Behavioural Processes 167 (2019) 103891

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Review Article

Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making

Angela J. Langdon*, Mingyu Song, Yael Niv

Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, 08544, United States

ARTICLE INFO

Keywords:
Learning
Decision making
Reward
Timing
Dopamine

ABSTRACT

We review the abstract concept of a 'state' – an internal representation posited by reinforcement learning theories to be used by an agent, whether animal, human or artificial, to summarize the features of the external and internal environment that are relevant for future behavior on a particular task. Armed with this summary representation, an agent can make decisions and perform actions to interact effectively with the world. Here, we review recent findings from the neurobiological and behavioral literature to ask: 'what is a state?' with respect to the internal representations that organize learning and decision making across a range of tasks. We find that state representations include information beyond a straightforward summary of the immediate cues in the environment, providing timing or contextual information from the recent or more distant past, which allows these additional factors to influence decision making and other goal-directed behaviors in complex and perhaps unexpected ways.

Many computational theories of learning and decision-making rely on the concept of a 'state' – a representation used by an animal, human or artificial agent that describes the current situation of the agent within an environment and which the agent uses to guide their behavior. In reinforcement learning (RL), where a typical task is to choose which of multiple possible actions to perform in order to obtain a possible reward, the current state selectively includes all current and past environmental information an agent treats as relevant for making their decisions to act (Sutton and Barto, 1998). A state representation thus encapsulates knowledge about the structure of a task, providing a map of discrete states that follow one another given events that occur during the task and the actions performed by the agent. While a state is a central concept for RL algorithms, it is an idea that can be ambiguous when applied to neurobiology and behavior. Here, we aim to orient readers unfamiliar with RL to key concepts in the definition of state, and discuss the assumptions (explicit and implicit) that arise when attempting to determine the representation of state internal to an agent acting within a task environment. Through this detailed discussion of the formal concept of 'state' we thus trace some of the features of state representations that underlie the complex behaviors of animals and humans in various learning and decision making tasks.

1. Reinforcement learning: Algorithms for action towards a goal

RL provides a diverse set of algorithms all designed to solve the problem of learning to obtain rewards (and avoid punishments) by taking actions that influence future events in an environment (Sutton and Barto, 1998). Distinct from supervised learning techniques, which learn from labeled examples of the target behavior, RL algorithms are designed to learn solely from trial-and-error experience of the outcomes that are provided by the environment. This makes RL an appealing framework for understanding the computational processes that support learning and decision making in animals and humans across many tasks; in both laboratory and natural environments, the task for a real *agent*, is to learn what actions to take in each circumstance in order to maximize reward. Reward in an RL setting can be flexibly defined, and does not encompass solely food or other consumable goods, but may track any signal provided by the environment that aligns with the goal of the agent, including money, warmth, or reaching a safe enclosure.

RL algorithms are built around the concept of a 'state', which provides a summary of the current situation of the agent within the environment. In the simplest interpretation, and in many classic RL models, the assumed state representation for a given task reflects the explicit configuration of the environment, essentially a representation of whatever unique cues or outcomes are known or designed to be predictive of obtaining reward (Montague et al., 1996; Schultz et al.,

* Corresponding author.
E-mail addresses: alangdon@princeton.edu (A.J. Langdon), yael@princeton.edu (Y. Niv).

https://doi.org/10.1016/j.beproc.2019.103891
Received 30 January 2019; Received in revised form 23 May 2019; Accepted 21 June 2019
Available online 02 August 2019
0376-6357/

1997; Suri and Schultz, 1999). However, in many tasks, there are features of the environment that are not readily or continuously observable by a learning agent, yet they are critical for guiding appropriate behavior (Kaelbling et al., 1998; Wilson and Niv, 2012). For example, during a task in which a brief cue at the start of a trial indicates whether a later action (e.g., a lever press) will be rewarded or not, the contextual information provided by the cue should be included as part of the state at the time of lever presentation even if the cue is no longer observable at that time. Further, there may be internal variables – satiety, fatigue – that are also relevant for describing the current situation of the agent and may factor into the state representation of a task (Berridge, 2004). Conversely, in more complex tasks as well as in many natural environments, there are numerous features that may be salient, yet are irrelevant for task performance (Niv et al., 2015). For the sake of efficiency, such irrelevant features should not be part of a state representation designed for learning and decision making in such a task. The state representation is therefore not trivial, and understanding learning across diverse settings will be incomplete without understanding the processes by which state representations are formed (Wilson et al., 2014).

Establishing what state representation is used by an animal to guide behavior during learning requires understanding what aspects of the environment, both external and internal, the animal is treating as relevant to achieving their immediate goal. What are the characteristics or limitations, if any, of the state representations used by real agents acting in real environments? And how do real agents acquire a state representation for a task they have never experienced before? In the following section, we review the theoretical basis for the concept of a 'state' in reinforcement learning theory, and use this theory to trace the features of hidden state representations that support learning and decision making across a range of tasks. Drawing on both neurobiological and behavioral evidence, we then uncover features of state representations that support reward prediction and prediction-error signaling, and survey the implications of these state representations for theories of learning and decision making.

2. Defining a 'state': states as summaries over relevant history

Reinforcement-learning problems are formulated within the general framework of *Markov decision processes* (MDPs; Bellman, 2013; Puterman, 2014). MDPs frame the problem of interacting with an environment in order to achieve a goal, where the goal in RL is typically to maximize the accumulated reward. Within the setting of a task, the environment provides cues of various types to the RL agent: these are the *observations*. In addition, the environment may yield an *outcome*, such as a reward. In turn, the agent can influence the environment by executing *actions*, in the hope of controlling the subsequent environmental outcomes in line with its reward-maximizing goal. The full sequence of observations, outcomes and actions for each moment in time up until the present is referred to as the *history*, as it provides a complete description of the observable events that have occurred during the task (Fig. 1).

In this setting, the *environment state* specifies the current configuration of the task: it summarizes the immediate observation and outcome, and specifies the mapping from the current state to its successor state, perhaps dependent on variables that are not observable to the agent, as well as the agent's action. Formally, this mapping is captured by a *transition matrix* – the probability of transitioning to each possible state given an initial state and action. Note that action here is very broadly defined: waiting, or a decision not to move, may also be considered an action. Moreover, state transitions need not display any action dependence in that several possible actions might all result in a transition to the same state. For instance, the probability of a cloudy sky transitioning to rain is independent of any action an agent may take. In general, both state transitions and outcomes are *stochastic* (that is, probabilistic), such that the same action, executed in separate visits to

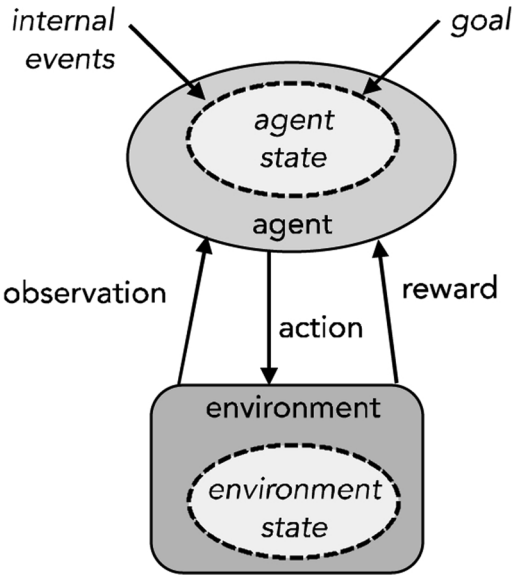

Fig. 1. The agent-environment interaction in RL. Reinforcement learning problems are defined given an agent, or learner, interacting with an environment. At each moment, the agent obtains observations from the environment, and decides which action it wishes to perform in order to influence the future state of the environment, potentially eliciting a reward. The generative state of the environment governs the possible observations, rewards and future states given the actions of the agent, and may comprise of multiple variables that govern the structure of the task. The agent has access to observations, but not necessarily to the true underlying generative state of the environment, as not all variables governing the structure of the task must be observable. Thus, the agent must infer a state representation – the agent state – that comprises all environment features it considers relevant for reaching their current goal.

the same state may result in a different subsequent state and/or outcome. Further, the environment state may depend on hidden variables that do not directly produce an observation that is available to the agent. For example, the probability of a cloudy sky transitioning to snow depends on a particular environmental state of temperature, pressure and humidity that may not be sensed directly by the agent, even though the resulting precipitation is readily observed.

Obviously, an agent can only base its decisions on an *internal* estimation of the current environmental state – the *agent state* – which it must infer without knowledge of the true generative properties of the environment. For the agent, constructing a Markov state representation of a task thus forms the bedrock for learning to act within a given environment. Given an internal representation of the state, the RL framework suggests various algorithms that learn the *value* of each state, defined as the cumulative future reward an agent can expect to receive when starting from that state and acting according to a given *policy* (i.e. rule for deciding between candidate actions). These algorithms comprise of learning rules for iterative updating of said state values through experience, so that reward outcomes that follow from a particular state accrue in the value of that state. RL algorithms also allow an agent to directly learn an action-selection policy for each state. A critical prerequirement for using RL is therefore for the agent to represent internally their current state. This state is what the agent will use to bind together previous experiences of the same situation, yielding reward predictions (values) that guide decisions about future action (policies).

Importantly, in an MDP, the probability of the environment transitioning to any successor state depends only on the current environment state and chosen action, and not on past states. For instance, when driving to work, the probability of getting to work on time given a certain action (e.g., going right or left at an intersection) depends on your current location (i.e., the current state), and not on how you arrived there. The conditional independence of transition probabilities at

the current state from all previous states and actions is known as the Markov property. For the Markov property to hold, the current state must therefore encompass all the information determining the probability of the next state transition. For instance, in all likelihood the probability of getting to work on time will also depend on the current time, and whether there are road closures – as these aspects are part of the environment state, they should also be included in your internal agent state as you decide what route to take. This so-called 'memoryless' property of Markov state representations, in which the current state is sufficient to determine the probability of future events, is central to RL algorithms, as it allows local learning that does not depend on history.

Formally, states are defined as equivalence classes over history (Minsky, 1967). As previously mentioned, history includes the entire set of environmental events, including outcomes, as well as all previous actions undertaken by the learning agent up until the current moment. History is not restricted to a single task; in theory, the history of the agent encompasses all experience and actions. Two or more histories form an equivalence class if, from the current moment forward, the agent responds to the subsequent sequence of events with exactly the same actions. To make this definition more concrete, consider the example of a GO/NO-GO task in which a cue at the start of a trial indicates whether a lever press will be rewarded or not (Fig. 2). If the cue is green, pressing the lever will be rewarded. If the cue is red, not engaging with the lever will lead to reward. Ideally, the internal state of the animal when contemplating the next action will separate the presentation of the lever into distinct states according to the past experience of observing a green or red cue, even though the cue is no longer present. That is, presentation of the lever for action should be treated as two different states based on the history of the environment, as the agent should respond to these two sequences of events with different actions. Conversely, all trials with a green cue should be classified to the same state regardless of the animal's location at the time of lever presentation (and therefore the retinal input of the lever), the elapsed time between the cue and the insertion of the lever, or the cues that preceded it. Between these extremes however, there are intermediate cases: should the fact that the previous trial was not rewarded (due to choosing an incorrect action) be part of the current state? And even if the true generative properties of the environment suggest that it shouldn't, is it?

The agent state and the environment state may not align, in that they may contain distinct subsets of information, though in many simple tasks it is assumed that they are equivalent. An agent state that accurately tracks the true environment state will naturally provide an accurate basis for learning. However, any approximation that obeys the Markov property will allow learning using RL algorithms. How closely the agent state hews to the true environment state limits how close an agent can approach optimal behavior in a task. For example, in the GO/NO-GO task of Fig. 2, an agent that correctly represents the occurrence of a light prior to the insertion of the lever, but fails to distinguish

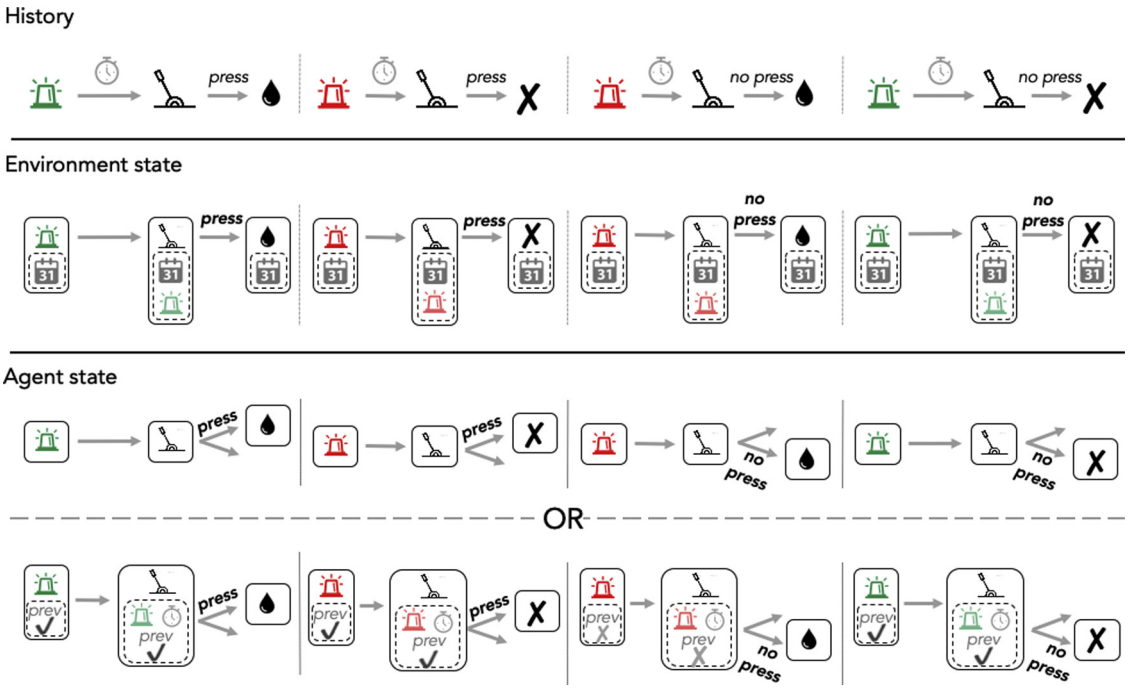

Fig. 2. Internal state representations constructed by an agent to solve a learning task. Top: A simple task in which the color of a light cue indicates whether an agent should engage with a lever when it is presented after a variable delay. A green light indicates engaging with the lever will produce a reward and not engaging will lead to no reward delivery, whereas a red cue indicates reward will be delivered only for not engaging with the lever and not delivered otherwise. The history of experience consists of all observable events: the entire set of cues, actions and outcomes (here, reward or no reward) for every timepoint throughout the task. Middle: the environment state is the true underlying state of the task, and contains all elements necessary for generating the events during the task (i.e. the 'rules'). As such, the environment state contains all features that are relevant for obtaining reward, in this case the current event in the trial, and the color of the cue prior to the lever. The environment state may also contain various features that are not relevant to the immediate goal of the agent, such as the day of the week (which perhaps determines whether rewards are given at all, or the test is in extinction). Bottom: The agent state can include different features and still provide a basis for learning. For example, the agent may represent the states of the task by tracking only the immediately observable event in the environment (top row). In this case, they represent a single 'lever' state on all trials, and need to learn a policy to guide action selection when in this one state. This state representation is adequate for a win-stay, lose-shift policy as in this sequence. Of course, given the true structure of the task, this policy may not be optimal in attaining maximal reward. Alternatively, the agent may include many distinct features, including features from past experience in their internal representation of the state (bottom row). Here, the lever presentation is represented by two distinct states (as is appropriate in this environment), separated according to whether they were preceded by a green or red light. However, there are features of the true state of the environment that the agent does not represent (i.e., day of the week). In addition, in the illustration, the agent includes task-irrelevant information in their representation of the current state of the task: unbeknownst to the agent, the generative properties of the environment are such that neither the delay between cue and lever presentation, nor the outcome of the previous trial, influence the likelihood of obtaining reward on each trial.

between red and green, will think the environment is more stochastic than it actually is. Conversely, an agent that represents many more features of the environment than are actually predictive of reward, such as the precise temperature in the room and the weather outside at the time of the lever, may learn accurately but much more slowly. Thus, grouping experience into a state representation useful for learning in a particular task requires the agent to be sensitive only to information from the environment that is relevant for their current goal. For example, if on each trial a 10 s cue is followed immediately by a reward, the onset of the cue should be recognized by a reward-seeking agent as a transition to a single 'reward is forthcoming' state. This is despite the fact that no two repetitions of the cue are truly identical; at the very least, each experience differs in the history of experience leading to that event (not to mention other perceptual differences). If each cue were treated as a distinct state, true to the full dependence of each occurrence on its history, learning about past states could not be used to guide future decisions as states would never be revisited. A critical step for tracking the current state is thus determining how similar two experiences need to be in order to be classified as the same state, which basically determines the extent of generalization of learning from one situation to another.

3. Inferring states in partially observable environments

Closely related to consideration of what comprises the state for a learning agent is the

concept of internal models in *partially observable* environments, in which the true generative state of the environment is hidden from the agent and must be inferred from observable events that occur during a task (Kaelbling et al., 1998; Rao, 2010). That is, in this case, a hidden state comprises of a set of features, some observable and some not, that determine the transitions between states given the agent's actions. An *observation model* describes the probabilistic structure or rules that govern the emission of observations given the true hidden state. The observation model also depends on the capabilities of the agent: while a cue may indicate a reward is imminent in a given environment state, this cue becomes an unobservable feature if the agent is blind. In such a setting, inferring the current hidden state of the environment involves parsing a sequence of observations (i.e. *history*) according to an internal model of how those observations are generated by the underlying sequence of hidden states.

Environments in which delays separate predictive cues from their associated outcomes inevitably become partially observable; in this setting, events are determined by timing processes that are not observable (unless a clock is present). The agent must rely on internal timing processes, and tracking the current environment state usually requires a representation of a (possibly short) history of the sequence of events and actions that are not observable at the current moment. A simple example of this is trace conditioning, in which a brief cue is followed by a reward after some finite, possibly variable, delay. Even if the cue is fully observable, the intervening delay to reward makes the underlying state at the time of the reward partially observable, through its dependence on the now-unobservable history of the environment. In particular, the reward is predicted with some probability after observing a cue (in the recent past), but with another probability after observing a reward (that is, in the inter-trial interval), and it behooves the agent to represent in its internal state the most recently observed event (cue or reward). In general, then, the summary representation provided by the state must be internally maintained across different timescales in the context of different goals, recruiting diverse neural timing and memory processes to support learning and decision making across a range of tasks (Paton and Buonomano, 2018).

In sum, when parsing the definition of state as discussed above, there are three components that stand out as important for understanding animal and human behavior during learning and decision making. First, past experience leading to the current moment can influence the current state of the agent, and thus their actions going forward. This potentially extends the internal state representation of a task beyond immediate events, beyond the current trial and even beyond the current task. Importantly, the history of the agent is not limited to events in the environment, but encompasses internal events as well as previous actions. Second, the ability to infer hidden structure in the constellation of past and present experience is crucial for building a state representation that aligns the external and internal states appropriately for the current task. Last, state representations are intrinsically goal-dependent. Whether an agent considers distinct historical experiences as equivalent, and thus treats them as a single state in the current moment, depends critically on the goal of the agent going forward.

4. Uncovering hidden state representations in neurobiology and behavior

The features of experience that comprise the state representation an agent uses to perform a certain task will critically constrain what features of the environment will accrue value (that is, become predictors of rewarding outcomes) and thus modulate decisions about actions in that task. That is, the components of the state are what the agent learns values for. In artificial scenarios, it can be relatively straightforward to determine what the most efficient state representation would be for solving a given learning task (but not always). In such settings, in which the goal of the agent is determined *a priori* as part of the design of the learning problem, it is possible to prescribe that the agent has access to the necessary signals from the environment and is able to disambiguate them clearly. That is, the agent can be endowed with knowledge of which subset of environmental features might be relevant for solving the task (if not which individual feature specifically), thus constraining the size of the state representation of the task, which ensures that learning will be efficient.

However, when discussing learning in animals and humans, the critical elements we outlined above for defining a state representation—how a state representation used by an animal or human specifically depends on the external and internal history and the goal of the agent—is ultimately unknown to us as experimenters. Nevertheless, we can trace the dependence of both behavior and neural activity on the history of both the task and the subject's own behavior, in order to uncover the summary features of experience that the subject is using as a state representation to guide learning.

4.1. Hidden state representations in the neurobiology of reward prediction

A number of now classic findings have identified phasic responses from dopamine neurons in the midbrain as correlates of the reward prediction error signal posited by reinforcement learning theories (e.g., Eshel et al., 2015; Schultz et al., 1997). This correspondence has been influentially modeled using *temporal difference reinforcement learning* (TDRL; (Niv, 2009)), typically assumed to operate on a state representation in which all relevant information is observable at each sequential timepoint (Sutton and Barto, 1998). This assumption implies that information about the current state is available to the agent continuously and unambiguously at each point in time. Given such a state representation, TDRL is able to update state value estimates using a prediction error signal computed on the difference between state value and reward at successive timepoint states (thus *temporal* difference).

The assumption that all relevant information is fully observable to the agent is frequently incorrect, in even the simplest of tasks. For instance, delays between predictive cues and rewarding outcomes introduce into the agent's state representation an internal model of timing, and its associated uncertainty. The fundamental role of internal timing in state representations for learning is evident in the sensitivity of dopamine prediction error signals to the timing of rewards. In tasks in which a delay intervenes between a predictive cue and reward

delivery, it is enough to remember the previous cue to predict *that* reward is coming. However, it turns out that dopaminergic activity is sensitive also to *when* reward should arrive. For example, delaying a reward (or omitting it altogether) results in suppression of dopamine firing (i.e. a negative prediction error) at (or slightly after) the expected time of reward even when no external cue marks this timepoint (Hollerman and Schultz, 1998). State representations for tasks with delay between predictive events and rewards must therefore include a model of elapsing time, in order to propagate the pre-delay (historical) state of the environment forward through the delay during which no observable cue remains to signal the state.

Typically, TDRL models use a temporal representation that consists of a sequence of momentary states (one state per timepoint; called a tapped delay-line) to propagate the history of the environment (i.e. the cue) forward in time, with separate reward predictions learned for each timepoint state (Montague et al., 1996). Such a temporal representation essentially assumes the state representation of the task includes a timing signal that 'remembers' the onset of the past cue perfectly and reliably tracks the elapsing duration from that time until reward delivery. More complex forms of temporal representation have been combined with TDRL to address fundamental properties of timing in reward learning, though many of the assumptions of the learning rule of TDRL (i.e., that values are aggregated and cached as a property of states, rather than computed online using a model of the state transitions in the environment) have typically been left intact (Ludvig et al., 2008, 2012). Whatever the precise nature of the temporal representation an agent recruits for their internal state, timing processes that act as a model of the past are a critical component of state estimation and thus a fundamental component of reward prediction and learning (Gallistel and Gibbon, 2000; Kirkpatrick, 2014).

A number of recent findings confirm that timing processes form a critical, and flexible, component of the state representations that shape reward prediction error computations in dopamine circuits. While TDRL models typically suppose a relatively precise state timing signal, at least one study has found only weak dependence of dopamine prediction errors on the delay to reward after learning, suggesting minimal erosion of reward predictions over elapsing delays (Fiorillo et al., 2008). A more recent study demonstrated that the profile of dopamine prediction errors at the delivery of a time-varying reward critically depends upon the experienced reliability of reward delivery throughout the task (Starkweather et al., 2017). This latter finding is strong evidence that features of past experience that go beyond directly observable cues shape the state representation within which reward predictions are learned. Interestingly, dopamine signals also report prediction errors related to the timing of relatively uninformative events, for instance, the length of the inter-trial interval, indicating that timing processes are recruited widely throughout a task, and not just during anticipation of an impending reward (Bromberg-Martin et al., 2010; Nomoto et al., 2010).

Beyond dopamine, neural activity in the striatum dynamically adjusts to span the temporal interval relevant for reward prediction, confirming that duration is flexibly represented (Mello et al., 2015). Importantly, recent work demonstrated that the temporal precision of dopamine reward prediction errors, and thus presumably the temporal component of state representations for reward prediction, depends on neural circuitry involving the ventral striatum (Takahashi et al., 2016). Finally, judgements about duration on the seconds-long timescale are causally impacted by perturbations of dopamine activity, suggesting dopamine activity itself can modulate the behavioral estimation of the current state (Soares et al., 2016). Together, these findings support theoretical work that has pointed to the critical interplay between timing and state inference in the neural computation of dopamine reward prediction errors (Daw et al., 2006; Langdon et al., 2018).

As mentioned, state representations for reward learning should include any information deemed by an agent as relevant for the task of predicting rewards and acting to attain them. Given that dopamine neurons report prediction errors associated with transitions between such states, task features that influence dopamine signals (beyond the reward itself) can be considered as components of the state representation the animal is using. For instance, temporally extended dopamine responses to perceptually ambiguous stimuli are consistent with inference about the likely hidden state of the task on the basis of the gradually resolved cue (Nomoto et al., 2010; Lak et al., 2017). Further, changes in the identity of equally preferred rewards lead to phasic dopamine responses despite no change in reward value (Takahashi et al., 2017), suggesting a predictive model underlying dopamine firing that operates over multiple dimensions (such as reward identity, not only reward amount; Langdon et al., 2018; Lau et al., 2017).

4.2. Hidden state representations in reward-guided behaviors

An essential window to the hidden state representations that guide reward learning and decision making is provided by behavior itself. In general, there is not a unique state representation that affords learning. As such, the appropriate state representation is usually assumed to be the *minimal* representation of the task that ensures actions can attain the assumed goal. This relates to the definition of equivalence classes introduced earlier: if two distinct sets of experience (histories) are succeeded by the same events *and* outcomes, then the agent should treat these histories as one and the same state. However, many apparently equivalent circumstances are clearly not treated as such by animals engaged in a range of tasks, leading to a rich diversity of behaviors even in controlled laboratory environments.

The interplay between timing and reward-driven behaviors again highlights the role of temporal representations in the hidden state representation of a task. It has long been known that the pattern of behavioral responses (for example, licking) dynamically evolves in the delay between a predictive event, such as the onset of a cue, and a reinforcer, such as a drop of juice, in a way that is strongly influenced by previous experience of the duration of this interval (Gibbon, 1977; Gallistel and Gibbon, 2000; Balsam et al., 2010). Further, delay discounting, in which choice preference between distinct cues is modulated by the expected delay to the associated outcome, attests to the fundamental role of temporal delay in valuation. Temporal discounting behavior is typically modeled by assuming a discounting parameter that controls the rate at which outcomes are devalued by time. Yet the discounting of outcomes according to delay is also susceptible to task features beyond the duration of the pre-reward interval: animal preference in intertemporal choice is also sensitive to post-reward delays, the delivery of additional rewards in the inter-trial interval as well as to the broader framing of the task (Blanchard and Hayden, 2015; Blanchard et al., 2013; Carter and Redish, 2016; Wikenheiser et al., 2013; Williams et al., 2017). Further, the influence of temporal delay on reward-guided behavior indicates animals and humans learn about the distribution of, and relationship between, the delay to different rewards in a task, consistent with the explicit learning of temporal durations as a property of the state representation of a task (De Corte et al., 2018; McGuire and Kable, 2013, 2015; Williams et al., 2017). Understanding states as summaries over history thus naturally recruits the concept of a *timescale* at which the state maintains a representation of previous events: these results imply that the timescale at which a state summarizes history is a malleable, and thus learnable, part of a state representation.

Beyond timing, context more broadly is an important component of state representations. For instance, in reversal learning experiments, rapid reversals are presumably possible because the animal encodes the phase of the experiment (A is rewarded or B is rewarded) as part of the state (Wilson et al., 2014). Representing alternate configurations of a reversal task as separate states allows inference about the current hidden reward contingencies to inform decision making during the task (Costa et al., 2015). This idea of latent 'context' states was tested in an

experiment in which pre-established fear conditioning was extinguished gradually (starting the "extinction" phase with several instances in which the cue was still paired with a foot shock, but with decreasing frequency). Gradual extinction was found to be more effective at ultimately reducing fear of the cue than a standard extinction paradigm where the cue was never followed by shock in the extinction phase, even though the animal actually experienced more shocks in the former case (Gershman et al., 2013; Shiban et al., 2015). These counter-intuitive findings could be explained by the animal forming a hidden state representation that grouped training and extinction into one context state based on the similarity between those two phases in the gradual extinction group, but separating training (the "dangerous" state) from extinction (the "safe" state) in classic extinction. This latter separation into two states would prevent learning in extinction from accessing and modifying the value of the "dangerous" state, thereby leading to the return of fear later in the test phase. In contrast, under the gradual extinction paradigm, because of the gradual decrease in shock frequency, the animal might group all its experiences into a single state and unlearn the prediction of shock during the extinction phase, resulting in minimal return of fear in the test phase.

5. Conclusion: uncovering the 'state' as the basis for learning and decision making

Oftentimes the state representation for a given task is considered a property of the world, as it contains the 'rules' by which the environment evolves given the actions of the agent. Accordingly, when modeling behavior on a learning task, one usually constructs a state representation that captures the known properties of the environment *a priori*, ignoring the fact that the agent may not know what information is necessary for solving the learning problem, or this information may not be available to the agent at the appropriate time to guide their decisions to act.

The central problem for translating concepts from reinforcement learning in artificial environments to understanding learning in animals and humans is that a real agent must first learn the structure of the environment in order to know what to include in its state representation. In studies of decision making in humans, this is a major goal of the instructions phase, however, in animals, this learning must also proceed by trial and error. Moreover, the requisite information for decisions may not be observable, therefore an agent can only base its actions on an internal estimation of the current state and its associated properties. The requirement to determine the current state in order to make decisions and learn effectively poses a particular challenge in environments in which the associative structure between predictors (such as cues) and rewards is hidden. In such settings, learning which features of past experience provide an accurate yet succinct summary relevant for future action is the key to building a useful state representation. Even when all information necessary to determine the current state is clearly observable in the environment, not all features may receive the same attentional focus, biasing learning towards some elements of the environment over others in a manner that can be highly variable across individuals and the duration of the task (Leong et al., 2017). Ultimately, how an agent behaves in the environment depends on what information influences the neurobiological and behavioral processes that support learning and decision making. Theories directed towards understanding how coherent hidden state representations are formed through experience are thus critical for understanding how the computations of learning and decision making are structured in real agents.

We do not yet have a general theory of how animals and humans come to appropriately represent the diverse tasks in which they routinely engage. By tracing the dependence of actions on the latent structure of a task environment we can uncover the dimensions extracted and exploited by real agents to make decisions in the pursuit of rewards. The concept of a state representation is eminently useful for applying theories from reinforcement learning to account for both reward-evoked neural activity and reward-guided behaviors. Theories of learning and decision making that purport to explain behavioral processes in real agents must therefore expand to address questions of how a state representation for any given task is itself learned through experience, why one state representation of a task comes to be used over other alternatives, and what features of the internal and external environment dominate state representations. Answering these questions about how an agent forms a summary of relevant history to support future reward-guided behavior is fundamental for understanding learning and decision making in the diverse tasks in which animals and humans routinely engage.

Funding

This work was supported by National Institutes of Health grant R01DA042065 from NIDA and by the Swartz Center for Theoretical Neuroscience at Princeton University.

Declaration of Competing Interest

The authors declare no competing financial interests.

References

Balsam, P.D., Drew, M.R., Gallistel, C.R., 2010. Time and associative learning. Comp. Cogn. Behav. Rev. 5, 1–22.
Bellman, R., 2013. Dynamic Programming (Courier Corporation).
Berridge, K.C., 2004. Motivation concepts in behavioral neuroscience. Physiol. Behav. 81, 179–209.
Blanchard, T.C., Hayden, B.Y., 2015. Monkeys are more patient in a foraging task than in a standard intertemporal choice task. PLoS One 10 e0117057.
Blanchard, T.C., Pearson, J.M., Hayden, B.Y., 2013. Postreward delays and systematic biases in measures of animal temporal discounting. PNAS 10446.
Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O., 2010. Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. Neuron 67, 144–155.
Carter, E.C., Redish, A.D., 2016. Rats value time differently on equivalent foraging and delay-discounting tasks. J. Exp. Psychol. Gen. 145, 1093–1101.
Costa, V.D., Tran, V.L., Turchi, J., Averbeck, B.B., 2015. Reversal learning and dopamine: a bayesian perspective. J. Neurosci. 35, 2407–2416.
Daw, N.D., Courville, A.C., Touretzky, D.S., 2006. Representation and timing in theories of the dopamine system. Neural Comput. 18, 1637–1677.
De Corte, B.J., Della Valle, R.R., Matell, M.S., 2018. Recalibrating timing behavior via expected covariance between temporal cues. ELife 7, e38790.
Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., Uchida, N., 2015. Arithmetic and local circuitry underlying dopamine prediction errors. Nature 525, 243–246.
Fiorillo, C.D., Newsome, W.T., Schultz, W., 2008. The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci. 11, 966–973.
Gallistel, C.R., Gibbon, J., 2000. Time, rate, and conditioning. Psychol. Rev. 107, 289–344.
Gershman, S.J., Jones, C.E., Norman, K.A., Monfils, M.-H., Niv, Y., 2013. Gradual extinction prevents the return of fear: implications for the discovery of state. Front. Behav. Neurosci. 7.
Gibbon, J., 1977. Scalar expectancy theory and Weber's law in animal timing. Psychol. Rev. 84, 279–325.
Hollerman, J.R., Schultz, W., 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. 1, 304–309.
Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and acting in partially observable stochastic domains. Artif. Intell. 101, 99–134.
Kirkpatrick, K., 2014. Interactions of timing and prediction error learning. Behav. Processes 101, 135–145.
Lak, A., Nomoto, K., Keramati, M., Sakagami, M., Kepecs, A., 2017. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Curr. Biol. 27, 821–832.
Langdon, A.J., Sharpe, M.J., Schoenbaum, G., Niv, Y., 2018. Model-based predictions for dopamine. Curr. Opin. Neurobiol. 49, 1–7.
Lau, B., Monteiro, T., Paton, J.J., 2017. The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. Curr. Opin. Neurobiol. 46, 241–247.
Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., Niv, Y., 2017. Dynamic interaction between reinforcement learning and attention in multidimensional environments. Neuron 93, 451–463.
Ludvig, E.A., Sutton, R.S., Kehoe, E.J., 2008. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput. 20, 3034–3054.
Ludvig, E.A., Sutton, R.S., Kehoe, E.J., 2012. Evaluating the TD model of classical conditioning. Learn. Behav. 40, 305–319.
McGuire, J.T., Kable, J.W., 2013. Rational temporal predictions can underlie apparent failures to delay gratification. Psychol. Rev. 120, 395–410.

McGuire, J.T., Kable, J.W., 2015. Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. Nat. Neurosci. 18, 760–766.
Mello, G.B.M., Soares, S., Paton, J.J., 2015. A scalable population code for time in the striatum. Curr. Biol. 25, 1113–1122.
Minsky, M.L., 1967. Computation: Finite and Infinite Machines. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.
Niv, Y., 2009. Reinforcement learning in the brain. J. Math. Psychol. 53, 139–154.
Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., Wilson, R.C., 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. 35, 8145–8157.
Nomoto, K., Schultz, W., Watanabe, T., Sakagami, M., 2010. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci. 30, 10692–10702.
Paton, J.J., Buonomano, D.V., 2018. The neural basis of timing: distributed mechanisms for diverse functions. Neuron 98, 687–705.
Puterman, M.L., 2014. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons).
Rao, R.P.N., 2010. Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci. 4.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Shiban, Y., Wittmann, J., Weißinger, M., Mühlberger, A., 2015. Gradual extinction reduces reinstatement. Front. Behav. Neurosci. 9.
Soares, S., Atallah, B.V., Paton, J.J., 2016. Midbrain dopamine neurons control judgment of time. Science 354, 1273–1277.
Starkweather, C.K., Babayan, B.M., Uchida, N., Gershman, S.J., 2017. Dopamine reward prediction errors reflect hidden-state inference across time. Nat. Neurosci. 20, 581–589.
Suri, R.E., Schultz, W., 1999. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience 91, 871–890.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: an Introduction. MIT press, Cambridge.
Takahashi, Y.K., Langdon, A.J., Niv, Y., Schoenbaum, G., 2016. Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. Neuron 91, 182–193.
Takahashi, Y.K., Batchelor, H.M., Liu, B., Khanna, A., Morales, M., Schoenbaum, G., 2017. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. Neuron 95 1395-1405.e3.
Wikenheiser, A.M., Stephens, D.W., Redish, A.D., 2013. Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. PNAS 201220738.
Williams, D.A., Todd, T.P., Chubala, C.M., Ludvig, E.A., 2017. Intertrial unconditioned stimuli differentially impact trace conditioning. Learn. Behav. 45, 49–61.
Wilson, R.C., Niv, Y., 2012. Inferring relevance in a changing world. Front. Hum. Neurosci. 5.
Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., Niv, Y., 2014. Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–279.