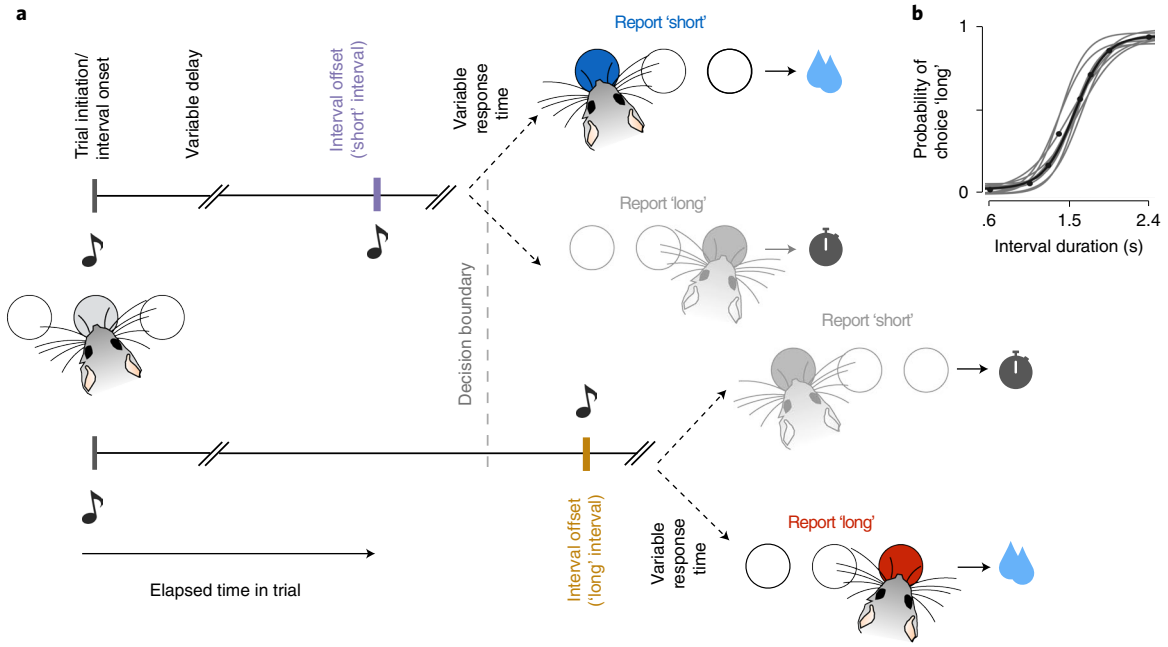

Fig. 1 | Rodents were trained to classify interval durations. a, Schematic illustrating the timeline of the main events during the interval discrimination task. Animals are presented with three ports and are required to initiate each trial in the central port. The central nose poke triggers a tone, and, after a variable interval, a second tone is presented. These intervals can be either longer or shorter than the decision boundary. Animals have to report 'short' or 'long' judgments in the two lateral ports based on their estimate of the time elapsed between the two tones. Correct choices result in a water reward and incorrect choices result in a timeout. **b**, Psychometric curves of animals performing the task. Gray lines indicate sigmoid fits to behavior of individual animals ($n = 6$), and black line and dots indicate average over all animals.

intervals approached the decision boundary, choices became more variable as captured by animals' psychometric functions (Fig. 1b).

Previous work has shown that animals' behavioral reports of elapsed time vary from trial to trial. Moreover, neural correlates of such variability have been found in the population activity of striatal circuits[17] and the activity of midbrain DA neurons in the substantia nigra pars compacta (SNc)[31]. However, the structure of internal representations that would be required to explain DA RPEs in relation to behavior during this task is unknown. To address this, we built several RL models with various internal representations of the task environment. We compared both behavior and RPEs produced by those models to the behavior of mice and activity of DA neurons in the SNc[31].

Reward expectations are modulated by task cues and internal estimates. TD RPEs can arise due to discrepancies among the probability, amount, or timing of expected and actual rewards or if an unpredictable change in the environment leads to a change in expected future rewards. During the interval discrimination task, the timing of trial onset was unpredictable[31]. Hence, the cue at interval/trial onset, which predicts a potentially forthcoming reward, should, thereby, cause changes in animals' reward expectations and concomitant RPEs. Consistent with this reasoning, we found that the tone marking interval/trial onset elicited a phasic DA response (Fig. 2a). Because the tone at interval onset was identical for all trials, the phasic DA response was not modulated by stimulus identity (Fig. 2a).

After interval onset, the task required animals to maintain an ongoing estimate of elapsed time to guide their decisions. This internal estimate may be used not only to guide decisions but also to encode time-varying expectations of future rewards. During the interval, expectations of future rewards should reflect average rewards expected from all intervals longer than the current estimate of the interval duration. At interval offset, however, animals' reward expectations should reflect an estimate of average rewards only from the currently estimated interval duration. Hence, at interval offset, the change in reward expectations should generate RPEs. Because reward expectations, both before and after interval offset, depend on interval duration, we expected DA responses at interval offset to also vary as a function of interval duration, and, indeed, they did (Fig. 2a). More specifically, two trends stood out. First, DA responses were larger for intervals farther from the decision boundary than for those close to the boundary. Second, the overall magnitude of responses was lower for 'long' intervals than 'short' intervals. Moreover, trial-to-trial variability in magnitude of DA responses also varied systematically with animals' reported judgments. For each interval duration presented, if trials are split into two groups based on the magnitude of DA response at interval offset, the psychometric function of trials with larger responses is shifted right relative to that corresponding to trials with lower responses (Fig. 2b). In other words, trial-to-trial variability in magnitude of DA response for each stimulus is predictive of the 'bias' in duration judgments.

At interval offset, because animals have not yet received any reward feedback, RPEs and DA responses must be based on internal variables. These internal variables include the choice accuracy or decision confidence, which is represented in cortical areas[32,33] and which influences DA activity[34]. The internal variables also include the hazard rate—that is, the probability of a cue occurring at a certain time, given that it has not occurred yet[35]. The hazard rate likewise influences DA activity[15,36,37].

Dopamine activity and its relation to choices cannot be predicted by directly combining choice accuracy and predictability of interval offset in time. If RPEs were influenced only by choice accuracy, we would expect RPEs to be lower for estimates of interval duration

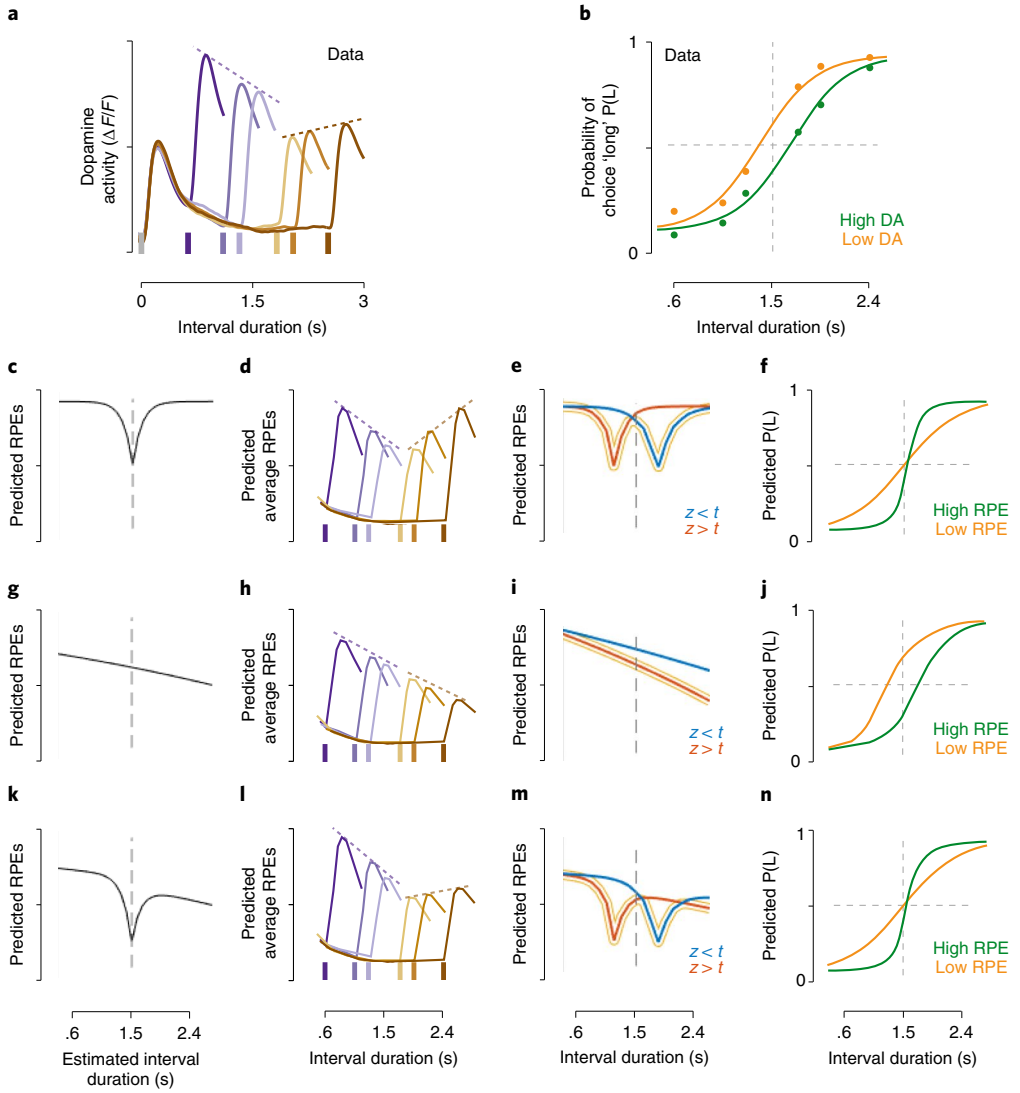

Fig. 2 | Reward prediction errors at interval offset can be modulated by choice accuracy and hazard rate of interval offset. a, Average dopamine responses for each of the intervals presented during the task. The initial peak occurs at interval onset (the time of interval onset is shown by the gray tick), and the following six peaks occur at each of the presented interval offsets (the times for each of the interval offsets used in the task are indicated using the colored ticks). The dashed lines highlight the overall profile of the magnitude of responses at interval offset. **b**, Psychometric functions for all trials in which high (green) or low (orange) DA responses were measured at interval offset. A clear difference in bias emerges from these two groups of trials. **c**, If we assume that animals do not maintain any prediction of the arrival times of interval offsets, but do encode estimates of choice accuracy for different estimates of elapsed time, the RPEs that we would predict would vary as a function of their internal estimates of elapsed time as shown here (for details, see Supplementary Fig. 1a–d). **d**, Hypothetical modulation of average RPEs at the six interval offsets presented in the task if these were based purely on estimates of an agent's choice accuracy. The averages in this case would be over the trial-to-trial variability in animals' estimates of elapsed time. The dashed lines highlight that, unlike in the data, the overall profile of magnitude of RPE at interval offset would be symmetric around the decision boundary in this case. **e**, Because animals' reward expectations evolve as a function of their internal estimates of elapsed time, shown here is how their estimates would evolve as a function of real time on two example trial types where the animal may overestimate (red) or underestimate (blue) elapsed time. For any interval, whether interval duration is overestimated or underestimated will systematically influence the magnitude of RPE at interval offset. Hence, for every timestep, the curve that corresponds to low RPE is highlighted in yellow. **f**, For every interval offset presented, if trials are split based on the magnitude of RPE, we would find that high RPE would go along with estimates of interval duration that are farther from the boundary than those trials on which RPE is lower. Hence, when trials are split into low-magnitude and high-magnitude RPE trials, the slope of the psychometric curves of the two groups would differ. **g**–**j**, Similar to **c**–**f** but for RPEs that are generated entirely due to temporal predictability of interval offsets. **k**–**n**, Similar to **c**–**f** and **g**–**j** but for RPEs that would be generated if the agent took into account both choice accuracy and temporal predictability of interval offsets. For more details regarding the shape of the predicted RPEs in **c**,**g**,**k**, see Supplementary Fig. 1; for more details regarding the predicted relationship between single-trial magnitude of RPE and choice shown in **f**,**j**,**n**, see Supplementary Fig. 2.

close to the boundary than those farther away (Fig. 2c and Extended Data Fig. 1a–d). In turn, the average RPEs predicted in the experiment would reflect this trend (Fig. 2d). Furthermore, for any given interval offset, low-RPE trials will correspond to trials wherein the agent's estimate of the duration is closer to the boundary and, hence, will be associated with lower choice accuracy and high choice variability. Consequently, grouping trials based on the magnitude of RPE within each interval, we would see a difference in the slope

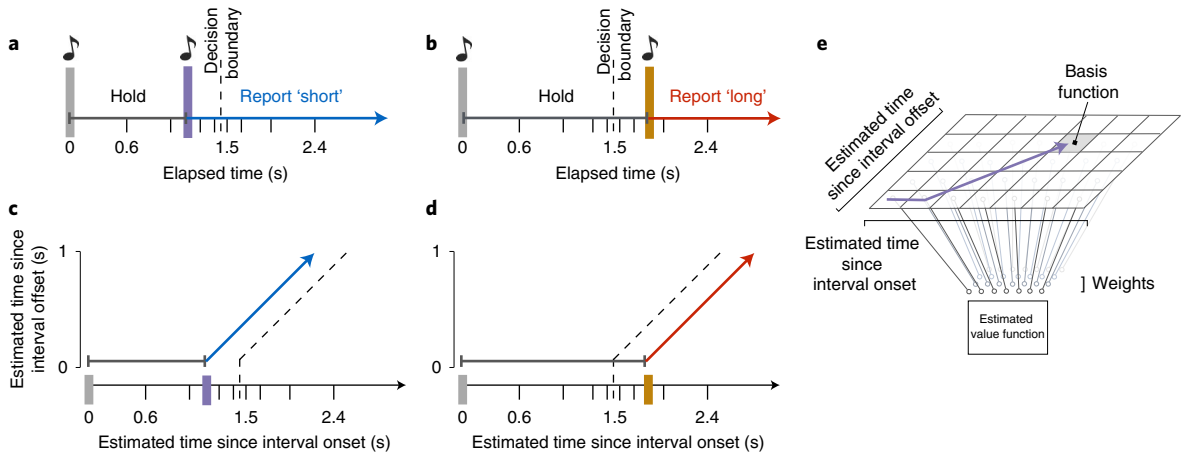

Fig. 3 | Task variables can be represented unambiguously using a two-dimensional state space. a, Timeline of an example trial in which an interval shorter than the decision boundary is presented. The color of the axis indicates the optimal action in that period of the trial (gray, wait; blue, short; red, long). **b**, Timeline of an example trial in which an interval longer than the decision boundary is presented. **c**, Illustration showing how the example trial in **a** is represented in the agent's state space. The agent's state is given by its internal estimates of time since interval onset (*x* axis) and time since interval offset (*y* axis). Because of variability in time estimation, these internal estimates are distinct from real time. The trajectory shows how the two-dimensional state variable changes over time and in response to interval onset and offset. The color of the arrows shows the correspondence between the segments of the trajectory in state space and the associated segments on the timeline of the example trial in **a**. Each location in the two-dimensional state space provides an unambiguous representation so that the agent can determine the optimal action. The internal decision boundary that would allow the agent to make optimal choices using this state representation is shown by the dashed line. **d**, Illustration showing how the example trial in **b** would be represented in the agent's state space, using the same conventions as in **c**. **e**, Because the state space is continuous, we approximate the value function by a linear combination of basis functions that are non-overlapping tile bases. The agent needs to estimate as many parameters or weights as there are basis functions. Here, the tiling of the basis functions is equally dense along the two axes shown, and, hence, the number of parameters that need to be estimated are N^2, where N is the number of basis functions tiling each dimension. This arrangement corresponds to the unambiguous representation used by the model. Shown in purple is an example trajectory through the state space. The basis function that would be active at the last timestep of this trajectory is highlighted in gray.

of the psychometric curve between low-RPE and high-RPE trials (Fig. 2e,f and Extended Data Fig. 2a–d).

Conversely, if RPEs were influenced only by the hazard rate of interval offsets, we would expect average RPEs to monotonically decrease with interval duration (Fig. 2g,h and Extended Data Fig. 1e–h). In this case, trial-to-trial variability in RPEs should predict a difference in bias in duration judgements (Fig. 2i,j and Extended Data Fig. 2e–h).

Most likely, RPEs are influenced by both choice accuracy and hazard rate. We would then expect that the respective effects combine (Fig. 2k and Extended Data Fig. 1i–l). When computed for the actual interval duration, rather than its estimate, average RPEs are qualitatively similar to DA responses at interval offset recorded in animals (as shown in Fig. 2l). However, trial-to-trial variability in RPEs is still dominated by effects of choice accuracy (as shown in Fig. 2m,n; for more details, see Extended Data Fig. 2i–l).

In other words, animals' choice accuracy alone, the hazard rate of interval offset alone, and the combination of the two predict distinct patterns of RPEs at interval offset, respectively. Furthermore, none of these three scenarios would seem to simultaneously explain the two key experimentally observed trends: the profile of average DA responses at interval offset and the trial-to-trial relationship between magnitude of DA for any single-interval offset and animals' choices. Average DA responses are best captured by computing reward expectations that take choice accuracy as well as hazard rate of interval offset into account (Fig. 2a is consistent with Fig. 2l). However, the differences in the psychometric functions for high and low DA are captured by computing reward expectations that take into account only the hazard rate of interval offset (Fig. 2b is consistent with Fig. 2j).

Because these simplified considerations on RPEs (and, thereby, DA responses) do not match the data fully, we hypothesized that our assumptions regarding how task variables may be encoded by animals may not be accurate. To gain a more detailed understanding of how differences in the encoding of task variables as well as animals' own behavior may influence RPEs, we simulated a reinforcement learning agent that could make choices at any time during the trial and was required to learn from trial and error, just like animals, to take the right action at every timestep to obtain rewards.

RL agents were modeled to keep track of time since task events. Because animals' choices are based on variable internal estimates of elapsed time, we modeled the RL agent to also have variable estimates of elapsed time (for details, see Methods, section 1). Previous experimental findings have shown that animals exhibit trial-to-trial variability in estimating elapsed time and that the standard deviation of timing estimates increases linearly with time, which is known as the scalar property in timing[38]. Hence, we constructed the agent's internal representation of elapsed time as varying from trial to trial in a manner that obeys this scalar property (for more details, see Methods, section 1, and Extended Data Fig. 3). The amplitude of the noise was adjusted so as to qualitatively match animals' overall task performance (determined using the psychometric function).

We assumed that the agent maintains noisy estimates of both elapsed time since interval onset and time since interval offset. With these two estimates, the agent has all the necessary information to estimate the length of the interval presented as well as the task epoch. The agent was allowed to report choices at any time during the task. Accordingly, it needs to learn to withhold choices during the interval and to report choices, based on its estimate of interval duration, after interval offset. Figure 3 shows how the agent traverses through the state space on two example trial types. In both examples, the agent encodes elapsed time since interval onset by advancing

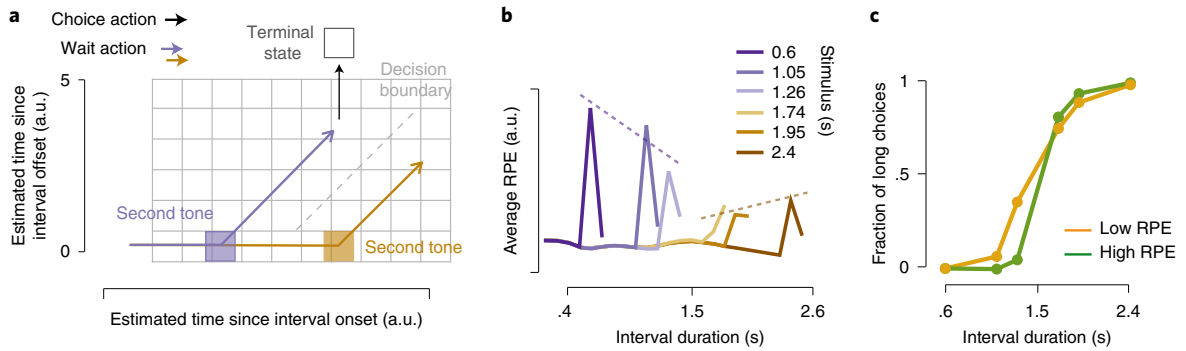

Fig. 4 | Unambiguous value function approximation cannot simultaneously reproduce average DA at interval offset and trial-to-trial relationship between DA magnitude and choice. a, Value function approximation used in the unambiguous representation. Each basis uniquely determines estimated elapsed time since interval onset and interval offset. **b**, Average RPEs elicited at interval offset in an agent that uses the unambiguous representation for interval discrimination. Compare with Fig. 2a. **c**, Psychometric curves of trials grouped based on the magnitude of RPE at each interval offset for an agent using the unambiguous representation. Compare with Fig. 2b. a.u., arbitrary units.

horizontally in the depicted state space. After interval offset, the agent additionally encodes elapsed time since interval offset and, therefore, traverses the state space along diagonals. Because time since interval onset will always be larger than time since interval offset, the agent will only visit states that are below the unity line in this state space. During the interval, the agent should learn to withhold choice, and, after interval offset, the optimal action depends on the x intercept of these diagonals—that is, on the difference between its estimates of elapsed time since interval onset and interval offset.

The agent learns using TD learning within an actor–critic architecture (for more details, see Methods, section 2). Actor–critic architectures have been commonly used to model dopamine activity as RPEs in tasks where outcomes depend on actions[39,40]. In this framework, the agent estimates reward expectations from each state (state–value function) and a state–action mapping (policy), that instructs the agent on which action to take, for all states it encounters. The agent learns to select actions that result in transitions to higher-value states, which then become more probable than transitions to lower-value states. Notably, the state–value function and policy must both be learned simultaneously and are both defined as functions of the location in state space. Because the agent's internal estimates of time are modeled as continuous variables, there are infinitely many locations in state space that the agent could be in. This makes the task of learning value functions for each state directly highly impractical. Hence, we use a function approximation scheme to estimate the value function and policy. Both of these functions are approximated using a set of basis functions or feature vectors (Fig. 3e shows a schematic of the function approximation scheme used). To keep this approximation as simple as possible, we used non-overlapping tile bases. This is equivalent to discretizing the continuous state space for value approximation.

RL agent based only on task requirements cannot reproduce the relationship between DA response and choice. We first modeled an RL agent that estimates the value function and policy by constructing a basis set that uniformly tiles both dimensions of the state space (Figs. 3e and 4a). After the agent learns the task, we found that the profile of average RPEs at each interval offset qualitatively captures that of average DA activity (compare Fig. 4b and Fig. 2a). However, the trial-to-trial relationship between the magnitude of RPEs and the agent's choices is inconsistent with the data (compare Fig. 4c and Fig. 2b). Rather, the psychometric functions of the high-RPE and low-RPE groups of trials show a change in bias as well as slope, as seen when we directly computed reward expectations based on choice accuracy and predictability of interval offset in time (Fig. 2n and Supplementary Fig. 2i,j). At first sight, these results suggest that there might be some aspect of dopamine activity that cannot be captured entirely by RPEs during this task. However, because RPEs are calculated from the agent's expectations of future rewards, given by the state–value function, the results could also suggest that animals are calculating expectations of future rewards in a way that does not match the true underlying structure of the task. Such a mismatch could come about if the representations that the animals are operating on are misrepresenting the statistical structure of the task.

To ask how the underlying state representation could be different, we note that the dynamics of the latent variables in the task are not made available to animals and that they need to infer how the task should be represented using only the sparse observations that they receive. Previous work has shown that population dynamics in the striatum during the interval encode elapsed time with high fidelity and that trial-to-trial variability in how activity evolves is predictive of animals' temporal judgments[17]. Hence, the model encoded time since interval onset with a high resolution. Other work, in the context of an interval reproduction task, has shown that cortical dynamics encode elapsed time since interval onset and offset in a similar manner[19]. Based on these findings, we assumed in the model implementation above that animals would represent elapsed time since interval onset and offset in a similar manner, with high fidelity, during an interval discrimination task as well. However, this assumption may not be well-aligned with how animals may be representing the interval discrimination task. Drawing on the principles of efficient coding, we reasoned that, in addition to maximizing the overall number of rewards obtained, animals may want to minimize the computational resources required to estimate the value function and policy over all possible states. In particular, we hypothesized that animals may encode time-varying value functions with higher resolution during the interval but not after the interval. Hence, we asked whether and how encoding elapsed time since interval onset and offset with different resolutions would affect task performance as well as reward expectations learned during the task.

RL agent with efficient representation can reproduce trial-to-trial relationship between DA response and choices. Let us consider the function approximation used to estimate the value function described in Fig. 4a. The size of each basis function (or tile) will determine how accurately changes in value from one location to another can be estimated. In turn, the number of basis functions

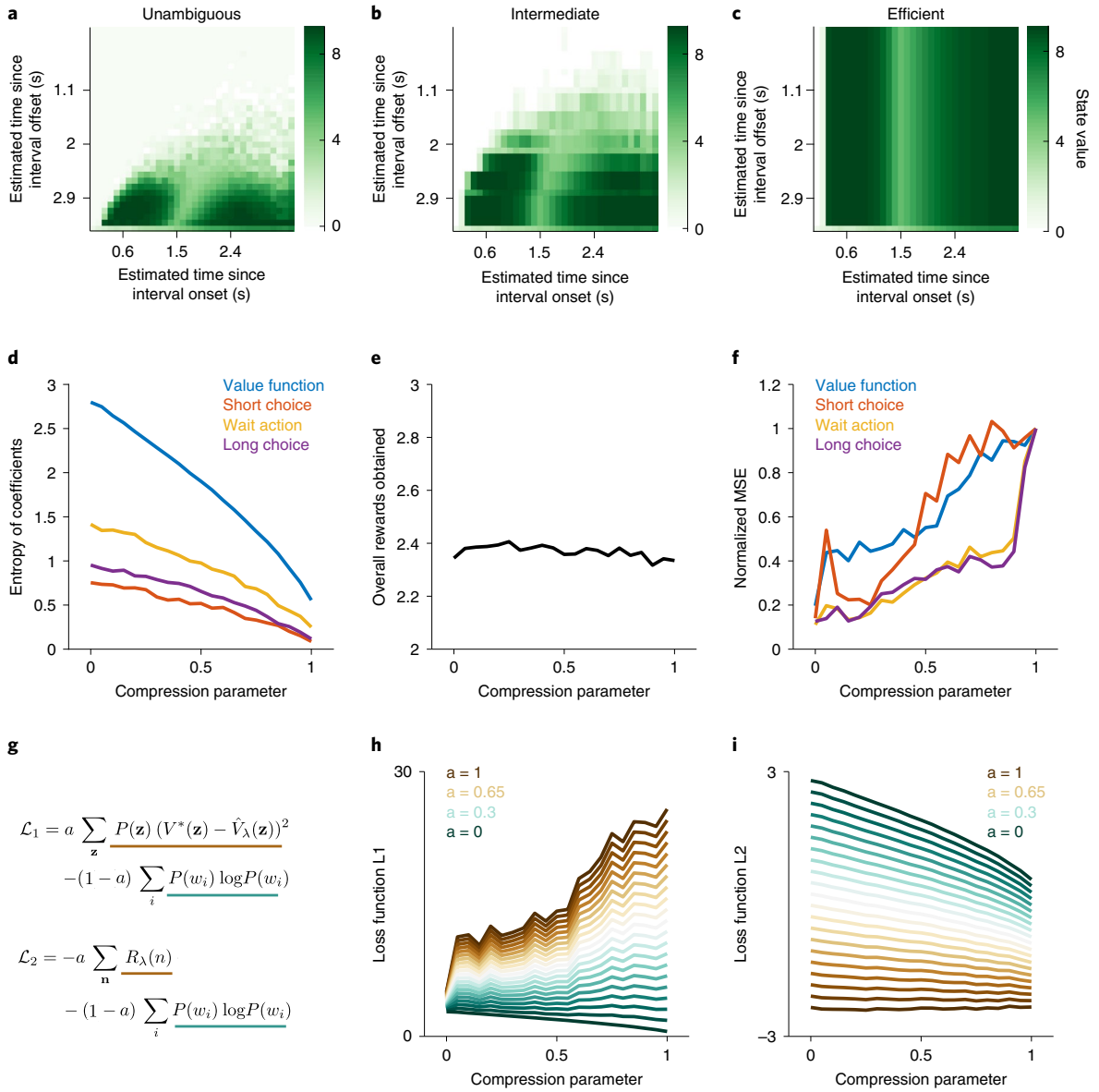

Fig. 5 | Reconstruction accuracy and representational efficiency. a–c, The heat maps in the top row show the approximated state-value function using the unambiguous (**a**), an intermediate (**b**), and the efficient (**c**) representation. All the plots here show results obtained using the following model parameters: $\alpha = .9, \sigma = .25, \epsilon = .1$ (for details, see Methods, section 4). **d**, Shown here are quantifications of the sparsity of the coefficients used to approximate the value and advantage functions when using representations with varying degrees of compression. As expected, the entropy of the coefficients smoothly varies as a function of the compression parameter, is lowest for the most efficient representation ($\lambda = 1$) and is highest for the unambiguous representation ($\lambda = 0$). **e**, Shown here are overall rewards obtained during the task when using representations with different degrees of compression. We see that the overall rewards are nearly unchanged as a function of the compression parameter. **f**, To understand how the reconstruction accuracy of the value function and advantage function changes as a function of the compression, we compute the mean squared error between the optimal value and advantage functions using the unambiguous representation ($\lambda = 0$) and the approximated value and advantage function using representations of varying degrees of compression (for details, see Methods, section 3). We compute the mean squared error for the state-value function (shown in blue) and advantage function for each of the three actions the agent can take (shown in red for short actions, yellow for wait actions and purple for long actions). We see that the reconstruction error steadily increases with the compression parameter and that the reconstruction error is largest when using the most efficient representation ($\lambda = 1$). **g**, Two potential loss functions containing terms for accuracy and efficiency but with different definitions for accuracy. L1 assumes a more traditional reconstruction error type accuracy term, defined in relation to the optimal value function that is learned using the unambiguous task reprepresentation, whereas L2 assumes that accuracy corresponds to total rewards obtained. Accuracy terms are underlined in brown; efficiency terms are underlined in green. **h**, Loss function 1 as a function of compression parameter, for different relative weightings of accuracy and efficiency terms. Green tones indicate greater relative weight given to efficiency; brown tones indicate greater relative weight given to accuracy. **i**, Same as for **h** but for the case of loss function 2. MSE, mean squared error.

(or tiles) will determine the computational cost of estimating the entire value function. Accordingly, more accuracy incurs more costs. An efficient coding scheme might stipulate that the resolution with which an optimal set of basis functions tile the two dimensions of the state space should depend on the degree to which the read-out varies along each of these dimensions[27]. More specifically, value

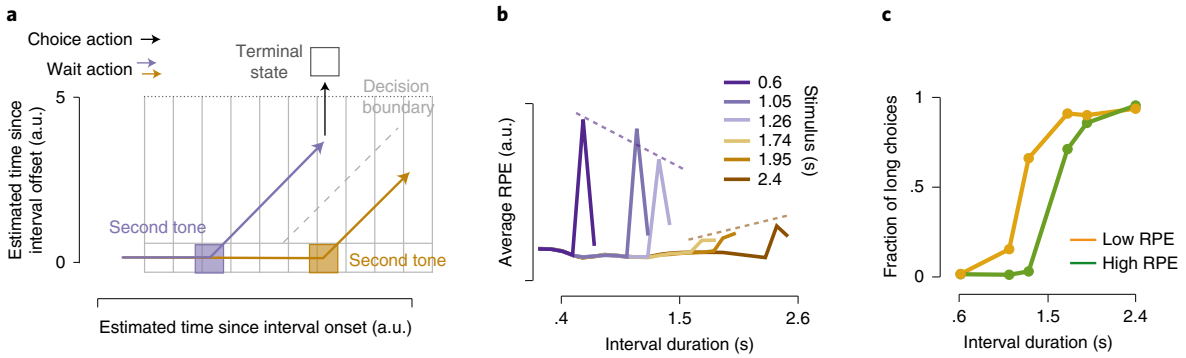

Fig. 6 | Efficient value function approximation can simultaneously reproduce average DA at interval offset and trial-to-trial relationship between DA magnitude and choice. a, Value function approximation used in the efficient representation. **b**, Average RPEs elicited at interval offset in an agent that uses the efficient representation for interval discrimination. Compare with Fig. 2a. **c**, Psychometric curves of trials grouped based on the magnitude of RPE at each interval offset. Compare with Fig. 2b. a.u., arbitrary units.

functions along the axis that represents elapsed time since interval onset may be encoded with high resolution, because doing so is necessary for the agent to accurately report choices. Value functions along the axis that represents elapsed time since interval offset, on the other hand, may be encoded with a lower resolution, because lack of accuracy here may not adversely affect the animal's ability to make the correct choice (Fig. 5a–c). In the extreme case, when the axis representing time since interval offset is encoded with the lowest possible resolution (while still encoding interval offset), the basis functions effectively encode only time since interval onset and whether or not interval offset has occurred (Fig. 6a). We will refer to the basis functions as described in Fig. 4a as the high-resolution or unambiguous representation and the other extreme shown in Fig. 6a as the low-resolution or efficient representation (efficiency quantification shown in Fig. 5f).

When we trained the RL agent using the efficient representation, we found that it is able to obtain similar overall rewards as an agent trained using the unambiguous representation (Fig. 5e). The efficient model also reproduces the profile of average DA responses at interval offset (compare Fig. 6b and Fig. 2a). In other words, the compression of the representation along the second axis did not adversely affect the agent's choice behavior, nor did it change the predictions for average RPEs at interval offset, even though reconstruction accuracy of the value function and policy gets worse as the degree of compression increases (Fig. 5d). Surprisingly, using the efficient representation, the model is also able to reproduce the trial-to-trial relationship between magnitude of DA and temporal judgments (compare Fig. 6c and Fig. 2b). This somewhat puzzling result was obtained only for very strong compressions of representations along the second axis and not for intermediate levels of compression. When we simulated the agent using several intermediate levels of compression in representing elapsed time since interval offset, we obtained results more similar to those using the unambiguous representation (Extended Data Fig. 4). Accordingly, only a large difference in the resolution with which elapsed time since interval onset and offset are encoded can explain the observed DA responses and their relation to behavior. These results were consistent for a wide range of other parameters with which the RL agent was simulated (for details, see Extended Data Figs. 5 and 6).

Although the current work does not focus on explicitly learning the task representation under a particular objective or loss function, we next wondered whether the ability of the efficient representation to capture a range of neural data in relation to behavior could, nonetheless, provide information about the characteristics of the objective used by the brain. Broadly speaking, efficient coding posits that neural systems are optimized to best encode meaningful information while minimizing cost. In **g**–**i** of Fig. 5, we formulate two example loss functions, one that combines reconstruction accuracy of the optimal value function with representational efficiency and one that combines overall rewards obtained with representational efficiency. We vary the relative weights of the two terms and plot the loss as a function of the compression parameter in the model to show how these losses vary from the efficient to the unambiguous representations. We see for all weightings, other than those that nearly ignore reconstruction error of the value function, that the efficient model (compression parameter = 1) does not optimize a loss that trades reconstruction error against efficiency of representation. However, for a loss that uses overall rewards obtained, the efficient representation does minimize the loss function, even if efficiency is given a relatively low weight. Accordingly, the RL agent that simultaneously captures the various trends in the data is consistent with a strategy of learning representations that preserves overall rewards obtained while penalizing representational cost.

Reward expectations at interval offset are markedly different when using representations of varying efficiency. To understand why our proposed efficient representation results in very different RPEs, we looked at the value function learned by the agents using the unambiguous versus efficient representations (as shown in the schematics in Figs. 4a and 6a, respectively). We note that RPEs at interval offset reflect the difference between the agent's reward expectation (given by the corresponding value function) during the interval and its reward expectation at interval offset, at which point the agent does have an estimate of the interval duration to be classified on that trial. For agents using the unambiguous representation, reward expectations before interval offset reflect the hazard rate of interval offset and choice accuracy, which increase as a function of elapsed time (Fig. 7a). After interval offset, the agent has acquired an estimate of the presented interval, and, therefore, its reward expectations reflect only choice accuracy (Fig. 7b). At any time, the difference between these two value functions determines the RPE if interval offset was presented at that time (Fig. 7c). Consequently, RPEs at interval offset reflect both the hazard rate of interval offset and the agent's choice accuracy.

Similarly, we can look at the value function learned using the efficient representation. As before, we see that reward expectations during the interval increase with the length of the interval (Fig. 7d). However, reward expectations after interval offset do not simply reflect the agent's choice accuracy. Instead, they exhibit a strong asymmetry around the decision boundary (Fig. 7e). We see that, on the long side of the boundary, reward expectations increase much more slowly as a function of distance from the boundary than on

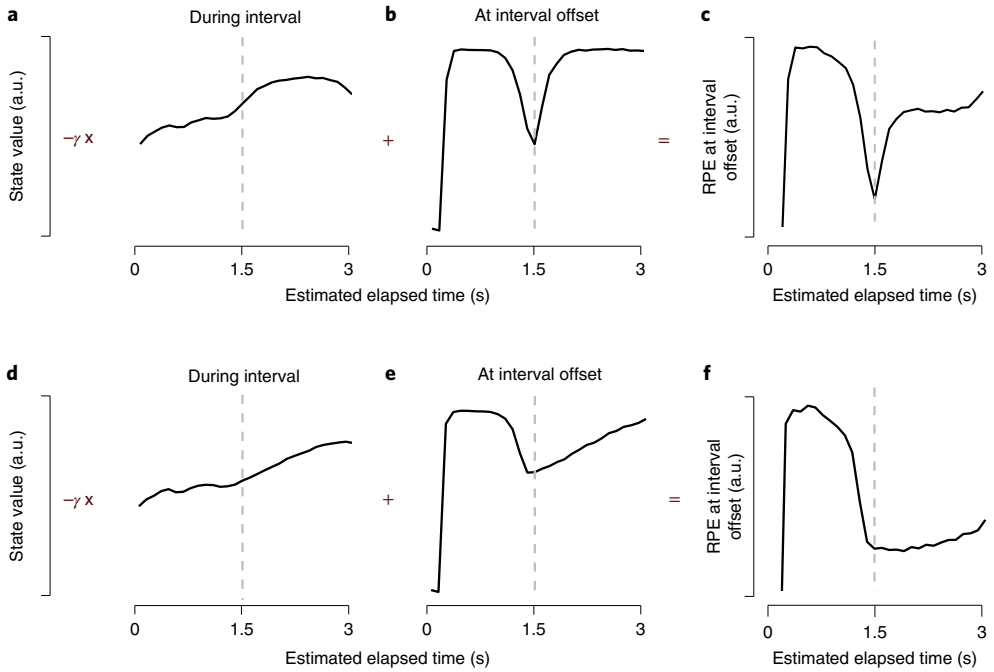

Fig. 7 | Reward expectations learned by the RL agent using the unambiguous and efficient representations differ most right after the decision boundary. Given reward expectations or the state–value function during the interval and that at interval offset, the RPE curve over all estimates of interval durations can be computed by taking the difference between value at interval offset and the discounted value (γV, where $\gamma = .95$ is the discount parameter) at the preceding timestep during the interval: $RPE(t) =$ (value at interval offset at t) $- \gamma$ (value during interval at $t-1$). Value function during the interval (**a**) and after interval offset (**b**) as a function of internal estimates of elapsed time since interval onset using the unambiguous representation for value estimation. **c**, Reward prediction error at interval offset as a function of all estimates of elapsed time since the interval onset using the unambiguous representation. Reward expectation during the interval (**d**) and after interval offset (**e**) as a function of internal estimates of elapsed time since interval onset using the efficient representation for value estimation. **f**, Reward prediction error at interval offset as a function of all estimates of elapsed time since the interval onset using the efficient representation. a.u., arbitrary units.

the short side of the boundary. Consequently, the resulting RPEs reflect this asymmetry and show a slower rise on the long side of the decision boundary compared to the short side (Fig. 7f). We found that this asymmetry substantially changes the trial-to-trial relationship between the animals' choice behavior and dopamine activity and allows the model to reproduce the change in bias observed in the psychometric curves of low and high DA trials in animals (for a detailed description of this relationship, see Extended Data Fig. 7).

Efficient representation predicts procrastination of choices for interval durations estimated as long and close to the decision boundary. We studied the origin of the asymmetry in the value function for the agent using the efficient representation. We found that the low resolution along the second axis leads to a systematic ambiguity in some parts of the state space in estimating value and the optimal action for those locations. For example, let us consider a basis function that spans a region in state space that the agent would visit right after the end of a long interval (marked with the gray rectangle in Fig. 8a). Although the agent can encounter this region directly after the end of a long interval, it could also be encountered if the agent was presented with a short interval but withheld choice for several timesteps (purple trajectory in Fig. 8a). Hence, the reward expectation associated with this basis will be estimated by averaging over the trials from both categories of interval durations—that is, when the agent's estimate of the interval presented is longer or shorter than the learned boundary. As a consequence, the agent would be impeded in learning the correct value of these states as well as the optimal action at these locations. Indeed, this is true for all the post-interval basis functions in the efficient representation. They encode elapsed time since interval onset and whether or not interval offset had occurred. Accordingly, the basis functions do not allow the agent to disambiguate between trials on which different interval durations were presented if it withholds choice for several timesteps.

The ambiguity in this efficient representation can be avoided if the agent reports choices immediately after interval offset, particularly for intervals estimated to be shorter than the decision boundary. Once the choice is made, the agent transitions into the inter-trial interval state and, thereby, avoids visiting other post-interval states. Indeed, we found that when agents use the efficient representation and when they estimate the interval to be shorter than the decision boundary, they report choices with very short response times (Fig. 8b). For intervals longer than the decision boundary, however, there is no urgency to respond after interval offset. In this case, if the agent waits after the interval offset, the correct action associated with the post-interval offset states will not change with the passage of time. Hence, delaying choice will not have any detrimental effect on choice behavior. Curiously, however, the efficient representation not only allows delaying near-boundary long choices but incentivizes it (for detailed discussion of why this is the case, see Extended Data Fig. 8).

In several two-alternative decision-making paradigms, animals generally take longer to respond when the decision variable is close to the decision boundary[41]. Several tasks in which response times are longer for harder stimuli require integration of noisy evidence, where the noise is uncorrelated over time. In these contexts, longer response times allow animals to integrate over more samples of noisy evidence and have a better estimate of the stimulus by averaging out the noise. In other words, longer response times result from increased deliberation when the stimulus category is more

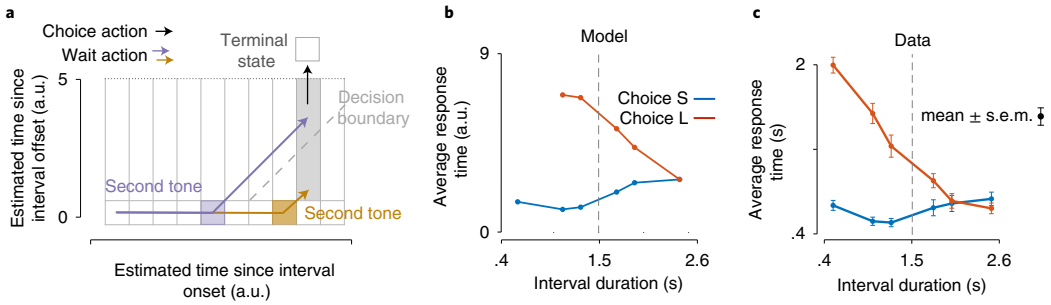

Fig. 8 | The efficient representation predicts an unusual profile of response times that closely matches animals' behavior. a, Efficient representation and two example trajectories through the state space. The gray rectangle indicates a post-interval offset basis. Note that this region of the state space can be reached after long intervals (yellow) as well as short intervals (purple) if the agent waits and withholds choice. The basis function is, therefore, ambiguous in informing the agent which interval duration was presented when the corresponding states are encountered on any given trial. **b**, Average response times of an RL agent that uses the efficient representation, conditioned on whether they indicated the interval to be short or long with respect to the decision boundary. **c**, Animals' average response times after each interval offset split based on choice. Data are presented as mean values ± s.e.m. pooled over all animals ($n = 6$ animals). a.u., arbitrary units.

ambiguous. However, the response times of the RL agent using the efficient representation in our task are markedly different, as they predict long responses only for difficult 'long' choices but not for difficult 'short' choices. In other words, the RL agent generates a highly non-trivial prediction that can be tested against data. The long response times we see for intervals that are perceived by the agent to be near-boundary 'long' appear to be a result of procrastination of difficult 'long' choices. Surprisingly, we found that animals also procrastinated as predicted by the model. The predicted pattern of response times from the model closely resembles that of animals during the task (Fig. 8c). We also found that this pattern of response times was not reproduced by the agent when using the unambiguous representation. Moreover, the profile of response times was observed only for highly compressed representations (Extended Data Fig. 9) and not for intermediate levels of compression, such as the one shown in Fig. 5b. These results are consistent over a wide range of model parameters with which the agent is simulated (Extended Data Fig. 10).

Furthermore, if we force the agent to not have variable response times, the profile of psychometric curves of trials grouped by magnitude of RPEs at interval offset in the agent matches those resulting from the unambiguous representation and do not match those in the data (Supplementary Fig. 1a,g,j). Thus, animals' pattern of response times provides further evidence to suggest that animals may, indeed, be using a representation similar to that captured by the efficient representation while solving this task.

Discussion

Understanding the principles that guide RL in value-based decision-making is crucial to further understanding of how neural circuits subserve adaptive behavioral control. Characterizing reward expectations that animals learn can inform us what variables they are representing, which, in turn, can reveal the constraints and strategies with which animals infer statistical regularities in their environments. Here, we studied the computations underlying a rigorously controlled time-dependent behavior in mice. We focused on two aspects of experimental data collected during this task: the recorded activity of dopamine neurons and their trial-to-trial relation to animals' choices. Using an RL framework, we investigated how varying internal representations, with which the agent was able to solve the task, changed RPEs at task cues. By comparing such RPEs with recorded DA responses, we were able to infer the nature of internal representations that animals might be using during this task. Our results are consistent with animals encoding a representationally efficient internal representation that does not accurately capture the statistics of the task or the optimal value function and policy. However, the efficient representation allows the agent to minimize representational costs while still learning the correct actions and collecting equivalent amounts of rewards. It also predicted a very specific behavioral strategy that matches animals' behavior during the task. Notably, we found that the behavioral strategy of the efficient RL agent is crucial to prevent ambiguities in the representation from affecting the accuracy of choices and, hence, the overall number of rewards that can be obtained. Our results also show why learning representations that directly maximize overall rewards may lead to unexpected interactions between how the environment is represented and what policy is learned.

In several sensory systems, the principle of representational efficiency has been used to characterize neural coding. In many of these systems, the variables that are to be represented, and, hence, subject to efficiency constraints, are usually well-defined. However, in the case of RL, it is unclear which variables used for value-based choices might be subject to constraints of representational efficiency in the brain[28]. For example, we may want to enforce efficiency constraints in how the structure of the environment is represented[14,42]. This approach may be considered to be the closest to that used for efficient coding in sensory systems. In this case, the problem can be stated as that of identifying statistical regularities or redundancies in the environment that can be leveraged to balance representational cost against the accuracy with which the statistics of the environment can be encoded. However, this is not the only approach that can be used for representational efficiency in RL. Representation constraints may be used to directly approximate optimal value functions[43] or action spaces[44]. In each of these cases, one has to define a space within which representations are subject to resource constraints as well as the quantity of interest that needs to be preserved—that is, the loss function that needs to be optimized. A key challenge in understanding how principles of efficient coding are applicable to the reward system is to identify the space within which representational constraints may be expressed as well as the loss function that may be optimized.

The interval discrimination task requires animals to generate and operate upon structured, time-evolving internal estimates to guide decisions. Hence, the representations that need to be learned in a task such as interval discrimination must be dynamic or time-varying in the absence of time-varying inputs or changes in the agent's environment. The problem of learning dynamics that maximize rewards under representational cost is, indeed, a more challenging learning problem than learning efficient static representations. Several lines of research may converge to this goal. A growing body of work asks

how dynamics in recurrent networks may be learned under different types of constraints, such as rank and energy constraints, and the consequences of these constraints on the computations performed by those networks[45,46]. In parallel, end-to-end RL using deep and recurrent neural networks that are optimized to maximize rewards directly have proven powerful in training agents on a range of tasks. The recent success of deep networks trained end to end with TD learning at playing various games has provided a demonstration of how effective this form of learning can be[47]. Moreover, previous work has shown that dopamine neurons project widely to a large number of neural circuits in the brain, and end-to-end learning has been shown to reproduce neural activity in prefrontal cortex in a wide range of tasks[48,49]. This success underscores the importance of understanding how representations learned directly to maximize overall rewards (or minimize reward prediction errors) may be different than those obtained by maximizing other quantities, such as reconstruction error. By combining the understanding of how time-varying representations can be learned to maximize rewards with various constraints on the parameters that describe dynamics, future work may allow for understanding how these representations may be learned from first principles. In the context of a rigorously controlled task, our work shows how representations that are inaccurate in encoding several aspects of the task, but allow the agent to preserve overall rewards obtained while being representationally efficient, can lead to behavior and reward expectations that are qualitatively different than those that would result from using representations that may best summarize the statistics of the environment or features of the optimal value function and policy.

In sum, by investigating behavior and DA activity during a time-based decision-making task using RL, we were able to reveal an efficient strategy that animals appear to be using to represent task variables. We demonstrate that constraints of representational efficiency affect the nature of reward expectations learned during this task and that the activity of DA neurons could be explained by the model using only this efficient representation. Finally, we show how animals' behavioral strategy interacts with the representation used to encode the task in an unexpected way and that this interaction was central for the RL agent to be able to reproduce animals' behavior and DA activity. These findings provide novel insights into the manner in which efficiency constraints might be expressed in the reward system and, more generally, provide insights into the principles underlying natural, intelligent behavior.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01085-7.

Received: 10 June 2020; Accepted: 26 April 2022;
Published online: 6 June 2022

References

1. Dayan, P. & Sejnowski, T. J. Td(λ) converges with probability 1. *Mach. Learn.* **14**, 295–301 (1994).
2. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
3. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
4. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
5. Schultz, W., Dayan, P. & Montague, P. R.A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
6. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
7. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Curr. Biol.* **24**, 2491–2500 (2014).
8. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
9. Niv, Y. & Langdon, A. Reinforcement learning with Marr. *Curr. Opin. Behav. Sci.* **11**, 67–73 (2016).
10. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).
11. Daw, N. D., Courville, A. C. & Touretzky, D. S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006).
12. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* **20**, 3034–3054 (2008).
13. Suri, R. E. & Schultz, W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* **91**, 871–890 (1999).
14. Botvinick, M. M., Niv, Y. & Barto, A. G. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
15. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).
16. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768 (2017).
17. Gouvêa, T. S. et al. Striatal dynamics explain duration judgments. *eLife* **4**, e11386 (2015).
18. Mello, G. B. M., Soares, S. & Paton, J. J. A scalable population code for time in the striatum. *Curr. Biol.* **25**, 1113–1122 (2015).
19. Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron* **98**, 1005–1019 (2018).
20. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* **21**, 102–110 (2018).
21. Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).
22. Lewicki, M. S. Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
23. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
24. Rieke, F., Bodnar, D. A., & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. Biol. Sci.* **262**, 259–265 (1995).
25. Machens, C. K., Gollisch, T., Kolesnikova, O. & Herz, A. V. M. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron* **47**, 447–456 (2005).
26. Reinagel, P. & Zador, A. M. Natural scene statistics at the centre of gaze. *Network* **10**, 341–350 (1999).
27. Salinas, E. How behavioral constraints may determine optimal sensory representations. *PLoS Biol.* **4**, e387 (2006).
28. Botvinick, M., Weinstein, A., Solway, A. & Barto, A. Reinforcement learning, efficient coding, and the statistics of natural tasks. *Curr. Opin. Behav. Sci.* **5**, 71–77 (2015).
29. Summerfield, C. & Tsetsos, K. Building bridges between perceptual and economic decision-making: neural and computational mechanisms. *Front. Neurosci.* **6**, 70 (2012).
30. Louie, K. & Glimcher, P. W. Efficient coding and the neural representation of value. *Ann. N Y Acad. Sci.* **1251**, 13–32 (2012).
31. Soares, S., Atallah, B. V. & Paton, J. J. Midbrain dopamine neurons control judgment of time. *Science* **354**, 1273–1277 (2016).
32. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
33. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
34. Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr. Biol.* **27**, 821–832 (2017).
35. Janssen, P. & Shadlen, M. N. A representation of the hazard rate of elapsed time in macaque area lip. *Nat. Neurosci.* **8**, 234–241 (2005).
36. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973 (2008).
37. Pasquereau, B. & Turner, R. S. Dopamine neurons encode errors in predicting movement trigger occurrence. *J. Neurophysiol.* **113**, 1110–1123 (2015).
38. Gibbon, J. & Church, R. M. Representation of time. *Cognition* **37**, 23–54 (1990).
39. Joel, D., Niv, Y. & Ruppin, E. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* **15**, 535–547 (2002).

40. Khamassi, M., Lachèze, L., Girard, B., Berthoz, A. & Guillot, A. Actor–critic models of reinforcement learning in the basal ganglia: from natural to artificial rats. *Adaptive Behavior* **13**, 131–148 (2005).
41. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
42. Wimmer, G. E., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *Eur. J. Neurosci.* **35**, 1092–1104 (2012).
43. Foster, D. & Dayan, P. Structure in the space of value functions. *Mach. Learn.* **49**, 325–346 (2002).
44. Solway, A. et al. Optimal behavioral hierarchy. *PLoS Comput. Biol.* **10**, e1003779 (2014).
45. Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623 (2018).
46. Kao, T.-C., Sadabadi, M. S. & Hennequin, G. Optimal anticipatory control as a theory of motor preparation: a thalamo-cortical circuit model. *Neuron* **109**, 1567–1581 (2021).
47. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
48. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* **6**, e21492 (2017).
49. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).