HHS Public Access

Author manuscript

Behav Neurosci. Author manuscript; available in PMC 2023 October 01.

Published in final edited form as:

Behav Neurosci. 2022 October ; 136(5): 383–391. doi:10.1037/bne0000516.

How do real animals account for the passage of time during associative learning?

Vijay Mohan K Namboodiri[1]

[1]Department of Neurology, Weill Institute for Neuroscience, Center for Integrative Neuroscience, Kavli Institute for Fundamental Neuroscience, Neuroscience Graduate Program, University of California at San Francisco, San Francisco, CA 94158

Abstract

Animals routinely learn to associate environmental stimuli and self-generated actions with their outcomes such as rewards. One of the most popular theoretical models of such learning is the reinforcement learning (RL) framework. The simplest form of RL, model free RL, is widely applied to explain animal behavior in numerous neuroscientific studies. More complex RL versions assume that animals build and store an explicit model of the world in memory. To apply these approaches to explain animal behavior, typical neuroscientific RL models make implicit assumptions about how real animals represent the passage of time. In this perspective, I explicitly list these assumptions and show that they have several problematic implications. I hope that the explicit discussion of these problems encourages the field to seriously examine the assumptions underlying timing and reinforcement learning.

Predicting rewards is essential for the sustained fitness of animals. Since animals (including humans) experience events in their life along the continuously flowing dimension of time, predicting rewards fundamentally requires a consideration of this timeline (Fig 1). Several models have been proposed for how animals learn to predict rewards based on their experience (Balsam et al., 2010; Brandon et al., 2003; Dayan, 1993; Gallistel and Gibbon, 2000; Gallistel et al., 2019; Pearce and Hall, 1980; Rescorla and Wagner, 1972; Schultz, 2016; Sutton and Barto, 1998; Wagner, 1981). Among these, the most widely used class of models in neuroscience is reinforcement learning (RL). The core concept of RL models is that animals make an initial prediction about upcoming reward, calculate a prediction error — the difference between the experienced reward and the predicted reward, and then update their prediction based on this prediction error. The critical issue for a real animal is that it must learn not just that a reward is predicted but also *when* it is predicted. While RL models were inspired by psychological models such as the Rescorla-Wagner (Rescorla and Wagner, 1972) or the Pearce-Hall (Pearce and Hall, 1980) models, mathematically rigorous versions of it have borrowed extensively from concepts in computer science. The fundamental issue that makes these newer models better suited to explaining animal behavior is that they attempt to solve the question of when the reward is predicted, while Rescorla-Wagner and Pearce-Hall models punt on this issue.

VijayMohan.KNamboodiri@ucsf.edu .

Explaining how these newer models account for the passage of time is the key issue considered in this perspective. Briefly, these RL models contend that the learning agent (e.g. real animals) represents the structure of their world in a "state space" that abides by simplified principles such as Markov chains (Niv, 2009; Sutton and Barto, 1998). A state is any abstract representation of observable or unobservable events in their world, and Markov chains are a special kind of state space in which transitions between states do not depend on the history of previously experienced states. Commonly used Markov chain models discretize the flow of time (Niv, 2009; Schultz et al., 1997) or assume temporal basis functions during intervals between events (Gershman et al., 2014; Ludvig et al., 2008, 2012; Petter et al., 2018). These formulations have an intrinsic mathematical simplicity to them, which makes rigorous mathematical calculations possible (e.g., the Bellman equation for value update). Here, I show that these simplifying assumptions have problematic implications when applied to learning in real animals, as they often do not naturally account for the timeline of experience of real animals. My hope is that the explicit treatment considered here stimulates serious considerations of these issues. While the solutions remain to be fully worked out, I believe there will be no progress until the problems are recognized.

Example illustrative task

Perhaps the simplest RL task for animals is cue-reward learning. Most commonly, this is studied in Pavlovian conditioning experiments in which an environmental cue is predictive of an upcoming reward (Pavlov, 1927). Often, there is a delay between when the cue turns off and the subsequent reward delivery (e.g. Bangasser et al., 2006; Beylin et al., 2001; Coddington and Dudman, 2018; Kobayashi and Schultz, 2008; Schultz et al., 1997). This variant of the task is known as trace conditioning. I will use this simple illustrative example throughout this perspective. The main reason for doing so is to show that even the simplest tasks require problematic assumptions. Indeed, the problems laid out here become more severe for tasks requiring reward predictions based on actions. Another reason is that this type of learning, i.e., cue followed by delay followed by reward, is highly ethologically relevant. For instance, for wild foragers, environmental landmarks can often act as "cues" predictive of a reward after some distance (or delay) (Chittka et al., 1995; Wystrach et al., 2019a, 2019b). Similarly, for many animals, cues reflecting the end of winter are predictive of an increased availability of food reward. It is then perhaps not surprising that even insects show evidence of such learning (Chittka et al., 1995; Dylla et al., 2013; Menzel, 2012; Toure et al., 2020; Wystrach et al., 2019a). I will first discuss the common mathematical formulation for representing state space in this task, before discussing implicit assumptions and their problematic implications.

Markov Chains

The mathematical concept of Markov chains is the building block for state space representations in RL. Briefly, a Markov chain is formed by a set of states, $S = \{1, 2, \ldots, n\}$. An implicit assumption is that all the relevant states in the world have been specified in S. The process is assumed to start from one state and successively moves to another state (possibly itself) with a probability p_{ij} (where i and j are indices for the starting and

ending states and can be equal). Each move is called a step. Each step results in a transition, which could be a self-transition to the same state. Crucially, the "transition probabilities" from a given state do not depend on the history of states. If we index the step number (a measure of time) by a subscript t, this means that the conditional probability $p(s_t|s_{t-1}, s_{t-2}, \ldots, s_1)=p(s_t|s_{t-1})$. This absence of history dependence is known as the Markov property and allows some convenient mathematical representations.

The state space in RL is typically such a Markov chain. In more realistic RL formulations, the animal can also take a set of actions $A=\{a_1, \ldots,a_m\}$ that transitions the agent from one state to another with a conditional probability of $p(s_j \mid s_i, a_k)$. These transition probabilities can collectively be represented by a transition matrix P. The state space for an RL agent is fully described by S, A and P. This more general state space that includes an ability of agents to interact with its states using actions is the Markov Decision Process used in RL.

For simplicity, I will only consider the example illustrative task discussed above, in which a reward follows a cue after a delay. Hence, I will omit considerations of actions and the dependence of P on actions. In this task, the states can be minimally specified as the cue state and the reward state. Representing these stimuli as states allows an animal to store the sensory properties of these states in memory. For instance, the animal could learn that an auditory cue has a specific set of sensory attributes such as frequency profile, loudness, duration etc. Similarly, the sensory properties of a type of reward can be represented as a reward state. These various attributes can be stored as part of the memory of that state. Additionally, it is assumed that animals learn a representation of a scalar value for reward. In RL, the reward values are typically denoted by $R(s, a)$, a scalar value associated with each state-action pair. For our purpose, I will denote the reward function by $R(s)$. For the cue and reward state formalism that I adopt, $R(cue)=0$ and $R(reward)=reward\ value$. Thus, S, P, and R completely describe the cue-reward task of interest.

Dealing with time in Markov chains

The biggest problem with the above state space model is that there is no representation of time. The task of interest contains a delay between the cue and reward, and a delay from the reward to the next presentation of the cue (typically called the intertrial interval or ITI). However, there is no representation of these delays in the above Markov chain.

Neuroscience-related RL models solve this problem using the idea of "microstates" (Fig 2). The simplest such model assumes that the delay from cue to reward is represented by a series of states of equal duration (example 1 in Fig 2). This is known as the complete serial compound model of the state space (Moore et al., 1998; Schultz et al., 1997; Sutton and Barto, 1990). Here, the set of states is $S=\{cue, delay^1, delay^2, \ldots, delay^n, reward\}$. This representation (with scalar reward values associated with each state) was used in early work to model temporal difference learning in tasks such as the example considered here (Schultz et al., 1997). An immediate problem with this model is that it does not represent the ITI, an interval that has been shown repeatedly to affect conditioning (Gibbon and Balsam, 1981; Holland, 2000; Kalmbach et al., 2019; Lattal, 1999). The ITI is almost always a random variable with a specified probability distribution. Since Markov chains assume that *all the*

states must be specified, there is no obvious way to break up the ITI into a fixed set of equal duration states obeying the Markov property, unless it is exponentially distributed (K Namboodiri et al., 2021). This problem is usually avoided by only modeling the "trial period", i.e., the delay between cue and reward. However, this is evidently an incomplete representation of the task, the stated goal of a state space. In fact, the moment at which a reward appears in the inter-cue interval can determine whether the observed conditioning to the cue is excitatory or inhibitory (Kaplan, 1984). Thus, modeling only the trial period is evidently insufficient. Nevertheless, this model has proven to be quite successful at explaining numerous aspects of conditioning and thus, has been referred to as a "useful fiction" (Ludvig et al., 2012; Sutton and Barto, 1998).

An extension of this model is to treat the states in the delay not as fixed duration states, but as a set of basis functions (also known as microstates or microstimuli) (Ludvig et al., 2008, 2012) (example 2 in Fig 2). A convenient idea is that the delay after a cue is spanned by a consecutive set of Gaussian states (Ludvig et al., 2008, 2012). In this view, each subsequent state has progressively smaller amplitude and larger width (to approximate scalar timing). This model of state space has benefits over the complete serial compound, as it allows efficient generalization and flexibility due to the non-zero value of many microstates at any given moment (Gershman et al., 2014; Ludvig et al., 2008, 2012; Petter et al., 2018). There is also some evidence for microstate-like activity patterns in brain regions such as the striatum (Mello et al., 2015), hippocampus (MacDonald et al., 2011; Pastalkova et al., 2008; Salz et al., 2016), and the prefrontal cortex (Tiganj et al., 2017). Remarkably, these neural representations can flexibly scale when the delays are altered (MacDonald et al., 2011; Mello et al., 2015). Thus, this microstate model is more consistent with neural data and is functionally advantageous over the complete serial compound. There are many variants of this general idea of a series of microstates (i.e. sequential set of delay states) following a cue (e.g., Brandon et al., 2003; Desmond and Moore, 1988; Grossberg and Schmajuk, 1989; Machado, 1997; Mondragón et al., 2014; Vogel et al., 2003; Wagner, 1981). I will not review these in detail here.

Implicit assumptions

The fundamental premise of the above models is that the delay between different environmental stimuli is a sequence of states in an animal's state space. By breaking the flow of time into such sequences of states, these models make some implicit assumptions. These are often not immediately obvious. I will list some here.

1. Every cue has its own associated set of microstates: the idea of microstates works only if separate cues have separate sets of microstates. Thus, if the animal is learning that cue1 predicts reward1 after delay1 and cue2 predicts reward2 after delay2, the set of microstates during delay1 must be different from the set of microstates during delay2. If not, value learning will be mixed up between the two cues and cannot appropriately assign credit.

2. The microstates are specified *before* value learning: this may be the most important assumption. The entire idea of RL (with value updates to satisfy the Bellman equation) works only if the state space is specified. Thus, before value

learning can occur, the set of sequential microstates following a cue must already exist. I will discuss the problems with this assumption in more detail in the next section.

3. Number and form of microstates are free parameters: another major assumption is that the number and form (e.g., are the basis functions Gaussian?) of microstates are treatable as free parameters for model-fitting. While the lack of principles for the formation of microstates is an obvious problem, this assumption is especially problematic as conditioning in the laboratory can occur over delays of milliseconds or even twenty-four hours (Etscorn and Stephens, 1973; Hinderliter et al., 2012; Kehoe and Macrae, 2002). Presumably, the resolution of microstates evoked over delays of milliseconds is very different from those evoked over hours. However, how does the brain know which microstates to trigger on the first experience of the cue? It is unclear what, if any, principles govern the formation of microstates in the brains of real animals over spans of five orders of magnitude.

4. The microstates during the cue to reward delay are fundamentally different from the microstates during the ITI: In the microstate framework, different delay periods that contain no external stimuli must be treated differently. Thus, the set of microstates during the cue to reward delay must be different from the set of microstates from the reward to cue delay. An explicit treatment of this formulation is found in the SOP model (Wagner, 1981).

5. Learning occurs in trials: another implicit assumption is that value learning occurs progressively by accumulation across trials. It is this trial duration that is assumed to be split into microstates. However, experiments such as the truly random control (Rescorla, 1967, 1968) throw the validity of this assumption into question. In this experiment, the rate of rewards is programmed to be the same during the presence or absence of the cue. Worse, because the events are Poisson processes, they are equally likely to occur at any moment in time. In this case, it is unclear what, if anything, can be treated as a "trial" in the animal's brain.

6. Microstates of a cue must be reproducible across repeated presentations: for learning to occur, if cue1 evokes a set of microstates on one presentation, the same set of microstates must be evoked on the next presentation, to ascribe value to the "correct" microstate.

In the next section, I take a deeper dive into these assumptions and show that the apparent simplicity of the microstate model belies a gargantuan complexity of representation imputed in animal brains. I am by no means the first to discuss some of the problematic implications of these assumptions (Gallistel et al., 2014, 2019; Hallam et al., 1992; Hammond and Paynter Jr, 1983; Luzardo et al., 2017). Nevertheless, the following section focuses on a particularly problematic aspect of these assumptions that has not received as much discussion in the literature.

How bad is the problem really?

The problem is brought into sharp relief when considering initial learning. Remember that the whole point of the formulation of a state space is to explain reward prediction learning. Thus, I will now critically examine the implications of these assumptions for initial learning.

Imagine an animal that is first experiencing a cue that will be followed later by a reward. On this first experience, the animal knows nothing of the significance of this cue (other than its general "salience" or intensity). Indeed, cues are galore in the environments of animals. Nearly every sensory feature of the world could in principle be a cue predictive of a future reward. For instance, maybe a sound is predictive of a future reward. If an animal indeed learns to predict this reward, the above RL algorithms would require the assumption that the sound evokes microstates until the reward *before first learning the relationship of the sound to reward*. This is the whole point of RL: it is a model of learning after all.

What does this imply? This implies that any cue that could *in principle* be predictive of reward must evoke microstates during every presentation. Every sensory stimulus could in principle be such a cue. Hence, for the microstate model to work, animal brains must produce microstates for every sensory stimulus in the experience of the animal. Worse, if cue1 was experienced on two separate days, the set of microstates that were evoked by cue1 should be the same. Thus, the brain must store in memory all the microstates for the nearly infinite number of sensory stimuli, and they must all be discriminable and reliably reproducible on repeated presentations of the stimuli.

The problem is actually significantly worse. This is because the animal does not know at what delay a reward will follow a cue on the first experience of the cue. Indeed, these delays can span five orders of magnitude (Etscorn and Stephens, 1973; Hinderliter et al., 2012; Kehoe and Macrae, 2002). As mentioned above, the data that are often taken as evidence of the existence of neural microstates show that these time representations remap when the delay changes (MacDonald et al., 2011; Mello et al., 2015). How then does the brain know what exact microstates to trigger on the first presentation of the cue, much before the delay to reward is known or learned? Worse still, the brain also must trigger microstates during the delay from the reward to the next cue, for every reward and cue, to learn the distribution of intertrial intervals. How does the brain produce distinct microstates during the ITI and delay to reward on the *first presentation of the cue and reward*? How does the brain know that two delay periods during which no external sensory stimuli exist are fundamentally different *before learning* that there is a relationship between cue and reward? How also does the brain know that delays between different cue-reward-cue triplets are different?

Finally, for simplicity, we have illustrated our main point using the simplest form of RL—one in which the selection of actions to maximize future rewards is not considered. The consideration of time delays becomes even more important for action selection. For instance, real animals often perform reward-related actions after a delay from the corresponding reward-related cues. Indeed, rewards are often predicted by (cue, action) pairs only when there is a specific temporal relation between these events (Miyazaki et al., 2020; Namboodiri et al., 2015; Narayanan and Laubach, 2009). Defining microstates to span these delays

worsens the combinatorial explosion of the state space, as these microstates need to then depend on both external cues and internal actions. Thus, the issues discussed here become even more vexing in this setting.

It is hopefully clear from this examination that the assumption of microstates, while seemingly simple, introduces an untenable solution for an animal brain. Solving these issues is crucial as these issues riddle application of RL to animal learning in even one of the simplest use cases considered here.

Do animals make such assumptions?

We cannot ask animals what assumptions they make during learning. Even in the case of humans where asking is possible, it is unclear if we have conscious access to the neural representations that our brains use for learning. Thus, in some sense, it is impossible to know the exact assumptions used by real animals. Hence, the best approach is likely to make the most parsimonious assumptions that solve the real-world problems faced by animals, in which the timescales of associative relationships are not known *a priori*.

To this end, the answer to a simple question can be illuminative. Is behavioral learning of real animals timescale-invariant or timescale-dependent? The microstate models are inherently dependent on a timescale since the shape and width of the microstates need to be specified prior to learning. In contrast, there is now considerable evidence accumulated over more than fifty years that behavioral learning is largely timescale-invariant. For instance, Fig 3 reproduces a meta-analysis of work in many labs showing that increasing the delay between a cue and outcome does not in fact increase the number of trials until acquisition if the outcome-to-outcome delays are correspondingly scaled by changing the ITI (Gallistel and Gibbon, 2000; Gibbon and Balsam, 1981). Another paper has shown that deleting 7 out of 8 trials does not reduce learning if the temporal spacing between the remaining trials is left unaltered (Gottlieb, 2008). In other words, decreasing the number of trials by a factor of 8 while increasing the ratio between outcome-outcome and cue-outcome intervals by 8 produces the same amount of conditioning. The timescale-invariance observed in these experiments makes little sense if animals make timescale-dependent assumptions about the associative structure of their environment.

Why is the ITI so important in conditioning? Here, I will present a simple intuition for this effect. Imagine that an animal is learning the association between a cue and a reward that follows 10 s later. Now imagine two extreme values of the ITI: 100 s and 1 s. When the ITI is 100 s, the structure of the world is indeed that the cue predicts the reward, as intended by the experimenter. However, when the ITI is 1 s, the delay between a reward and the next cue is much shorter than the delay between a cue and the next reward. In this case, the reward predicts the cue rather than the other way around. Thus, what predicts what depends fundamentally on both the trial and intertrial intervals in an experiment. Once it is clear that the causal structure of a cue-reward association depends on the ITI, it is also relatively straightforward to see that this is likely timescale-invariant. For instance, say we agree that when a cue-reward delay is 10 s, a “reasonably” long ITI to interpret the experiment as cue predicting reward is 100 s. What would such a reasonably long ITI be when the cue-reward

delay is 20 s instead of 10 s? A simple answer is that the corresponding ITI must be increased to 200 s for the structure to remain consistent. This is because when the ITI is scaled to 200 s when the cue-reward delay is increased to 20 s, the corresponding timeline can be perfectly superimposed after scaling on the timeline with 10 s cue-reward delay and 100 s ITI. Thus, what matters to the interpretation of the causal structure is not the absolute value of the cue-outcome delay, but the ratio between that delay and the outcome-outcome delay. This informal intuitive argument rationalizes the observations of timescale-invariance discussed in the previous paragraph.

Similarly, the notion of what constitutes a trial also depends on the ITI. For instance, in the previous examples, when the ITI is 1 s and the cue-reward delay is 10 s, the “experiment” can be thought of as backward conditioning (i.e., reward predicts cue), in which case, the “trial” is demarcated by the reward-to-cue delay instead of the cue-to-reward delay. Indeed, though Rescorla’s work was one of the earliest popularizers of the idea of a trial, he knew that the notion of a trial is unlikely to apply to his subjects, as has been discussed in detail (Gallistel, 2021).

A discussion of alternative frameworks and current limitations

At a superficial level, a major issue is that the assumption of microstates is not a realistic representation of the passage of time for real animals. This issue has been pointed out multiple times (Elman, 1990; Gershman et al., 2014; Ludvig et al., 2008) and can be addressed using more realistic models of timekeeping in the brain (e.g., Buonomano and Merzenich, 1995; Mauk and Buonomano, 2004; Namboodiri et al., 2016; Petter et al., 2018; Simen et al., 2011). However, the more important issue I raise here is regarding the assumptions needed for learning, when timescales are not specified *a priori*. In this light, it is useful to consider a recent highly successful model for conditioning that combines a Rescorla-Wagner rule applied to a drift diffusion model of timing (Luzardo et al., 2017). In this model, cues are postulated to initiate an accumulating timer with a fixed threshold and an adaptable slope. A learning rule adapts the slope based on the knowledge of when the reward happens, thereby adapting the slope of the accumulator to eventually time the cue-reward delay appropriately. This model explains an impressive array of phenomena. It also has a major advantage over the microstate models as it does not postulate an arbitrary number of microstates that span time delays. Nevertheless, it too suffers from similar issues as above when applied to initial learning. For it to work for initial learning, there must be a timer for every cue that could in principle be predictive of reward. As we laid out above, there are almost an infinite number of such cues, each with an infinite set of possible delays to reward. Further, when a timer is initiated at cue onset on the first time that the cue was experienced, how does the timer know that it is timing a specific upcoming reward? What if the cue does predict reward at a fixed delay but there are other intermittent cues in this delay? What if this cue was only predictive of another cue and not a reward or was not predictive of anything in particular at all? What if the reward predicted by the cue is not the immediately following reward? How does the timer get feedback about exactly which interval it is supposed to time? These issues are solvable only if the animal knows that it is timing the interval between a specific cue state and a reward state, or in other words, *after*

learning that a specific cue and a specific reward may be related. Thus, this approach does not prescribe how an animal can initially learn the association between a cue and reward.

Another formalism within the general framework of RL is a semi-Markov/Markov renewal process model that explicitly learns the distribution of time intervals between consecutive state transitions (Bradtke and Duff, 1994; Daw et al., 2006; Namboodiri, 2021). Though very similar, there is one difference between a semi-Markov and Markov renewal process-based state space. In a simple cue-reward task, a semi-Markov state space treats the state of the world as the cue state (or the interstimulus interval state) during the delay between cue and reward (Bradtke and Duff, 1994; Daw et al., 2006). As the value of states is tied to the currently active state in a semi-Markov model, the value is stationary for the duration of this entire state. However, in a Markov renewal process, the value function can be defined in continuous time, as the states and transition times are treated separately (Namboodiri, 2021). While these models avoid any issues with breaking up the flow of time into states, they nevertheless suffer from some limitations. The key limitation is that by only learning the intervals between consecutive events in the world, such learning is very sensitive to the presence of distractors. Though existing evidence supports the prediction that distractor states impede trace conditioning (Carter et al., 2003; Clark et al., 2002; Han et al., 2003; Manns et al., 2000), it is unclear how such learning can adapt to the real world where the delay between most cues and their predicted reward is filled with other sensory stimuli. How could an animal learn to treat these intermittent sensory stimuli as distractors and not themselves predictive of subsequent outcomes? If multiple Markov chains occur simultaneously in a mixture distribution, the above learning algorithms will not be able to demix them without additional mechanisms. Hence, these algorithms still have major limitations as currently prescribed.

Another possible solution to the problem is to move completely away from RL and propose that other quantities control learning in tasks such as the one considered here. A set of models that propose that animals learn contingency (defined as normalized gain in available information) between the *timing* of reward predictors and rewards belongs to this class (Balsam et al., 2010; Gallistel et al., 2014, 2019; Ward et al., 2012). These models are successful at explaining numerous aspects of the learning of conditioned responses in relation to the various time intervals. Further, they can work for initial learning in a timescale invariant fashion. This model solves essentially all concerns listed above regarding initial learning by proposing that animals store their timeline of experience in memory and operate on this temporal map to uncover associations. However, whether a complicated computation such as a timescale-invariant mutual information can be calculated in an online manner by a neural network needs to be demonstrated both theoretically and experimentally. Indeed, it has been suggested that such computations can only be performed offline by intracellular molecular machinery (Gallistel, 2017) (see (Akhlaghpour, 2022; Gallistel et al., 2020; Thornquist et al., 2020) for example models of such computations). Further, this theory makes the strong prediction that cue-outcome associations with any arbitrarily long interceding delay can be learned provided the outcome-to-outcome delays are sufficiently long. This strong claim remains to be quantitatively tested in a manner that separates learning from the magnitude of performance. Currently, some evidence suggests that there may be some limitations to scale invariance. For instance, trace eyeblink conditioning has

not been successfully demonstrated beyond a few seconds of trace interval (unclear if long trace interval experiments have been attempted with correspondingly long outcome-outcome delays), and recent data suggest that cue-reward associations with a 60s delay are not well-learned even when the inter-reward interval is more than 4000 s (Thrailkill et al., 2020) (though this study did show measurable behavioral learning and was not focused on analyzing the onset of acquisition). Overall, the strong prediction of this information theoretic model remains to be fully tested. Lastly, this model has not as yet been extended to sequences of stimuli that predict reward, and does not provide a clear explanation for the prediction error signals observed in midbrain dopamine neurons, which are known to drive learning in animals (Cohen et al., 2012; Kim et al., 2020; Mohebi et al., 2019; Schultz, 2016; Schultz et al., 1997).

Lastly, a recent theoretical framework proposes a solution to learn timescale-invariant predictions of the future using timescale-invariant representations of the past (Goh et al., 2021). This model uses a Laplace transform to store recent history in a timeline of events (Shankar and Howard, 2012) and uses this timeline to make a projection into the future from the present moment. This model effectively implements a neural network instantiation of the temporal map that was posited as a necessary requirement for learning (Balsam and Gallistel, 2009). Just like the contingency of timing model, this learning model also avoids the concerns raised above regarding initial learning. Since this model has been directly extended into a neural network model, it is biologically plausible based on current neuroscientific understanding. It also shows early promise in explaining dopamine dynamics (Goh et al., 2021). Despite these many appeals, it remains to be tested whether it can quantitatively fit behavioral learning and neural signals for learning such as dopamine release in the striatum.

Conclusions

Here, I explicitly list the assumptions made by well-known RL models that account for the passage of time. I show that the apparent superficial simplicity of these models belies the extraordinary complexity required to execute them. These problems are often not recognized, as researchers define a convenient state space for each experiment using free parameters. These assumptions are especially problematic when applied to initial learning, a stated goal of reinforcement learning. The fundamental issue is that breaking up the flow of time into discrete states is a problematic way to learn associations across temporal delays. This is because it is unclear prior to initial learning which intervals are worth timing. Further, learning temporal relationships only between consecutive events raises concerns as to how initial learning would be possible in the real world where many other sensory stimuli span such delays. While the solutions remain to be worked out, I hope that this perspective highlights the issues with current models. Until we acknowledge the problems, there will not be any solutions.

Acknowledgments and Funding

I thank J. Berke, C.R. Gallistel, S. Mihalas, M. Howard, and members of the Namboodiri lab for helpful comments that instructed this perspective. This work was funded by grants from the National Institute of Mental Health

(R00MH118422), and the Brain and Behavior Research Foundation (NARSAD Young Investigator Award) to V.M.K.N.

References

Akhlaghpour H (2022). An RNA-based theory of natural universal computation. Journal of Theoretical Biology 537, 110984.

Balsam PD, and Gallistel CR (2009). Temporal maps and informativeness in associative learning. Trends Neurosci 32, 73–78. [PubMed: 19136158]

Balsam PD, Drew MR, and Gallistel CR (2010). Time and Associative Learning. Comp Cogn Behav Rev 5, 1–22. [PubMed: 21359131]

Bangasser DA, Waxler DE, Santollo J, and Shors TJ (2006). Trace Conditioning and the Hippocampus: The Importance of Contiguity. J Neurosci 26, 8702–8706. [PubMed: 16928858]

Beylin AV, Gandhi CC, Wood GE, Talk AC, Matzel LD, and Shors TJ (2001). The role of the hippocampus in trace conditioning: temporal discontinuity or task difficulty? Neurobiol Learn Mem 76, 447–461. [PubMed: 11726247]

Bradtke SJ, and Duff MO (1994). Reinforcement learning methods for continuous-time Markov decision problems. In Proceedings of the 7th International Conference on Neural Information Processing Systems, (Cambridge, MA, USA: MIT Press), pp. 393–400.

Brandon SE, Vogel EH, and Wagner AR (2003). Stimulus representation in SOP: I. Theoretical rationalization and some implications. Behav Processes 62, 5–25. [PubMed: 12729966]

Buonomano DV, and Merzenich MM (1995). Temporal Information Transformed into a Spatial Code by a Neural Network with Realistic Properties. Science 267, 1028–1030. [PubMed: 7863330]

Carter RM, Hofstötter C, Tsuchiya N, and Koch C (2003). Working memory and fear conditioning. PNAS 100, 1399–1404. [PubMed: 12552137]

Chittka L, Geiger K, and Kunze J (1995). The influences of landmarks on distance estimation of honey bees. Animal Behaviour 50, 23–31.

Clark RE, Manns JR, and Squire LR (2002). Classical conditioning, awareness, and brain systems. Trends Cogn Sci 6, 524–531. [PubMed: 12475713]

Coddington LT, and Dudman JT (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat. Neurosci 21, 1563–1573. [PubMed: 30323275]

Cohen JY, Haesler S, Vong L, Lowell BB, and Uchida N (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482, 85–88. [PubMed: 22258508]

Daw ND, Courville AC, Tourtezky DS, and Touretzky DS (2006). Representation and timing in theories of the dopamine system. Neural Comput 18, 1637–1677. [PubMed: 16764517]

Dayan P (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation 5, 613–624.

Desmond JE, and Moore JW (1988). Adaptive timing in neural networks: The conditioned response. Biol. Cybern 58, 405–415. [PubMed: 3395634]

Dylla KV, Galili DS, Szyszka P, and Lüdke A (2013). Trace conditioning in insects-keep the trace! Front Physiol 4, 67. [PubMed: 23986710]

Elman JL (1990). Finding Structure in Time. Cognitive Science 14, 179–211.

Etscorn F, and Stephens R (1973). Establishment of conditioned taste aversions with a 24-hour CS-US interval. Physiological Psychology 1, 251–259.

Gallistel CR (2017). The Coding Question. Trends in Cognitive Sciences 21, 498–508. [PubMed: 28522379]

Gallistel CR (2021). Robert Rescorla: Time, Information and Contingency. Revista de Historia de La Psicología 42, 7–21.

Gallistel CR, and Gibbon J (2000). Time, rate, and conditioning. Psychol Rev 107, 289–344. [PubMed: 10789198]

Gallistel CR, Craig AR, and Shahan TA (2014). Temporal contingency. Behav Processes 101, 89–96. [PubMed: 23994260]

Gallistel CR, Craig AR, and Shahan TA (2019). Contingency, contiguity, and causality in conditioning: Applying information theory and Weber's Law to the assignment of credit problem. Psychol Rev 126, 761–773. [PubMed: 31464474]

Gallistel CR, Johansson F, Jirenhed D-A, Rasmussen A, Ricci M, and Hesslow G (2020). Quantitative Properties of the Creation and Activation of a Cell-Intrinsic Engram (bioRxiv).

Gershman SJ, Moustafa AA, and Ludvig EA (2014). Time representation in reinforcement learning models of the basal ganglia. Front. Comput. Neurosci 7.

Gibbon J, and Balsam P (1981). Spreading associations in time. In Autoshaping and Conditioning Theory, Locurto CM, Terrace HS, and Gibbon J, eds. (New York: Academic), pp. 219–253.

Goh WZ, Ursekar V, and Howard MW (2021). Predicting the future with a scale-invariant temporal memory for the past. ArXiv:2101.10953 [Cs, q-Bio].

Gottlieb DA (2008). Is the number of trials a primary determinant of conditioned responding? J Exp Psychol Anim Behav Process 34, 185–201. [PubMed: 18426303]

Grossberg S, and Schmajuk NA (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. Neural Networks 2, 79–102.

Hallam SC, Grahame NJ, and Miller RR (1992). Exploring the edges of Pavlovian contingency space: An assessment of contingency theory and its various metrics. Learning and Motivation 23, 225–249.

Hammond LJ, and Paynter WE Jr (1983). Probabilistic contingency theories of animal conditioning: A critical analysis. Learning and Motivation 14, 527–550.

Han CJ, O'Tuathaigh CM, van Trigt L, Quinn JJ, Fanselow MS, Mongeau R, Koch C, and Anderson DJ (2003). Trace but not delay fear conditioning requires attention and the anterior cingulate cortex. PNAS 100, 13087–13092. [PubMed: 14555761]

Hinderliter CF, Andrews A, and Misanin JR (2012). The Influence of Prior Handling on the Effective CS-US Interval in Long-Trace Taste-Aversion Conditioning in Rats. Psychol Rec 62, 91–96.

Holland PC (2000). Trial and intertrial durations in appetitive conditioning in rats. Animal Learning & Behavior 28, 121–135.

K Namboodiri VM, Hobbs T, Trujillo-Pisanty I, Simon RC, Gray MM, and Stuber GD (2021). Relative salience signaling within a thalamo-orbitofrontal circuit governs learning rate. Current Biology.

Kalmbach A, Chun E, Taylor K, Gallistel CR, and Balsam PD (2019). Time-scale-invariant information-theoretic contingencies in discrimination learning. Journal of Experimental Psychology: Animal Learning and Cognition 45, 280.

Kaplan PS (1984). Importance of relative temporal parameters in trace autoshaping: From excitation to inhibition. Journal of Experimental Psychology: Animal Behavior Processes 10, 113–126.

Kehoe EJ, and Macrae M (2002). Fundamental behavioral methods and findings in classical conditioning. In A Neuroscientist's Guide to Classical Conditioning, (Springer), pp. 171–231.

Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, Zhang Y, Li Y, Watabe-Uchida M, Gershman SJ, et al. (2020). A Unified Framework for Dopamine Signals across Timescales. Cell 183, 1600–1616.e25. [PubMed: 33248024]

Kobayashi S, and Schultz W (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci 28, 7837–7846. [PubMed: 18667616]

Lattal KM (1999). Trial and intertrial durations in Pavlovian conditioning: issues of learning and performance. J Exp Psychol Anim Behav Process 25, 433–450. [PubMed: 17763570]

Ludvig EA, Sutton RS, and Kehoe EJ (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Computation 20, 3034–3054. [PubMed: 18624657]

Ludvig EA, Sutton RS, and Kehoe EJ (2012). Evaluating the TD model of classical conditioning. Learning & Behavior 40, 305–319. [PubMed: 22927003]

Luzardo A, Alonso E, and Mondragón E (2017). A Rescorla-Wagner drift-diffusion model of conditioning and timing. PLOS Computational Biology 13, e1005796. [PubMed: 29095819]

MacDonald CJ, Lepage KQ, Eden UT, and Eichenbaum H (2011). Hippocampal "Time Cells" Bridge the Gap in Memory for Discontiguous Events. Neuron 71, 737–749. [PubMed: 21867888]

Machado A (1997). Learning the temporal dynamics of behavior. Psychol Rev 104, 241–265. [PubMed: 9127582]

Manns JR, Clark RE, and Squire LR (2000). Awareness predicts the magnitude of single-cue trace eyeblink conditioning. Hippocampus 10, 181–186. [PubMed: 10791840]

Mauk MD, and Buonomano DV (2004). The neural basis of temporal processing. Annu. Rev. Neurosci 27, 307–340. [PubMed: 15217335]

Mello GBM, Soares S, and Paton JJ (2015). A Scalable Population Code for Time in the Striatum. Current Biology 25, 1113–1122. [PubMed: 25913405]

Menzel R (2012). The honeybee as a model for understanding the basis of cognition. Nature Reviews Neuroscience 13, 758–768. [PubMed: 23080415]

Miyazaki K, Miyazaki KW, Sivori G, Yamanaka A, Tanaka KF, and Doya K (2020). Serotonergic projections to the orbitofrontal and medial prefrontal cortices differentially modulate waiting for future rewards. Sci Adv 6.

Mohebi A, Pettibone JR, Hamid AA, Wong J-MT, Vinson LT, Patriarchi T, Tian L, Kennedy RT, and Berke JD (2019). Dissociable dopamine dynamics for learning and motivation. Nature 570, 65–70. [PubMed: 31118513]

Mondragón E, Gray J, Alonso E, Bonardi C, and Jennings DJ (2014). SSCC TD: A Serial and Simultaneous Configural-Cue Compound Stimuli Representation for Temporal Difference Learning. PLOS ONE 9, e102469. [PubMed: 25054799]

Moore JW, Choi J-S, and Brunzell DH (1998). Predictive timing under temporal uncertainty: the time derivative model of the conditioned response. Timing of Behavior: Neural, Psychological, and Computational Perspectives 3–34.

Namboodiri VMK (2021). What is the state space of the world for real animals? BioRxiv 2021.02.07.430001.

Namboodiri VMK, Huertas MA, Monk KJ, Shouval HZ, and Hussain Shuler MG (2015). Visually cued action timing in the primary visual cortex. Neuron 86, 319–330. [PubMed: 25819611]

Namboodiri VMK, Mihalas S, and Hussain Shuler MG (2016). Analytical Calculation of Errors in Time and Value Perception Due to a Subjective Time Accumulator: A Mechanistic Model and the Generation of Weber's Law. Neural Comput 28, 89–117. [PubMed: 26599714]

Narayanan NS, and Laubach M (2009). Delay activity in rodent frontal cortex during a simple reaction time task. J Neurophysiol 101, 2859–2871. [PubMed: 19339463]

Niv Y (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology 53, 139–154.

Pastalkova E, Itskov V, Amarasingham A, and Buzsáki G (2008). Internally Generated Cell Assembly Sequences in the Rat Hippocampus. Science 321, 1322–1327. [PubMed: 18772431]

Pavlov IP (1927). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex (Oxford, England: Oxford Univ. Press).

Pearce JM, and Hall G (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87, 532–552. [PubMed: 7443916]

Petter EA, Gershman SJ, and Meck WH (2018). Integrating Models of Interval Timing and Reinforcement Learning. Trends in Cognitive Sciences 22, 911–922. [PubMed: 30266150]

Rescorla RA (1967). Pavlovian conditioning and its proper control procedures. Psychol Rev 74, 71–80. [PubMed: 5341445]

Rescorla RA (1968). Probability of shock in the presence and absence of cs in fear conditioning. Journal of Comparative and Physiological Psychology 66, 1–5. [PubMed: 5672628]

Rescorla RA, and Wagner AR (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Current Research and Theory 2, 64–99.

Salz DM, Tiganj Z, Khasnabish S, Kohley A, Sheehan D, Howard MW, and Eichenbaum H (2016). Time Cells in Hippocampal Area CA3. J. Neurosci 36, 7476–7484. [PubMed: 27413157]

Schultz W (2016). Dopamine reward prediction error coding. Dialogues Clin Neurosci 18, 23–32. [PubMed: 27069377]

Schultz W, Dayan P, and Montague PR (1997). A Neural Substrate of Prediction and Reward. Science 275, 1593–1599. [PubMed: 9054347]

Shankar KH, and Howard MW (2012). A scale-invariant internal representation of time. Neural Comput 24, 134–193. [PubMed: 21919782]

Simen P, Balci F, deSouza L, Cohen JD, and Holmes P (2011). A Model of Interval Timing by Neural Integration. J. Neurosci 31, 9238–9253. [PubMed: 21697374]

Sutton RS, and Barto AG (1990). Time-derivative models of pavlovian reinforcement.

Sutton RS, and Barto AG (1998). Introduction to Reinforcement Learning (Cambridge, MA, USA: MIT Press).

Thornquist SC, Langer K, Zhang SX, Rogulja D, and Crickmore MA (2020). CaMKII Measures the Passage of Time to Coordinate Behavior and Motivational State. Neuron 105, 334–345.e9. [PubMed: 31786014]

Thrailkill EA, Todd TP, and Bouton ME (2020). Effects of conditioned stimulus (CS) duration, intertrial interval, and I/T ratio on appetitive Pavlovian conditioning. Journal of Experimental Psychology: Animal Learning and Cognition 46, 243–255. [PubMed: 32175762]

Tiganj Z, Jung MW, Kim J, and Howard MW (2017). Sequential Firing Codes for Time in Rodent Medial Prefrontal Cortex. Cerebral Cortex 27, 5663–5671. [PubMed: 29145670]

Toure MW, Young FJ, McMillan WO, and Montgomery SH (2020). Heliconiini butterflies can learn time-dependent reward associations. Biology Letters 16, 20200424.

Vogel EH, Brandon SE, and Wagner AR (2003). Stimulus representation in SOP:: II. An application to inhibition of delay. Behavioural Processes 62, 27–48. [PubMed: 12729967]

Wagner AR (1981). SOP: A model of automatic memory processing in animal behavior. Information Processing in Animals: Memory Mechanisms 85, 5–47.

Ward RD, Gallistel CR, Jensen G, Richards VL, Fairhurst S, and Balsam PD (2012). CS Informativeness Governs CS-US Associability. J Exp Psychol Anim Behav Process 38, 217–232. [PubMed: 22468633]

Wystrach A, Buehlmann C, Schwarz S, Cheng K, and Graham P (2019a). Avoiding pitfalls: Trace conditioning and rapid aversive learning during route navigation in desert ants. BioRxiv 771204.

Wystrach A, Schwarz S, Graham P, and Cheng K (2019b). Running paths to nowhere: repetition of routes shows how navigating ants modulate online the weights accorded to cues. Anim Cogn 22, 213–222. [PubMed: 30684062]

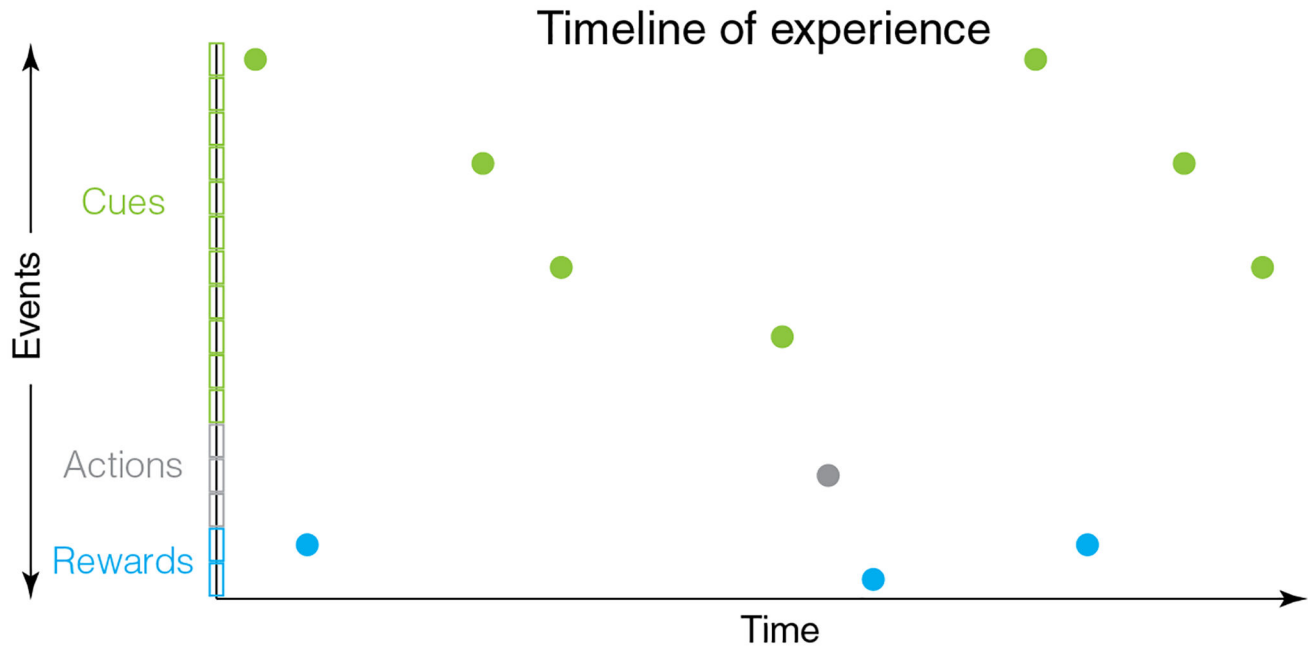

Fig 1.

Animals experience events in their life in a timeline along the continuously flowing dimension of time. Thus, prediction of rewards requires a consideration of the flow of time. Here, external cues, internally generated actions and rewards are shown by separate colors. Distinct types of events within these groups are shown by individual boxes along the y-axis.

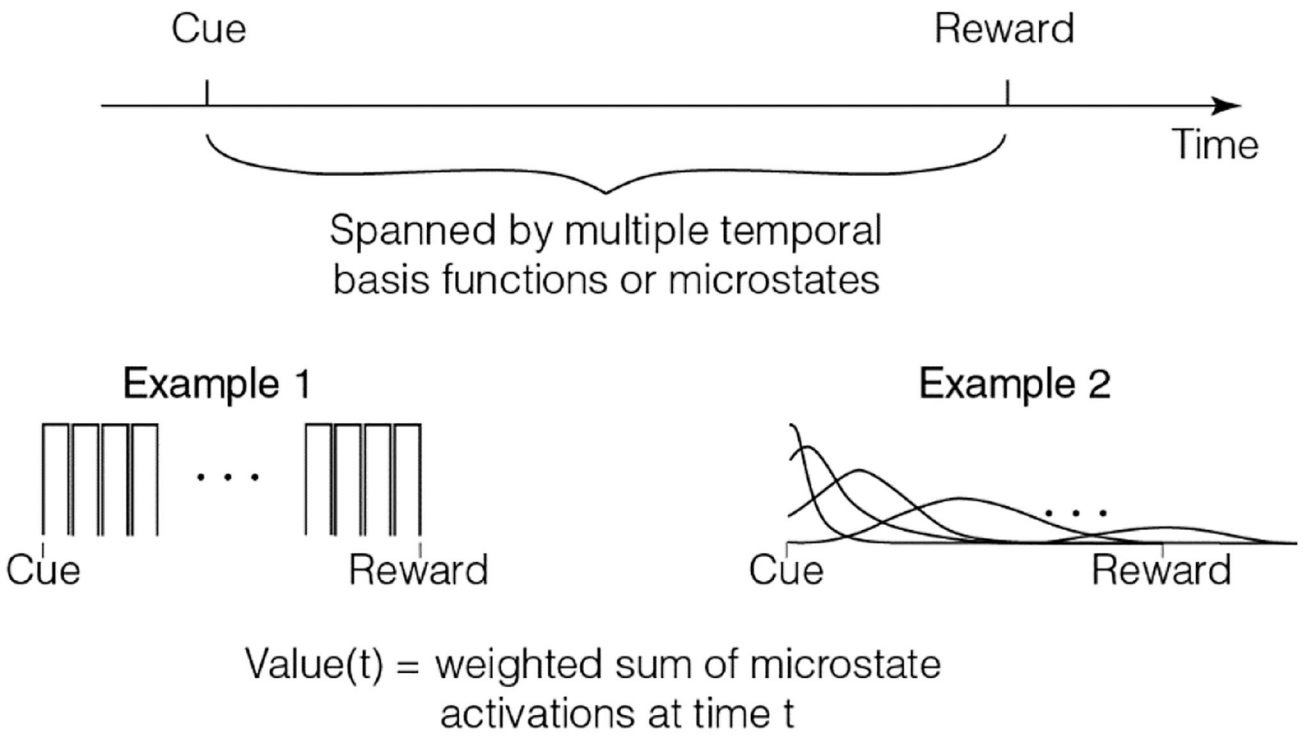

Fig 2.

Common models for dealing with delays between cue and reward assume that such delays are spanned by multiple microstates. Two examples are shown here (see text). As can be seen, these formulations assume that the delay periods themselves are represented by many states to which an RL algorithm can attach value.

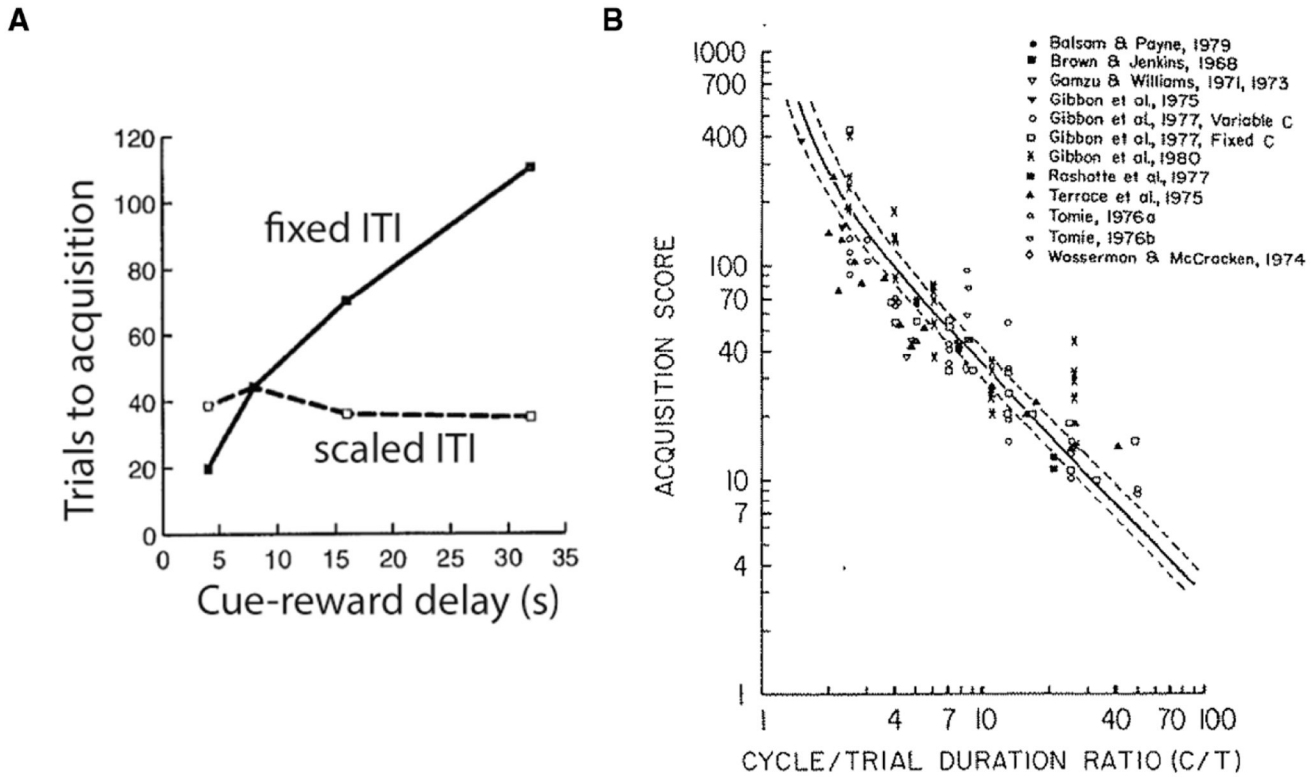

Fig 3.

Timescale-invariance of behavioral learning. A. Reproduced from Gallistel and Gibbon, 2000. When the cue-reward delay is increased, the number of trials to acquisition increases only when the ITI is fixed. When the ITI is correspondingly scaled, the number of trials to acquisition remains largely constant. B. Reproduced from a meta-analysis published in Balsam et al. 1981. When the ratio between outcome-outcome delay (called cycle duration) and cue-outcome duration (called trial duration) is changed, the number of trials to acquisition varies in a predictable manner.