Psychonomic Bulletin and Review (2020) 27:1166–1194
https://doi.org/10.3758/s13423-020-01749-0

A-learning: A new formulation of associative learning theory

Stefano Ghirlanda[1,2] · Johan Lind[2] · Magnus Enquist[2]

Published online: 6 July 2020

Abstract

We present a new mathematical formulation of associative learning focused on non-human animals, which we call A-learning. Building on current animal learning theory and machine learning, A-learning is composed of two learning equations, one for stimulus-response values and one for stimulus values (conditioned reinforcement). A third equation implements decision-making by mapping stimulus-response values to response probabilities. We show that A-learning can reproduce the main features of: instrumental acquisition, including the effects of signaled and unsignaled non-contingent reinforcement; Pavlovian acquisition, including higher-order conditioning, omission training, autoshaping, and differences in form between conditioned and unconditioned responses; acquisition of avoidance responses; acquisition and extinction of instrumental chains and Pavlovian higher-order conditioning; Pavlovian-to-instrumental transfer; Pavlovian and instrumental outcome revaluation effects, including insight into why these effects vary greatly with training procedures and with the proximity of a response to the reinforcer. We discuss the differences between current theory and A-learning, such as its lack of stimulus-stimulus and response-stimulus associations, and compare A-learning with other temporal-difference models from machine learning, such as Q-learning, SARSA, and the actor-critic model. We conclude that A-learning may offer a more convenient view of associative learning than current mathematical models, and point out areas that need further development.

Keywords Associative learning · Pavlovian conditioning · Instrumental conditioning · Mathematical model · Conditioned reinforcement · Outcome revaluation

In this theoretical paper, we introduce a new mathematical formulation of associative learning theory, which we call A-learning as it focuses on non-human animals. A-learning builds on current theory in psychology and machine learning, and offers two main improvements over existing mathematical models. First, it shows that associative learning can take into account the future value of stimuli and responses, which enables the acquisition and extinction of complex sequences of behavior. Second, A-learning suggests how instrumental and Pavlovian learning can be integrated into one mathematical model, which is ultimately necessary as most learning situations include both Pavlovian and instrumental contingencies (Mackintosh & Dickinson, 1979; Mackintosh, 1983).

The paper is structured as follows. This Introduction summarizes the accomplishments of associative learning theory, and its limitations. The latter concern the relationship between Pavlovian and instrumental conditioning, which associations underlie these two kinds of learning, and how these associations determine behavior. The Introduction concludes with a brief rationale for A-learning's use of temporal-difference (TD) learning algorithms (Sutton & Barto, 2018). The Theory section presents A-learning and compares it to current theory. The Results section applies A-learning to associative phenomena, such as the acquisition of instrumental and Pavlovian responses and sequences of responses, matching, behavioral contrast, Pavlovian-to-instrumental transfer (PIT; Cartoni et al. (2016)), and outcome revaluation. Finally, the Discussion summarizes how

This work is dedicated to the memory of Nick Mackintosh, whom we knew only superficially in person, but whose work we have studied in depth. This research was supported by the Knut and Alice Wallenberg Foundation, grant 2015–0005, and by a fellowship leave to SG from Brooklyn College. Simulations of the partial reinforcement effect were developed in collaboration with Brittany Lozada and Yoni Zomick within the Brooklyn College Neuroscience REU, NSF award #1461099.

✉ Stefano Ghirlanda
drghirlanda@gmail.com

[1] Brooklyn College and Graduate Center, CUNY, New York, NY, USA

[2] Stockholm University, Stockholm, Sweden

A-learning may contribute to a more complete associative learning theory, and what developments are still necessary.

It is widely acknowledged that neither Pavlovian nor instrumental learning can be understood with a single kind of learning variable, be it stimulus-response (S-R), response-stimulus (R-S), or stimulus-stimulus (S-S) associations (Hall, 2002; Holland, 2008; Pearce, 2008). A-learning takes the logical next step and, similarly to two-factor theories (Mowrer, 1960; Mackintosh, 1983), attempts to reproduce Pavlovian and instrumental learning using two kinds of learning variables: S-R values (akin to S-R associations) and stimulus values (blending elements of S-S associations and conditioned reinforcement). We aim to show that A-learning can reproduce a wide variety of findings from simple learning principles.

State of the art

Over more than one hundred years, psychologists have built an impressive understanding of associative learning, covering many animal learning phenomena and implicit learning in humans. Associative learning theory also underpins some of the most successful applications of psychology to animal welfare and training (McGreevy & Boakes, 2011), and to human health (Bernstein, 1999; Haselgrove & Hogarth, 2013; Schachtman & Reilly, 2011). Current understanding of associative learning is based on the following elements.

Well-established learning phenomena and experimental paradigms These include acquisition and extinction of responding, stimulus control (discrimination and generalization), associative competition (blocking, overshadowing, etc.), conditioned reinforcement, and outcome devaluation (Pearce, 2008; Bouton, 2016).

Theoretical reasoning around learning phenomena Such reasoning comprises ideas about how associations change with experience and, even more fundamentally, what associations are learned. Current theory is mostly phrased in terms of stimulus-response (S-R), stimulus-stimulus (S-S), and response-outcome (R-O) associations, although more complex constructs such as S-R-O associations are sometimes employed (Balleine & Dickinson, 1998; Pearce, 2008; Hall, 2002; Bouton, 2016).

A pervasive distinction between Pavlovian and instrumental learning The traditional view that Pavlovian and instrumental learning are separate processes (Konorski & Miller, 1937; Skinner, 1937) is still widely held among non-specialists, and constitutes the backbone of all contemporary textbooks (Pearce, 2008; Frieman & Reilly, 2015; Bouton, 2016). Although modern theorists recognize major similarities between Pavlovian and instrumental learning, such as the effects of contingency and associative competition (Mackintosh, 1994; Pearce, 2008; Bouton, 2016), there is a continuing effort to tease apart Pavlovian and instrumental contributions to learning (Mackintosh, 1994; Hall, 2002).

Mathematical models that yield insight into many phenomena Mathematical models are most developed for Pavlovian conditioning (Pearce, 2008; Bouton, 2016), although the possibility of using similar principles for instrumental learning is recognized (Bush & Mosteller, 1951; Blough, 1975). Despite efforts, there is no accepted mathematical theory that includes both Pavlovian and instrumental learning, although there are verbal models (Dickinson, 1980; Balleine & Dickinson, 1998; Mackintosh, 1983; Hall, 2002).

Altogether, this combination of empirical data, concepts, and models delivers a remarkably detailed understanding of associative learning. At the same time, various phenomena resist a satisfactory explanation, which we believe stems from the unresolved foundational issues discussed next.

The Pavlovian-instrumental distinction is not fully understood

The distinction is clear-cut procedurally: Pavlovian procedures arrange stimulus-stimulus (S-S) contingencies (e.g., food is delivered after ringing a bell), while instrumental procedures employ response-stimulus (R-S) contingencies (e.g., food is delivered after a lever press; Mackintosh (1994)). It has long been recognized that this apparent simplicity hides several pitfalls. Arranging a Pavlovian contingency may also introduce instrumental ones. Thus, a dog might salivate in response to a bell that signals food because salivation improves chewing or the taste of food—a possibility that can be discounted only after extensive investigation (Coleman & Gormezano, 1979). Likewise, an instrumental contingency may induce Pavlovian ones: a rat that learns to press a lever for food is also exposed to the S-S contingency between the sight of the lever and the food (Trapold & Overmier, 1972). These unintended contingencies are important because most responses are sensitive to both Pavlovian and instrumental contingencies (Mackintosh, 1983). Even salivation—the prototypical Pavlovian response—can be reinforced instrumentally (Miller & Carmona, 1967; Shapiro & Herendeen, 1975), and even lever pressing—the prototypical instrumental response—can be modified by Pavlovian contingencies (Atnip, 1977). Indeed, from the animal's perspective, it is hard to know whether an experiment is "Pavlovian" or "instrumental," and thus what to learn about. It follows that, unless very special procedures are adopted, most learned behavior reflects both Pavlovian and instrumental contingencies (Hearst, 1975; Mackintosh, 1983). These facts are often stated (Mackintosh, 1994; Balleine, 2011; Bouton, 2016), yet we lack a mathematical model that includes both Pavlovian and instrumental effects. Without such a model, some of the most interesting associative phenomena remain difficult to understand because they rely on Pavlovian-instrumental interplay, such as outcome devaluation and Pavlovian-to-instrumental transfer.

Uncertainty about associative structures

It is clear that associative learning cannot be explained using only one kind of association (e.g., S-R or S-S), but debate continues about the role of different kinds of associations. Both S-R and S-S associations are deemed important in Pavlovian conditioning, with each kind prevailing under certain, incompletely understood circumstances (Hall, 2002; Holland, 2008). For example, first-order Pavlovian conditioning is understood as producing primarily S-S associations, while second-order conditioning appears to result mainly in S-R associations (Rescorla (1973) and Rescorla (1980); see Discussion).

Associative structures in instrumental learning are also not fully understood. Early S-R theories (Thorndike, 1911; Guthrie, 1942) have given way to conceptual models that emphasize the role of response-reinforcer associations (R-S), although some findings still fit the old theories better (see Hall (2002) and Bouton (2016), and Outcome Revaluation below). In particular, antecedent stimuli must play a role in initiating instrumental responses (e.g., the sight of a lever), which is not captured by the notion of a response-reinforcer association. This role is commonly described as "setting the occasion" for responses (Skinner, 1938; Bouton, 2016), but, lacking a mathematical model, it remains unclear how exactly stimuli participate in instrumental behavior (Mackintosh & Dickinson, 1979).

Current mathematical models neglect sequential phenomena and response rules

Current mathematical models of associative learning leave out phenomena that we know are important in learned behavior. In particular, second-order Pavlovian conditioning shows that a CS can function as an "unconditioned" stimulus, albeit one whose value is learned rather than inborn (Pearce, 2008; Bouton, 2016). Such a conditioned (learned) reinforcer is also effective on instrumental behavior. For example, if a sound is paired with food, a rat can learn to press a lever just to hear the sound (Skinner, 1938; McGreevy & Boakes, 2011). Conditioned reinforcement is well-characterized conceptually, and it is also an important tool to investigate associative learning (Rescorla, 1980; Williams, 1994a), but it is absent from mainstream mathematical models. This omission is noteworthy for several reasons. First, conditioned reinforcement appears fundamental in the acquisition of behavioral sequences (Skinner, 1938; Enquist et al., 2016), with applications to both clinical practice (Pierce & Cheney, 2013) and animal training (McGreevy & Boakes, 2011). Second, conditioned reinforcement straddles the Pavlovian-instrumental dichotomy, as it can affect instrumental learning while originating from Pavlovian contingencies (Frieman & Reilly, 2015). Modeling conditioned reinforcement may thus be helpful in understanding Pavlovian-instrumental interactions. Third, even experiments that are not meant to investigate conditioned reinforcers may effectively establish such reinforcers. For example, Skinner (1934) observed that the sound of a food delivery apparatus, which is incidentally paired with food in many experiments, can reinforce behavior such as approach and lever pressing even when no food is provided. In principle, any contextual or incidental stimulus may become a conditioned reinforcer capable of affecting learning. Thus, a mathematical model that includes conditioned reinforcement may yield insight into many learning situations.

Lastly, we note that theoretical efforts have focused on how associations are formed, but less on how they determine behavior. Few attempts have been made to progress beyond Rescorla and Wagner (1972)'s statement that a stronger association has a stronger influence on behavior. Even if behavioral responses are of interest mainly as indices of learning (Hall, 1994; Mackintosh, 1994), our inferences will be uncertain if learning processes are only loosely connected to behavior. When responding is assumed to depend on a single association, a linear transformation between associative strength and response rate can be adequate (Harris, 2011; Ghirlanda & Ibadullayev, 2015; Ghirlanda & Enquist, 2019). More elaborate proposals exist both for single (Harris, 2011; Honey et al., 2019) and multiple associations (Stout & Miller, 2007; Ghirlanda & Ibadullayev, 2015; Ghirlanda, 2018), but have not gained acceptance so far, and do not consider the interaction of Pavlovian and instrumental learning.

Augmenting psychological theory with temporal-difference learning

In this paper, we aim to present a new theoretical direction that may contribute to a more complete understanding of the remaining gray areas in associative learning. Our work is influenced by machine learning algorithms known as "temporal difference" (TD) models, which, similarly to animals, can learn to obtain rewards and avoid punishment by observing and interacting with an unknown environment. These models have drawn significant inspiration from animal learning theory (Sutton & Barto, 2018), but have so far exerted little influence over it. While TD models are influential in behavioral neuroscience (Dayan & Niv, 2008; Balleine et al., 2009), they have not been fully adapted to animal learning, and few tests have been conducted (see Sutton and Barto (1981), Ludvig et al. (2012), Enquist et al. (2016), and Comparison With Other TD Models below). Nevertheless, TD models may contribute to animal learning theory in at least two ways. First, they enable us to take into account explicitly the sequences of stimuli and responses which form the raw data for associative learning. For example, TD models can simulate the learning of lever pressing for food as a sequence that includes orienting toward and approaching the lever, pressing it, moving from the lever to the food location, and finally ingesting the food, even if only this last event is directly reinforcing. TD models can also learn about discriminative stimuli that arise at any point in the sequence. The analysis of behavioral sequences is a cornerstone of experimental psychology (Skinner, 1938; Mackintosh, 1983; Pierce & Cheney, 2013; Baum, 2017), yet it is poorly supported by current mathematical models, which mainly deal with the formation of single associations.

A second advantage of TD models is that, partly because of their historical connection to psychology, they are easy to understand in terms of familiar psychological concepts. They make current theory more precise, rather than replacing it with an entirely different framework. A-learning, in particular, is a mathematical synthesis of ideas about S-R and S-S associations, conditioned reinforcement, and value-based decision-making. At the same time, A-learning prompts us to rethink some traditional ideas about associative learning. To justify this effort, we report below novel computational and conceptual analyses of associative learning phenomena that are difficult to understand with current theory.

Theory

A-learning combines stimulus-response values, stimulus values, and value-based decision-making

A-learning has been introduced in Enquist et al. (2016) (focusing on behavioral sequence learning, violation of expectation, and genetic predispositions) and further analyzed in Lind (2018a) (focusing on planning), and Lind et al. (2019) (focusing on social learning). Here we focus on the model's conceptual structure with reference to associative learning theory. The elementary experience based on which the model learns is a triplet consisting of a stimulus s, a behavior b used in response to s, and the next stimulus s':

$$s \to b \to s' \tag{1}$$

Animals are assumed to learn from each experience through two learning processes that operate simultaneously. The first is **stimulus-response (S-R) value learning**, which uses the experience (1) to estimate the value of responding with b to s, written $v(s \to b)$. For example, a rat undergoing instrumental conditioning may experience sequences like:

$$\begin{aligned} &\text{lever} \to \text{press} \to \text{food} \\ &\text{lever} \to \text{rear} \to \text{no food} \end{aligned} \tag{2}$$

in which case the model will learn to assign a higher value to lever→press than to lever→rear. The equation governing such learning, and what constitutes "value" will be discussed below. The second learning process is **stimulus value learning**, which uses the experience $s \to b \to s'$ to estimate the value of stimulus s, written $w(s)$. For example, if s is a Pavlovian conditioned stimulus (CS) and s' the unconditioned stimulus (US), the model learns to attribute to s' a similar value as it attributes to s, as in the classic example

$$\text{bell} \to * \to \text{food} \tag{3}$$

where $*$ signifies that any behavior can intervene between bell and food. Stimulus value learning operates also in instrumental situations, in which case the value attributed to s can depend on how the animal behaves. For instance, in example (2), the lever stimulus will acquire value only if the animal actually presses the lever.

The two learning processes just introduced are linked by the assumption that stimulus values influence the learning of S-R values. This is best seen in the model's learning equations. The learning equation for stimulus values is:

$$\Delta w(s) = \alpha_w \left[u(s') + w(s') - w(s) \right] \tag{4}$$

where $\Delta w(s)$ is the change in $w(s)$ caused by experience (1), α_w is a positive learning rate, and $u(s')$ is the genetically determined *primary value* of s' (often referred to as innate or unconditioned value). For example, food ingestion will generally have positive u value, while pain a negative u value. Potentially, value can be attached to any stimulus, such as sexual and social stimuli (Curio et al., 1978; Mineka & Cook, 1988; Lind et al., 2018), and may depend on the animal's state. For example, female rats lever-press for paper strips only around parturition, when they need nest material (Oley and Slotnick (1970); see also Hauser and Gandelman (1985)).

The learning equation for the S-R value $v(s \to b)$ is very similar to Eq. 4:

$$\Delta v(s \to b) = \alpha_v \left[u(s') + w(s') - v(s \to b) \right] \tag{5}$$

where α_v is a second learning rate. The presence of $w(s')$ in this equation formalizes the assumption that learned (conditioned) stimulus values can reinforce behavior in the same way as primary value. A subtle difference between

Eqs. 4 and 5 is that the experience $s \rightarrow b \rightarrow s'$ leads to updating only the $v(s \rightarrow b)$ for the behavior b that is actually used. Thus, in example Eq. 2, the experience lever → press → food would update $v(\text{lever} \rightarrow \text{food})$ but not $v(\text{rear} \rightarrow \text{food})$. On the other hand, any choice of behavior leads to update the same $w(s)$. For example, $w(\text{lever})$ is updated after both lever→press→food and lever→rear→no food.

To select which behavior is used in response to a stimulus, A-learning adopts the softmax decision-making rule, according to which the probability of choosing b in response to s is:

$$\Pr(s \rightarrow b) = \frac{e^{\beta v(s \rightarrow b)}}{\sum_{b'} e^{\beta v(s \rightarrow b')}} \tag{6}$$

where the sum runs over all possible behaviors. According to this equation, behavior b is selected more often the higher the corresponding $v(s \rightarrow b)$ value. The parameter $\beta > 0$ regulates the trade-off between choosing the behavior with the highest estimated value and exploring other behaviors. With a high enough β, only the most valued behavior is likely to be chosen, while with lower β other behaviors are also tried out. We will see in the Results section that Eq. 6 is consistent with many findings, such as matching (Herrnstein, 1974; Baum, 1974) and behavioral contrast (Reynolds, 1961; Williams, 2002).

The learning Eqs. 4 and 5 are error-correction equations like the Rescorla and Wagner (1972) equation, whereby over successive experiences $w(s)$ and $v(s \rightarrow b)$ approach asymptotically the expected value of the following stimulus, including its primary and stimulus values. More precisely, it can be proved that, over many learning experiences in an environment with fixed statistical properties, $w(s)$ approaches:

$$w(s) \rightarrow \sum_{s'} \Pr(s') \left(u(s') + w(s')\right) \tag{7}$$

where $\Pr(s')$ is the probability that s' follows s, and the sum runs over all possible s' (Enquist et al., 2016; Sutton & Barto, 2018). Behaviors used in between stimuli do not appear explicitly in Eqs. 4 and 7, but they may nevertheless influence stimulus values by partly determining the sequence of experienced stimuli, such as in instrumental situations like example (2). Similarly to Eq. 7, $v(s \rightarrow b)$ approaches:

$$v(s \rightarrow b) \rightarrow \sum_{s'} \Pr(b \rightarrow s') \left(u(s') + w(s')\right)$$

where $\Pr(b, s')$ is the probability that b is chosen and then s' is experienced (Enquist et al., 2016; Sutton & Barto, 2018).

The sequential nature of learning

A crucial aspect of A-learning, shared with other TD models, is that stimulus value and S-R value learning can gather knowledge about the future value of stimuli and responses (Wiering, 2005; Enquist et al., 2016; Sutton & Barto, 2018). This enables the model to predict the occurrence of future reinforcement based on stimuli that are not immediately contiguous with it, and to learn extended sequences of actions in order to obtain reinforcement. As an example, consider the partial reinforcement-extinction effect (PREE, see Mackintosh (1974), Pearce (2008), and Bouton (2016)). This is the finding that responses that are reinforced only some of the time extinguish more slowly than responses that are reinforced every time, as shown in Fig. 1 based on a simulation with A-learning. This persistence seems paradoxical because responding is less valuable under partial reinforcement, but it may emerge as follows.

During learning, an animal trained under continuous reinforcement experiences sequences of stimuli of the form:

$$\ldots \rightarrow s \rightarrow s_+ \rightarrow s \rightarrow s_+ \rightarrow s \rightarrow s_+ \rightarrow s \rightarrow s_+ \rightarrow \ldots \tag{8}$$

where s represents all stimuli in the experimental situation, such as a Skinner box with a lever, and s_+ a reward, such as a food pellet. In between each s and s_+, a trained animal performs a response b, such as a lever press. We have not indicated responses in Eq. 8 to facilitate reasoning about stimulus values. Although only s_+ is rewarding, the sequence in (8) leads to increasing stimulus values for both $w(s)$ and $w(s_+)$. The growth of $w(s)$ is caused by the fact that s precedes s_+, while the growth of $w(s_+)$ is secondary to the growth of $w(s)$. Once $w(s) > 0$, in fact, Eq. 4 begins to attribute additional value to s_+ as a predictor of the now valuable s. In other words, s_+ becomes valuable because it predicts, via an intermediate step, the future occurrence of another s_+.

Under partial reinforcement, a third stimulus occurs, which we write s_0 and which represents the animal's perceptions after a non-rewarded response, such as the sound of the lever press without the sounds that accompany food delivery (see Pearce et al. (1997), for a similar suggestion). Thus, the stimulus sequence experienced during partial reinforcement is:

$$\begin{aligned} &\ldots \rightarrow s \rightarrow s_+ \rightarrow s \rightarrow s_0 \rightarrow s \rightarrow s_+ \rightarrow \\ &\ldots \rightarrow s \rightarrow s_0 \rightarrow s \rightarrow s_+ \rightarrow \ldots \end{aligned} \tag{9}$$

where we have assumed for illustration that every other response is rewarded (fixed-ratio 2 schedule). In this case, A-learning predicts that s_0 should acquire stimulus value, because it also predicts the future occurrence of

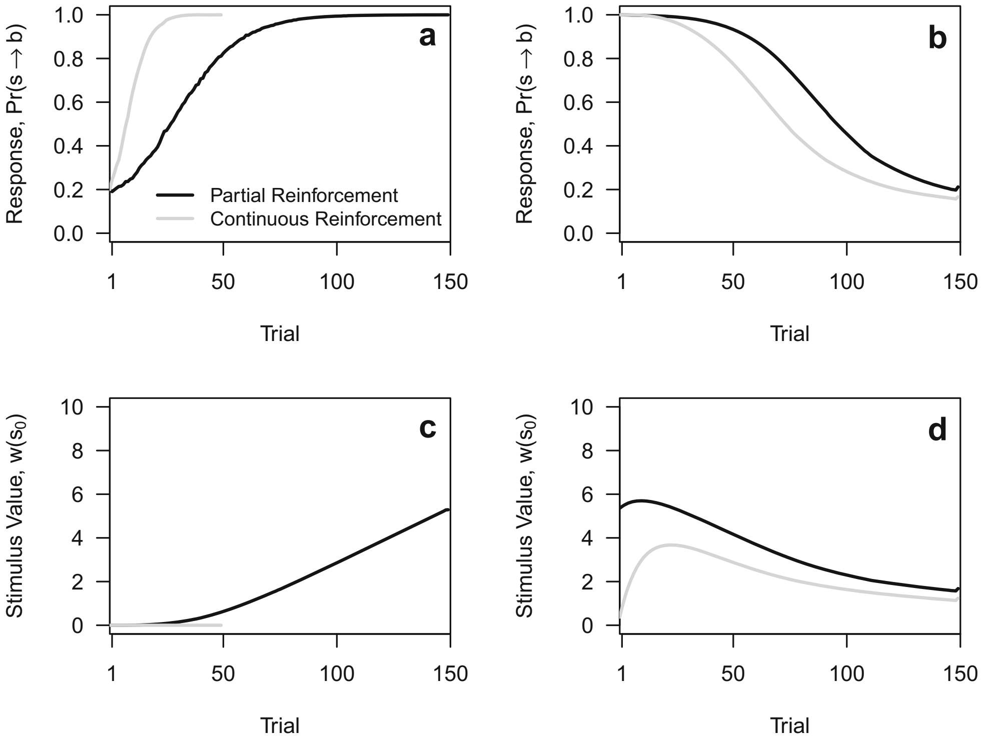

Fig. 1 Model simulation of the partial reinforcement-extinction effect (PREE) in instrumental learning. Partial reinforcement leads to both slower acquisition (**a**) and slower extinction (**b**). The latter effect derives from absence of reward (stimulus s_0, see text) acquiring stimulus value, under partial but not under continuous reinforcement (**c**). In extinction, $w(s_0)$ reverts to 0 under both partial and continuous reinforcement, because of the absence of primary reinforcement (**d**). A fixed-ratio 2 schedule is used for partial reinforcement, requiring two responses for each reward. Simulation script and model parameters are available online

reward (Fig. 1c). Crucially, s_0 is also the stimulus that follows responses during extinction. Thus, animals trained under partial reinforcement will perceive the extinction experiences $s \to b \to s_0$ as rewarding, until $w(s_0)$ is driven to zero by the absence of primary rewards in extinction (Fig. 1d). The rewarding effect of $w(s_0)$ causes $v(s \to b)$ to extinguish more slowly than for animals trained under continuous reinforcement, for which $w(s_0) = 0$ at all times (1). We refer to Capaldi (1971), Capaldi (1994), and Eisenberger (1992) for similar accounts and extended discussion; our aim is simply to use the PREE as an example of sequential effects in learning.

(During extinction, $w(s_0)$ is predicted to increase temporarily for the continuous reinforcement group, see Fig. 1c. The reason is that s_0 is followed by s, which had acquired stimulus value during training.)

Associative competition

Associative competition refers to the fact that learning about one stimulus may be affected by concurrent or previous learning about other stimuli, such as in overshadowing, blocking, and related phenomena (Pearce, 2008; Bouton, 2016). According to the landmark Rescorla and Wagner (1972) model, associative competition occurs because a US can condition only a finite amount of associative strength across all CSs. Thus, CSs trained concurrently will accrue a weaker association than if had been trained each on their own (overshadowing), and a CS will not gain associative strength when accompanied by other CSs that are already strongly associated with the US (blocking). This account can be interpreted in terms of predictiveness: because a US cannot be predicted more than perfectly, CSs compete for a finite amount of predictive power. A-learning embodies the same computational principles, which are readily expressed in terms of value: the total S-R value that a US can condition cannot exceed its own value (including learned stimulus value). Formally, we first assume that the v value of a compound stimulus is computed as the sum of the v values of its components:

$$v(s_1, s_2, \ldots, s_n \to b) = \sum_i^n v(s_i \to b) \tag{10}$$

where $s_1, s_2, \ldots, s_n$ denotes stimuli s_1 to s_n presented simultaneously. The same sum rule applies to u and w values. Next, we assume that an experience with a compound stimulus leads to learning about each of its components, according to:

$$\Delta v(s_i \to b) = \alpha_v \left[u(s') - v(s_1, s_2, \ldots, s_n \to b) \right] \tag{11}$$

and similarly for w values (see Enquist et al. (2016), for extensions to stimuli with different salience and to continuous stimulus dimensions). As anticipated, Eqs. 10 and 11 follow the same principles as the corresponding equations in the Rescorla and Wagner (1972) model, which A-learning adopts also for instrumental conditioning. As a consequence, A-learning covers the same phenomena of associative competition as the Rescorla and Wagner (1972) model, such as overshadowing, blocking, conditioned inhibition, and relative validity (Pearce, 2008; Bouton, 2016), and readily reproduces the observation that the same phenomena also arise in instrumental learning. The Rescorla and Wagner (1972) picture of associative competition is not perfect (Miller et al., 1995; Pearce & Bouton, 2001), which we leave to future work.

Comparison with current theory

The S-R values and stimulus values in A-learning can be compared with more traditional concepts. S-R values can be considered a mathematical formalization of S-R associations, as the following fundamental properties of S-R values are also true of S-R associations:

- S-R value $v(s \to b)$ is updated only when response b is used to stimulus s.
- A larger $v(s \to b)$ translates into a higher probability of performing b in response to s (via Eq. 6).[1]
- The magnitude of $v(s \to b)$ reflects the amount of reinforcement that is expected from responding to s with b (see text around Eq. 7).

The main differences between S-R values and traditional S-R associations are that S-R values are defined mathematically, both in what they represent (the expected value of the next stimulus), and in how they lead to behavioral decisions (through Eq. 6). Additionally, we have assumed that learned stimulus values (w) affect S-R values in the same way as primary values (u).

A-learning is not a simple S-R theory, however, because it also includes stimulus values. Above we introduced stimulus values as a mathematical formalization of conditioned reinforcement, but they also share properties with traditional S-S associations. Most importantly, stimulus values are learned based on S-S contingencies (Pavlovian contingencies) and reflect the value of forthcoming stimuli in a similar way as the strength of S-S associations reflects CS-US contingencies and US value. For example, in Pavlovian conditioning with only one CS and US, Eq. 7 reduces to

$$w(\mathrm{CS}) = pu(\mathrm{US}) \tag{12}$$

where p is the probability that the US follows the CS, and we have assumed for simplicity that the US has no conditioned value. Thus $w(\mathrm{CS})$ will be stronger for a US of a higher value, and for a consistently presented US (higher p).

The main difference between stimulus values and S-S associations is that stimulus values influence behavior indirectly, by influencing S-R values (Eq. 5) rather than by entering the decision function (Eq. 6). This leads to several unique predictions, as discussed below in Pavlovian Acquisition and Outcome Revaluation. Another difference between stimulus values and S-S associations is that each stimulus s has only one value, $w(s)$, whereas it could participate in many S-S associations. The latter allows a richer representation of the environment, but it also introduces theoretical complications. For example, under partial reinforcement a CS is followed by both reinforcing (US) and non-reinforcing stimuli (no-US, see previous section). Should a CS→no-US experience result in the decrease of the CS-US association, as in the Rescorla and Wagner (1972) model, or in the increase of a CS-no-US association, as considered by Konorski (1967) and Pearce and Hall (1980)? The argument holds even more strongly for different amounts of the same reinforcement: Does a food pellet remain the same stimulus if its size is doubled or halved, or is it treated as a new stimulus entering its own associations? A-learning avoids these difficulties because any stimulus that follows s affects, unambiguously, the single stimulus value $w(s)$.

In summary, A-learning's largest departures from current animal learning theory lie in giving more weight to S-R associations (in the form of S-R values) and in replacing S-S associations with stimulus values. We will show in Results that the interplay of S-R values and stimulus values generates a remarkably rich phenomenology.

Simulation software and statistical analyses

The simulations below were performed with lesim2, a learning simulator that implements several learning models and provides a scripting language for the specification of intra- and inter-trial events. The simulator is available at https://learningsimulator.org, while simulation scripts are available at https://osf.io/b8mez. Typical values of model parameters are displayed in Table 1.

[1] Early S-R theories considered a single stimulus-response link and were formulated directly in terms of response probability rather than S-R associations (Estes, 1950; Bush & Mosteller, 1951; Atkinson & Estes,). The latter have prevailed, however, because they extend more readily to situations with many stimuli (Rescorla & Wagner, 1972; Mackintosh, 1983).

Table 1 Typical values of simulation parameters. Exceptions are noted in the text. Exact values for all simulations are in the scripts available online. Behavior costs are subtracted to primary values and stimulus values in Eqs. 4 and 5

Parameter	Values
Number of subjects per simulation	100–500
Learning rate for stimulus values, α_w in Eq. 4	0.05, 0.1
Learning rate for S-R values, α_v in Eq. 5	0.05, 0.1
Decision making parameter, β in Eq. 6	1–1.5
Reinforcement value of positive stimuli (e.g., food)	5–10
Reinforcement value of negative stimuli (e.g., illness, shock)	−400
Cost of behavior (e.g., lever pressing, CR)	0–2

Results

In this section, we illustrate how Eqs. 4, 5, and 6 reproduce a diversity of Pavlovian and instrumental phenomena. In the impossibility of covering the whole field of associative learning, we have selected findings according to following criteria:

- Fundamental findings, such as Pavlovian and instrumental acquisition.
- Findings showing how A-learning may overcome the limits of traditional S-R theories even though it lacks S-S and R-S associations, such as the effect of non-contingent reinforcement on instrumental responding, and outcome revaluation.
- Findings that reveal surprising features of A-learning that, nevertheless, may agree with data. An example is A-learning's account of Pavlovian acquisition in terms of S-R associations shaped by genetic predispositions. Another example is A-learning's explanation of outcome revaluation as deriving from the action of stimulus values, rather than from S-S associations.
- Findings that challenge current theory because they reflect the action of multiple associations (or, in A-learning, multiple S-R values, and stimulus values). Examples are the acquisition and extinction of instrumental chains and Pavlovian higher-order conditioning, Pavlovian-to-instrumental transfer, and, again, outcome revaluation.

Our central message is that augmenting S-R learning with stimulus value learning and a well-defined decision function provides a transparent explanation for many major findings.

Acquisition of responses

Since the work of Konorski and Miller (1937) and Skinner (1937), psychologists have distinguished between learning about response-reinforcer contingencies (instrumental) vs. CS-US contingencies (Pavlovian). A-learning explores the hypothesis that both arise from the same principles, S-R value learning and stimulus value learning, with the only difference that Pavlovian learning is more strongly determined by genetic influences on learning and decision-making. This departure from current thinking has been motivated in part by a reanalysis of Pavlovian conditioning data by Ghirlanda and Enquist (2019), which we summarize in Pavlovian Acquisition below (see also Gallistel et al. 2004).

Instrumental acquisition

A-learning's account of simple instrumental learning is similar to classic theories such as the law of effect (Thorndike, 1911; Baum, 1973). Namely, experiences of the form $s \to b \to$reward increase the S-R value $v(s \to b)$, which, owing to the decision-making rule in Eq. 6, results in increased probability of choosing b in response to s (Fig. 2a). Once popular, this account fell out of favor due mainly to three observations (Hall, 2002). The first is that a pure S-R relationship would not contain any information about the reinforcer, which contrasts with observations that animals can be sensitive to changes in reinforcer value. This objection is covered in Outcome Revaluation below. A related objection is that animals react visibly to omissions of expected reinforcers, which would seem impossible based on S-R associations alone (Pearce, 2008). A learning equation such as Eq. 5, however, assumes that the animal can compute the difference between expected and actual value (Rescorla & Wagner, 1972), say $\delta = u(s') + w(s') - v(s \to b)$. This difference can be assumed to elicit specific behaviors, such as search of the missing reinforcer or aggression toward a conspecific (Dollard et al., 1939; McFarland, 1971; Mackintosh, 1974). For example, adding δ to the S-R value for search behavior would result in the decision function (6) selecting such behavior more often after the omission of a reinforcer (Enquist et al., 2016, see also). In other words, while A-learning does not "expect" a specific forthcoming stimulus s', it does "expect" a specific value for this stimulus.

The third objection to S-R accounts of instrumental learning regards the effects of "free" reinforcement, that is, reinforcement delivered independent of the animal's behavior. For example, Hammond (1980) trained rats to lever-press for water, and then observed a reduction of lever-pressing when rats began receiving free water as well. Because the delivery of free water does not change the lever→press association, this finding seems to defy an S-R account. However, free water can increase the association between the lever stimulus (which is still perceived when free water is delivered) and other behavior. In A-learning,

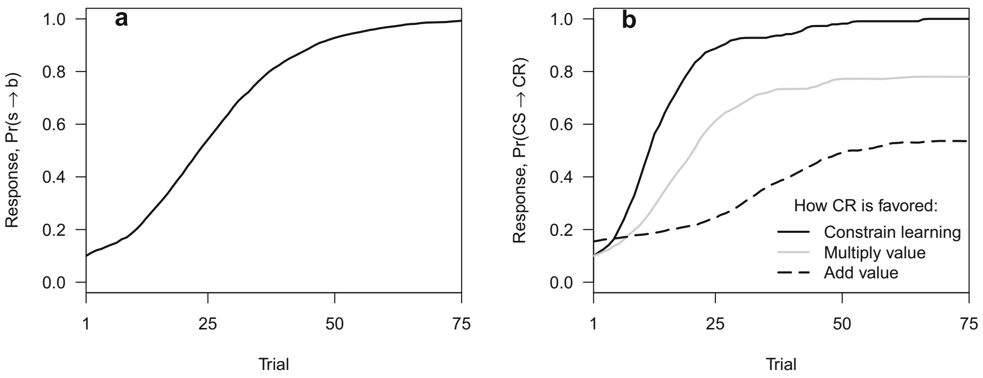

Fig. 2 Acquisition of instrumental and Pavlovian responses in A-learning. **a** Instrumental responses are learned by trial and error based on the growth of S-R values according to Eq. 5. **b** Pavlovian conditioned responses (CRs) are also learned according to Eq. 5, but which response is learned is determined more strongly by genetic predispositions. Three acquisition curves are shown, corresponding to genetic influences that constrain learning to the appropriate CR, that multiply the value of the CR, and that add to the value of the CR. See text around Eq. 14 for details. Simulation scripts and model parameters are available online

other behavior competes with lever pressing according to Eq. 6, which in this case yields

$$\Pr(\text{lever} \to \text{press}) = \frac{e^{\beta v(\text{lever}\to\text{press})}}{e^{\beta v(\text{lever}\to\text{press})} + e^{\beta v(\text{lever}\to\text{other})}} \quad (13)$$

where "other" describes all behavior different from lever-pressing. Writing $x = e^{\beta v(\text{lever}\to\text{press})}$ and $y = e^{\beta v(\text{lever}\to\text{other})}$ simplifies this equation into $x/(x+y)$, showing that an increase in the value of other behavior (y) leads to a decrease in the probability of lever pressing, even if the value of lever pressing remains unchanged.

Taking into account decision-making may also explain the results by Dickinson and Charnock (1985) shown in Fig. 3a. These authors trained two groups of rats to lever-press for food, and then introduced free food in two ways. In group Signaled, free food was always preceded by a white noise, while in group Random the noise and free food were uncorrelated. Lever pressing decreased in both groups, but the decrease was smaller in group Signal. Current theory offers the following account (Dickinson & Charnock, 1985; Pearce, 2008). First, free food is assumed to reduce lever pressing by strengthening a context→reinforcer association that interferes with a lever pressing→reinforcer association. Second, the noise stimulus is assumed to overshadow the context→reinforcer association, decreasing its interference with lever pressing. This explanation allows S-S associations (context-reinforcer) to compete with R-S associations (lever pressing-reinforcer), which is not entirely satisfactory because in other situations S-S associations have been assumed to potentiate R-S associations rather than compete with them (see Pavlovian-to-Instrumental Transfer).

A-learning offers a simpler account based on overshadowing between stimuli. Figure 3b shows our replication of Dickinson and Charkov's (1985) data. In addition, Fig. 3c and d show that $v(\text{lever} \to \text{press})$ attains the same value in both groups, but $v(\text{lever} \to \text{other})$ is lower in group Signaled. Hence, in group Signaled, lever pressing has less competition from other behavior when the decision rule in Eq. 6 is applied, leading to higher rates of lever pressing. The reason why $v(\text{lever} \to \text{other})$ is lower in group Signaled is that, in this group, the signal co-occurs with the lever on all trials, which leads to the signal overshadowing the lever as a cause of free food. In group Random, this overshadowing effect is smaller because the lever co-occurs with the signal only on half of the trials.

Pavlovian acquisition

We consider three theoretical issues around Pavlovian acquisition: what events trigger learning, what is learned, and what form the CR takes. A-learning's account of Pavlovian learning will be unfamiliar at first, and we ask the reader to keep an open mind. In current theory, Pavlovian acquisition has two properties. First, the CR primarily reflects the strength of a CS-US association. Second, the CS-US association is updated at every CS-US experience. Neither is true in A-learning. First, Pavlovian responses reflect the growth of S-R values, in the same way as instrumental responses. These values are made sensitive to CS-US contingencies, rather than to response-reinforcer contingencies, by an appropriate choice of model parameters. Second, S-R values are updated only when the CR occurs, rather than at every CS-US experience. In the remainder of this section we show how A-learning reproduces Pavlovian acquisition, including CR form, omission training, and autoshaping. Other Pavlovian phenomena are discussed later, such as outcome revaluation and higher-order conditioning.

Pavlovian acquisition as S-R value learning Empirical data and simulation results regarding the effect of signaled vs.

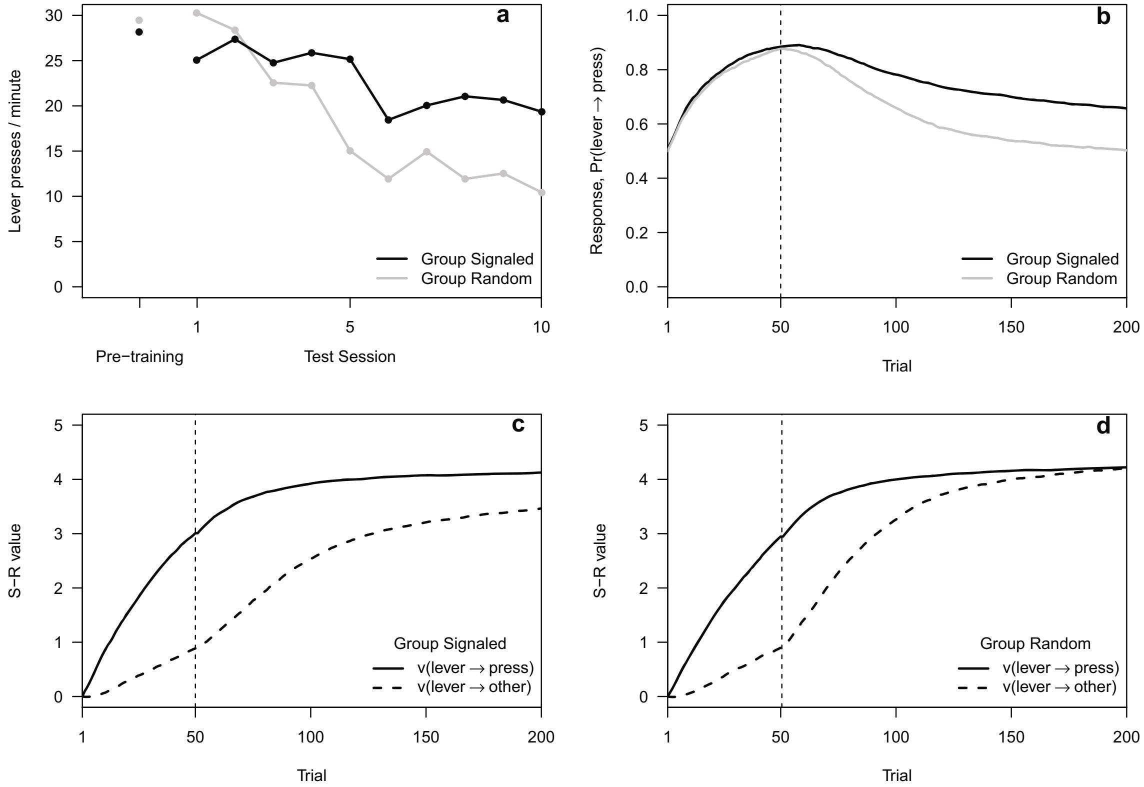

Fig. 3 Effect of signaled vs. unsignaled free reinforcement on instrumental responding. **a** Data from Dickinson and Charnock (1985), experiment 2, in which unsignaled free reinforcement was found to depress instrumental responding more than signaled free reinforcement (see text for details). **b** Model simulation of the same experiment. The *dotted line* indicates the introduction of free reinforcement. **c** Changes in S-R values v(lever→ press) and v(lever→ other) in group Signaled **d** Changes in the same S-R values in group Random. Simulation scripts and model parameters are available online

unsignaled free reinforcement on instrumental behavior. This is not sufficient in Pavlovian learning. Suppose, for simplicity, that only two responses to the CS are available: CR and no-CR. Because the US occurs regardless of behavior, it would reinforce equally CR and no-CR and lead to at most 50% CR frequency—yet data commonly show frequencies near 100%. Moreover, if $n > 2$ responses are possible, all would be reinforced, and the CR would attain a frequency of only $1/n$. We can avoid this incorrect result by setting model parameters appropriately. Several kinds of settings can produce a high frequency of CRs. For example, we can set different values of the learning rate α_v in Eq. 5 for CR and no-CR:

$$\begin{aligned}\alpha_v(\text{CR}) &> 0\\ \alpha_v(\text{no-CR}) &= 0\end{aligned} \tag{14}$$

This choice allows the US to reinforce only the CS→CR value, leading to a robust CR as shown in Fig. 2b. A similar outcome can be obtained by biasing decision-making rather than learning. For example, if $\beta(\text{no-CR}) = 1$ in Eq. 6 and $\beta(\text{CR}) > 1$, then the CR will be chosen more often and it will accrue more value (Fig. 2b). Decision making can also be biased by adding a fixed value to the CR, as suggested by Baum (1974) to account for bias in instrumental responding (see Eq. 18 and surrounding text). In summary, appropriate settings of model parameters enable the US to reinforce the CR primarily or exclusively, leading to high CR frequency even though the US follows other responses as well.

The course of CR acquisition Between the 1930's and 1970's, a consensus emerged that the acquisition of CS-US associations proceeds based on the experience of CS-US pairings, regardless of whether CRs occur. In A-learning, however, changes in the S-R value $v(\text{CS} \to \text{CR})$ occur only when a CR is performed. As a consequence, A-learning predicts larger gaps between CRs early in learning, when CR probability is low, compared to a theory in which CR probability increases with every trial. In Ghirlanda and Enquist (2019), we developed quantitative methods to distinguish between these two scenarios and applied them to acquisition data from pigeons, rabbits, and rats. The data were most compatible with the hypothesis that CR probability increases only when a CR is performed, rather than on every trial. We also showed that traditional arguments against this hypothesis (from response prevention studies, omission training, and sensory preconditioning) are not sufficient to reject it, either because the data is ambiguous or because of known conceptual flaws (Mackintosh, 1974; 1983). Overall, A-learning's prediction that CRs are necessary for Pavlovian learning, while unorthodox, appears compatible with available data (Ghirlanda & Enquist, 2019, for details, see). Note that the smooth acquisition curves graphed in the present paper represent average response probability (for example: Fig. 2b). In individual acquisition curves, response probability increases only on trials when a CR is performed.

CR form and behavior systems The parameter settings that yield Pavlovian conditioning in A-learning may appear ad hoc, yet they reflect the biological organization of learning (Domjan 1993; Timberlake 1983, 1994; Fanselow 1989, 1994; Shettleworth 1994), as revealed by the fact the form of the CR depends jointly on the CS and US in a manner that makes functional sense. For example, gastric distress (US) following food consumption (CS) regularly produces food aversion (CR), but not other possible responses such as startle or approach to the food location. Moreover, the CR need not be identical to the UR. For example, Holland (1977) found that, in rats, the CR to a light that signals food is inspection of the light (similar to the UR to food), while the CR to a tone CS is head jerking. These CRs appear functional, as head-jerking is useful to locate the source of a sound, while inspecting visual cues associated with food is conducive to finding food. These and similar findings are only loosely integrated with the view that Pavlovian conditioning depends primarily on a CS-US association. If this were the case, the CRs that a CS can elicit would be identical to the URs elicited by the US (Hall, 2002), or at least to some of them (Wagner (1981); but see Honey et al. (2019), for a recent proposal). In A-learning, on the other hand, we can structure model parameters to produce the appropriate CR to every CS-US pair. For example, to reproduce the results in Holland (1977) we can set:

$$\begin{aligned} \alpha_v(\text{tone, head-jerk, food}) &> 0 \\ \alpha_v(\text{tone, inspect, food}) &= 0 \\ \alpha_v(\text{light, head-jerk, food}) &= 0 \\ \alpha_v(\text{light, inspect, food}) &> 0 \end{aligned} \tag{15}$$

These settings prevent the growth of $v(\text{tone} \to \text{inspect})$ and $v(\text{light} \to \text{head-jerk})$ when the tone and light are followed by food, leading to the observed CR specificity (Fig. 4). We can account for constraints on instrumental learning in the same way, such that a reinforcer is most effective on behavior that is naturally related to that reinforcer (Shettleworth 1975, 1978; Domjan 1993; Roper 1983).

Omission Training and Autoshaping By structuring model parameters, A-learning can also exhibit Pavlovian phenomena that are believed to contradict S-R theories, such as omission training and autoshaping (Mackintosh, 1983; Pearce, 2008). In omission training, a CR can persist despite preventing a positive US. For example, a rat can persist in licking a tube through which sucrose is delivered, even if doing so cancels sucrose delivery (Patten and Rudy (1967); see Gormezano and Hiller (1972), Lucas (1975), and Locurto et al. (1976) for other examples). Omission training, however, can also be effective in decreasing or abolishing a CR (Locurto et al., 1976; Eldridge & Pear, 1987; Sanabria et al., 2006; Poling & Poling, 1978). This variability in outcomes is not well understood, partly because omission training pits Pavlovian and instrumental contingencies against each other, which current theory does not handle easily. In exploratory simulations with A-learning, we have produced various levels of CR maintenance. For example, the CS may acquire stimulus value on no-CR trials (ending with the US), which can reinforce the CR or behaviors that precede the CR, such as approach to the CS location (Lucas, 1975; Eldridge & Pear, 1987). Additionally, the US may have asymmetrical effects on different behaviors. For example, approach to a response key may be more easily reinforced by food than withdrawal from the key, hence the effect of non-reinforcement may be smaller than those of reinforcement. If these conditions do not hold, however, omission training is predicted to be effective and to lead to CR extinction. Some of these preliminary results are in Fig. 5, and may be the starting point for a future, more systematic investigation.

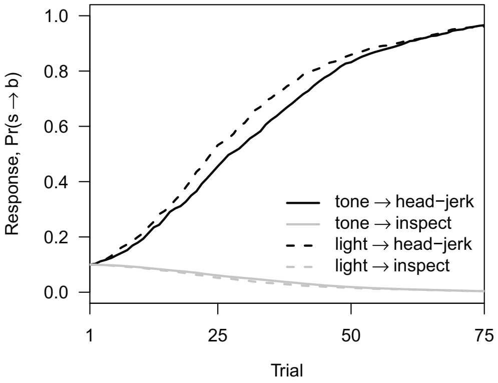

Fig. 4 CR form as a function of the CS-US pair in a replication of Holland (1977). In this experiment, a light and a tone predicted the same US (food pellets), but rats acquired different CRs to the two stimuli: head-jerk to the tone and inspection to the light. The graph shows this result in model simulations with the learning rates in Eq. 15. Simulation scripts and model parameters are available online

In autoshaping, a response to a CS develops even though it is not required to yield the US. For example, pigeons will come to peck a key that is lit just before food delivery (Brown & Jenkins, 1968), and rats will come to contact a lever that is introduced just before

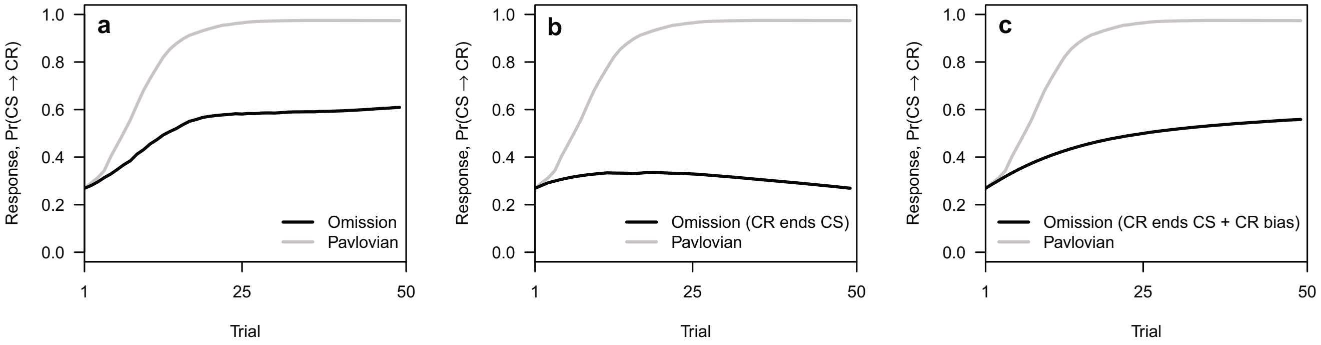

Fig. 5 Responding under omission training in A-learning, compared to responding under standard Pavlovian conditioning. In all cases, the α_v value for $v(\text{CS} \rightarrow \text{CR})$ is 0.1 and the α_v value for $v(\text{CS} \rightarrow \text{no-CR})$ is 0.01 (see text around Eq. 14). **a** CRs do not terminate the CS. **b** CRs terminate the CS. **c**. CRs terminate the CS, but the CR has a baseline higher probability of being selected, compared to alternative responses. Simulation scripts and model parameters are available online

food delivery (Boakes, 1977). Autoshaping is generally consistent with the behavior system approach as autoshaped CRs make functional sense: a pigeon typically benefits from pecking visual stimuli that precede the ingestion of food (these usually come from the food itself), and a rat from approaching them. Mechanistically, autoshaped CRs can arise in A-learning through the same kinds of predispositions discussed above in the cases of Pavlovian acquisition and omission training. For example, pecking can be selected more often than other responses if it is given a higher β value in Eq. 6, even if it has the same S-R value as other responses, similarly to what is shown in Fig. 2b, gray line. These remarks do not cover the extensive literature on omission training and autoshaping, but are encouraging as they emerge from a coherent pattern of setting model parameters to reflect genetic influences on learning (see also Enquist et al. 2016).

Avoidance responses

An avoidance response prevents or postpones a negative event, like a lever press that postpones an electric shock. Theories of avoidance learning must explain how a response can be reinforced by the absence of an event that would have otherwise occurred (see Herrnstein (1969) and Mackintosh (1983), for extended discussion). For this reason, avoidance responses have been sometimes considered as Pavlovian CRs that are automatically triggered by CS-US experiences (Bolles, 1970). For example, a stimulus that signals shock may cause increased activity in a rat, which may facilitate the accidental performance of the response that cancels the shock. At least two findings, however, prevent a purely Pavlovian account. First, animals can learn avoidance responses that do not resemble CRs to the US, such as lever pressing (although these may be are harder to learn; Mackintosh (1983)). Second, in most cases, much higher response rates are obtained when the response is actually instrumental in avoiding the negative consequence (Herrnstein, 1969). If only the Pavlovian contingency mattered, the same rate of avoidance CRs would occur regardless of whether they avoid the US. Partly for these reasons, two-factor theories of avoidance have been popular (Mowrer, 1960; Herrnstein, 1969; Mackintosh, 1983; Maia, 2010). In these theories, a stimulus signaling a negative consequence would first become a Pavlovian CS triggering a CR consisting of a negative subjective state, such as fear or anxiety. Then, the avoidance response would be reinforced instrumentally, because it removes the CS, and with it the negative state. However, avoidance responses can be learned even if they do not terminate a CS, at least not an overt one. For example, rats can learn to press a lever in order to postpone unsignaled shocks that would otherwise occur at regular intervals (Sidman, 1953).

A-learning can learn avoidance responses with no special assumptions. Figure 6a shows results from simulations in which an initially neutral stimulus, referred to as the CS and lasting two time steps, signals the delivery of a negative US. One response, referred to as the avoidance response, terminates the trial and cancels US delivery. Four other response are also available, which have no effect. The avoidance response is acquired not because it is directly reinforced, but because every other response is punished by the negative US (Bolles, 1970). This is sufficient to produce robust avoidance because A-learning's decision function, Eq. 6, can select with high probability a response with zero value (or even negative value), provided the alternatives have even lower values (Fig. 6b).

A-learning also predicts the finding that an avoidance response is more difficult to learn if it does not terminate the CS (Fig. 6c). The reason is that the CS acquires negative value because it predicts the US. Thus, an avoidance response that does not terminate the CS is punished by the continuation of the CS, even if not as much as the other responses, which are punished both by the CS and the US.

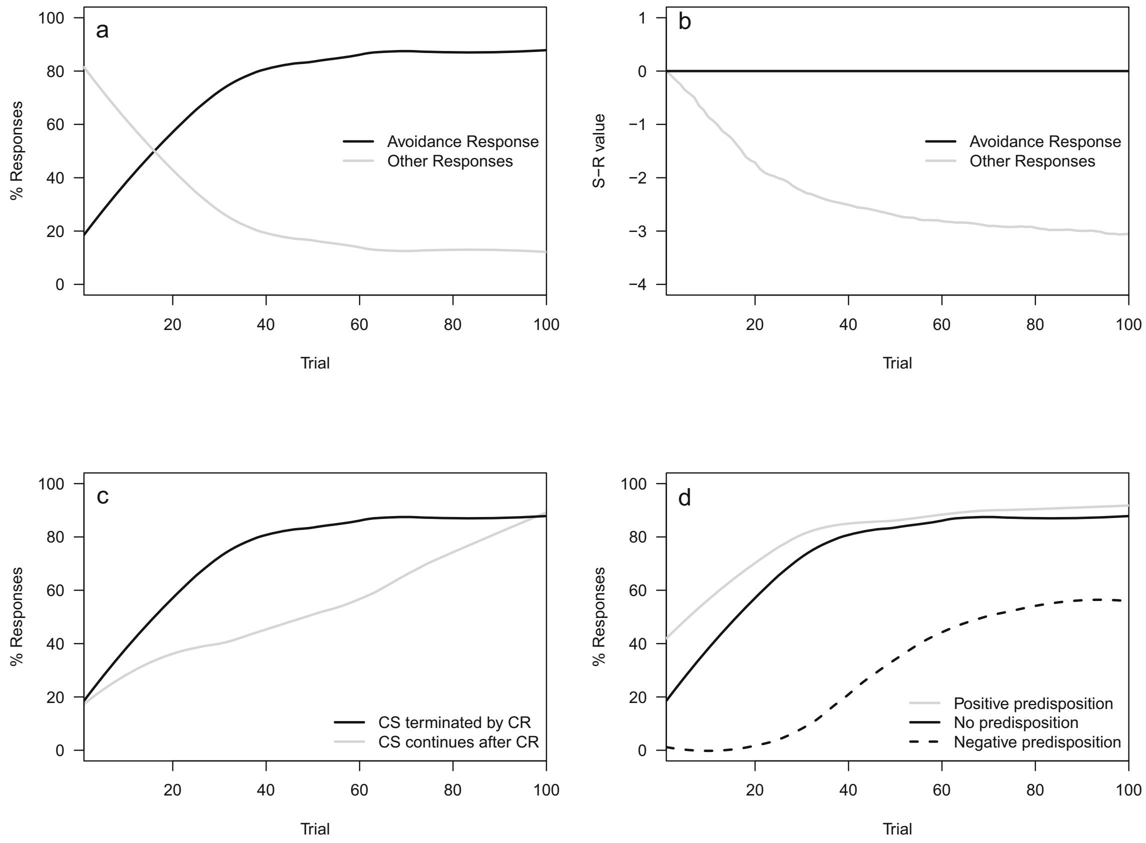

Fig. 6 Avoidance learning. **a** Acquisition of an avoidance response that terminates a trial, at the end of which a negative stimulus would have occurred. **b** S-R values underlying the results in the previous panel. **c** Comparison between learning a response that both cancels a negative stimulus and terminates a warning CS (*black line*, same simulation as panel **a**) and a response that cancels the negative stimulus but lets the warning CS continue for the entire scheduled duration (*gray line*). **d** Different speed of acquisition of an avoidance response that is genetically predisposed (*gray line*), one that is not predisposed (*black line*, same simulations as in **a** and **c**) and one that is opposed by genetic predispositions. In all simulations, a trial lasts three time steps, the first two presenting a warning CS and the last one a negative stimulus ($u = -10$). A positive genetic predisposition is modeled as adding a fixed value of 1 to the S-R value of the avoidance response. A negative predisposition subtracts a value of 5. Simulation scripts and model parameters are available online

Lastly, A-learning readily accommodates the finding that unconditioned responses can either promote or interfere with avoidance learning. The argument will be familiar to the reader by now: as discussed for Pavlovian and instrumental acquisition, stimulus-specific learning rates and baseline response probabilities may influence what responses are tried out in specific motivational states, and what responses can be learned quickly (Lind et al., 2019). If the avoidance response selected by the experimenter is among the predisposed ones, then avoidance will be easier to learn than if the experimenter had chosen a response that runs counter to the animal's predispositions (Bolles, 1970; Mackintosh, 1983). We show this effect in Fig. 6d, which compares avoidance responses that are either favored or opposed by a genetic predisposition with an avoidance response that is neither favored nor opposed.

While we do not claim to resolve all issues around avoidance learning with these preliminary investigations, the results are encouraging. The above mechanisms may also be complemented with an internal fear or anxiety motivational state, as suggested by two-factor theory (Mowrer (1960); see Maia (2010) for an implementation in terms of TD learning).

Matching and behavioral contrast

Our discussion of signaled and unsignaled free reinforcement and of avoidance learning highlights that a model of decision-making may be as necessary as a model of learning for a satisfactory theory. The decision-making rule in A-learning, Eq. 6, also generates two major phenomena: matching and behavioral contrast. While extensive empirical knowledge exists around these phenomena (Herrnstein, 1974; Baum, 1974; Reynolds, 1961; Williams, 2002), they are outside the scope of current theory because they depend on establishing a precise link between learning variables and responding.

Matching refers to choosing behavior in proportion to the amount of reinforcement that is gained from it (Herrnstein, 1974). Formally, the most general form of the matching law can be written as

$$\frac{\Pr(s \to b_1)}{\Pr(s \to b_2)} = k \left(\frac{r_1}{r_2}\right)^a \tag{16}$$

where r_i is the amount of reinforcement obtained from behavior b_i, while a and k are positive parameters (Baum, 1974). Parameter a describes whether the individual matches behavior to reinforcement perfectly ($a = 1$) or whether it displays a preference for the lower-yielding activity (undermatching, $a < 1$) or for the higher-yielding one (overmatching, $a > 1$). Parameter k describes whether the individual has a reinforcement-independent bias for b_1 or b_2 ($k > 1$ or $k < 1$, respectively), or whether it has no bias ($k = 1$). Eq. 6 implies immediately the matching law in the absence of bias, since it leads to:

$$\frac{\Pr(s \to b_1)}{\Pr(s \to b_2)} = \left(\frac{e^{v(s\to b_1)}}{e^{v(s\to b_2)}}\right)^\beta \tag{17}$$

which is identical to Eq. 16 with $k = 1$, $a = \beta$, and $r_i = e^{v(s\to b_1)}$, the latter meaning that theoretical values should measure reinforcement on a logarithmic scale. The case $k \neq 1$ is not covered explicitly by Eq. 6, but it can be taken into account as indicated by Baum (1974), that is by introducing aspects of value that are intrinsic to each activity rather than learned. In fact, if v_i is the intrinsic value of b_i, such that its total value in situation s is $v_i + v(s \to b_i)$, Eq. 6 yields:

$$\frac{\Pr(s \to b_1)}{\Pr(s \to b_2)} = \left(\frac{e^{v_1}}{e^{v_2}}\right)^\beta \left(\frac{e^{v(s\to B_1)}}{e^{v(s\to b_2)}}\right)^\beta \tag{18}$$

so that the bias term is expressed as $k = e^{\beta(v_1-v_2)}$. Furthermore, Williams (1994b) shows that the matching law can encompass several dimensions of reinforcement in addition to its magnitude, namely the probability of reinforcement and the delay between the behavior and reinforcement. These effects arise naturally in A-learning because S-R values reflect the average reward resulting from a behavior. This property of S-R values takes into account the magnitude and probability of reward by definition, and can also take into account delay because the effect of delay is to reduce value. That is, of two behavioral options resulting in the same reward, but with different delays, A-learning will learn to prefer the one with the shortest delay because this leads to a higher reward rate, provided that A-learning perceives a cost for time that passes without a reward.

Equation 6 can also reproduce behavioral contrast phenomena, i.e., changes in the probability of a behavior due to changes in the probability of other behaviors (Reynolds, 1961; Williams, 2002). In fact, Eq. 6 implies that $\Pr(s \to b)$ will change in the opposite direction as any of the $\Pr(s \to b')$ that appear at the denominator of the fraction ($b' \neq b$). This results in at least two kinds of contrast effects. First, a behavior will increase (decrease) in probability whenever other behaviors that are possible in the same situation become less (more) probable. For example, in a Skinner box with two levers, the probability of pressing lever 1 can be written as

$$\Pr(\text{box} \to \text{press 1}) = \frac{e^{\beta v(\text{box}\to\text{press 1})}}{e^{\beta v(\text{box}\to\text{press 1})} + e^{\beta v(\text{box}\to\text{press 2})}} \tag{19}$$

where, for simplicity, we have omitted other behaviors. Thus, a change in the value of pressing lever 2 leads to an opposite change in the probability of pressing lever 1 (note the similarity with our account of the effects of free reinforcement, Eq. 13, and of avoidance learning).

A second kind of behavioral contrast emerges when the probability of responding to a stimulus changes because of a change in the probability of responding to a different stimulus (Reynolds, 1961; Williams, 2002). Suppose, for example, that an animal is initially rewarded equally for behavior b_1 in response to s_1, and for behavior b_2 in response to s_2. After equal responding to both stimuli is established, the reward for responding to s_2 is withdrawn. As a consequence, responding to s_1 frequently increases. This effect arises within A-learning because extinguishing responding to s_2 also extinguishes behavior toward contextual stimuli that influence responding to s_1. That is, the two stimulus situations can be conceptualized as s_1, c and s_2, c, with c summarizing common stimuli. Because A-learning assumes that $v(s_i, c \to b_j) = v(s_i \to b_j) + v(c \to b_j)$, the probability of responding to s_1, c is:

$$\Pr(s_1, c \to b_1) = \frac{e^{\beta v(s_1\to b_1)+\beta v(c\to b_1)}}{e^{\beta v(s_1\to b_1)+\beta v(c\to b_1)} + e^{\beta v(s_2\to b_2)+\beta v(c\to b_2)}} \tag{20}$$

where, again, we have included only behaviors b_1 and b_2 for simplicity. In addition to lowering $v(s_2 \to b_2)$, unrewarded presentations of s_2, c lower $v(c \to b_2)$, which appears at the denominator of Eq. 20 and thus increases $\Pr(s_1, c \to b_1)$. This effect is demonstrated in Fig. 7.

Behavioral sequences

In both instrumental and Pavlovian procedures, animals can learn about extended sequences of events. In instrumental settings, animals can learn chains of actions to obtain rewards or avoid punishment (Skinner, 1938; Mackintosh, 1983), and in Pavlovian settings they can learn to perform a CR in response to a CS that predicts other CSs, rather

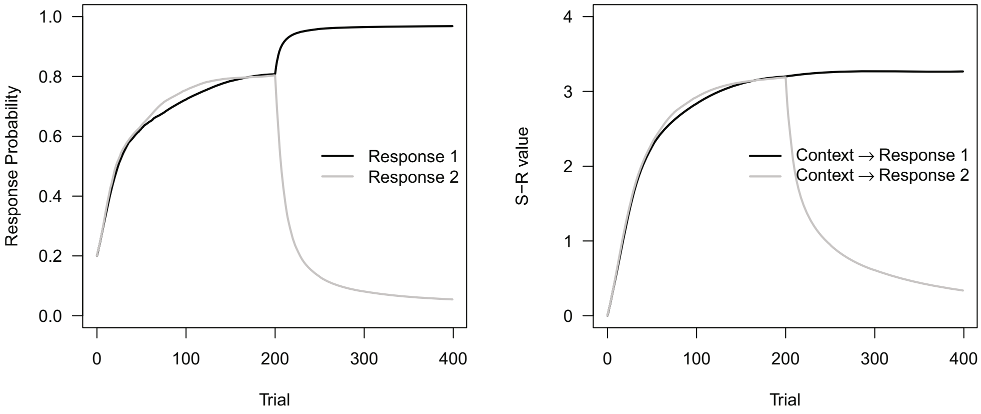

Fig. 7 Behavioral contrast in A-learning. **a** Responses b_1 and b_2 to stimuli s_1 and s_2, respectively, are initially equally rewarded in the same context c. After trial 200, the reward to s_2 is withdrawn and, as a consequence, responding to s_1 increases. **b** The effect is mediated by decrease in the S-R value $v(c \to b_2)$. The magnitude of the effect depends on the salience of the context, which is modeled as a learning rate for S-R values. In the figure, the learning rate for c is higher than that for s_1 and s_2. Simulation scripts and model parameters are available online

than directly the US (Pavlov, 1927; Rescorla, 1980). In this section we show how A-learning reproduces instrumental and Pavlovian sequential learning, noting that stimulus value learning plays a crucial role in both.

Acquisition of instrumental chains

We considered instrumental chains in some detail in Enquist et al. (2016), for example, replicating field observations by Inoue-Nakamura and Matsuzawa (1997) regarding learning of tool use in chimpanzees and “misbehavior” phenomena described by Breland and Breland (1961). How A-learning learns instrumental chains can be summarized as follows. Suppose the sequence to be learned is:

$$s_1 \to b_1 \to s_2 \to b_2 \to \text{reward}$$

meaning that response b_1 to stimulus s_1 brings about stimulus s_2, and response b_2 to s_2 brings about a reward. Learning of this chain by A-learning is illustrated in Fig. 8a, with the second response (more proximal to the reward) being learned more rapidly. The reason for the delayed learning of the first response is that s_2 is not initially reinforcing, and therefore even the correct performance of b_1 does not increase the S-R value $v(s_1 \to b_1)$. However, every time the chain is performed correctly (which initially occurs by chance), s_2 accrues stimulus value, $w(s_2)$, because it is followed by reward (Fig. 8c). Once $w(s_2) > 0$, performing b_1 in response to s_1 is rewarding and $v(s_1 \to b_2)$ starts to grow. This account is conceptually identical to the one by Skinner (1938), with stimulus value as a mathematical model of conditioned reinforcement. It is consistent with many findings about instrumental chains, such as that the speed of learning can be improved by teaching the chain backwards (backward chaining), by adding temporary rewards at intermediate steps (forward chaining), or by adding conditioned reinforcers at intermediate steps (e.g., clicker training; see McGreevy and Boakes (2011), for a survey of animal training techniques, and Enquist et al. (2016), for model analysis).

Acquisition of Pavlovian higher-order conditioning

Let us now consider the Pavlovian analogue of instrumental chains, that is, second- and higher-order conditioning. In a typical second-order conditioning experiment, the animal is first exposed to $\text{CS}_1 \to \text{US}$ experiences, and then to $\text{CS}_2 \to \text{CS}_1$ experiences. During the latter, a CR to CS_2 develops, which eventually extinguishes in the absence of further US presentations. Figure 8b shows that A-learning reproduces this typical finding. Figure 8d shows the growth of the stimulus value $w(\text{CS}_1)$ during the first phase of the experiment, when CS_1 predicts the US, and its extinction during the second phase, in which the US no longer appears. This dynamics of $w(\text{CS}_1)$ is what drives the acquisition and eventual extinction of a CR to CS_2 during the second phase. In this simulation, we have used assumption Eq. 14 to avoid learning of behavior other than the CR. Different CRs to CS_1 and CS_2 can be explained as above by setting learning rates appropriately.

Extinction of instrumental chains

The extinction of learned behavior (Pavlovian or instrumental) is one of the most researched topics in associative learning because of its clinical relevance (Bouton et al., 2012) and theoretical interest. The latter partly stems from the fact that early applications of the Rescorla and Wagner (1972) model characterized extinction as the erasure of learned associations, and were thus unable to explain why ostensibly extinguished behavior can reappear (see Rescorla, 2002;

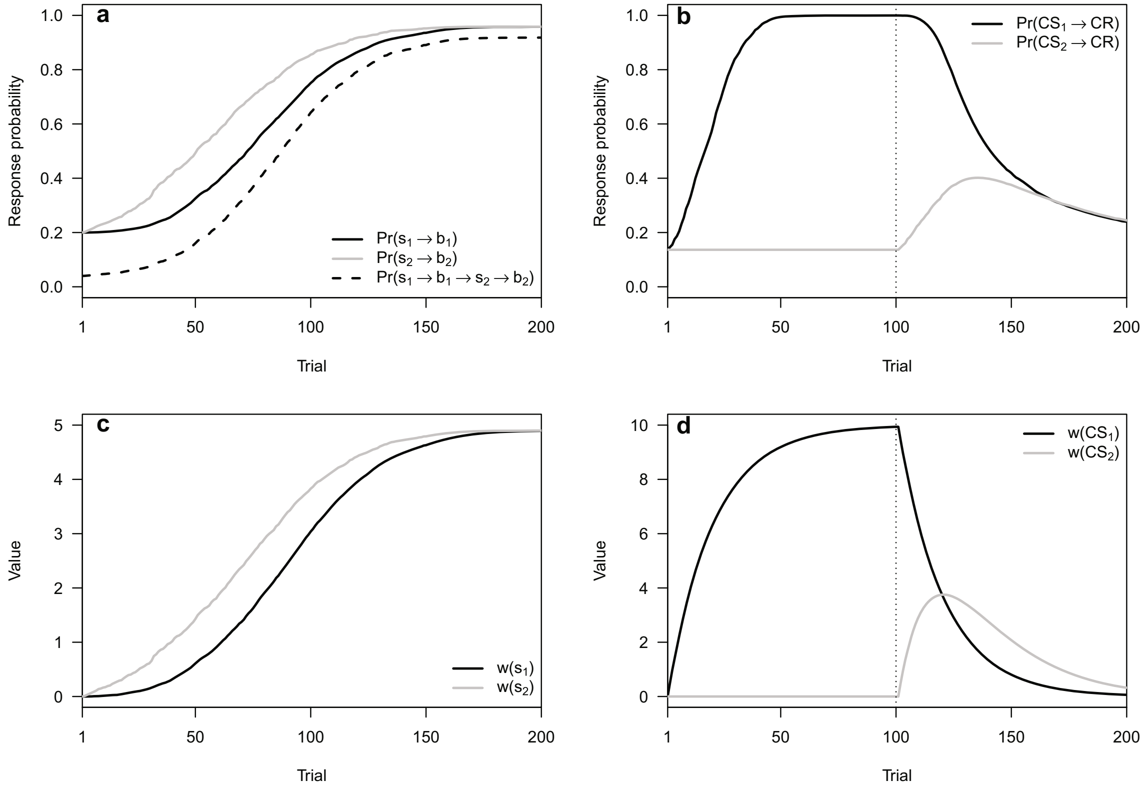

Fig. 8 Simulations of instrumental chaining and Pavlovian second-order conditioning. In instrumental chaining **a**, rewards are earned when completing a chain of two behaviors. In Pavlovian second-order conditioning **b**, experiences of $\mathrm{CS}_1 \rightarrow \mathrm{US}$ are followed by experiences of $\mathrm{CS}_2 \rightarrow \mathrm{CS}_1$. Panels **c** and **d** show the stimulus values changes that occur during training. Simulation scripts and model parameters are available online

Bouton et al., 2012, for details). However, Delamater and Westbrook (2014) show that the reappearance of extinguished behavior is consistent with the Rescorla and Wagner (1972) model, provided one allows learning about contextual stimuli that accompany nominal cues. Because these results depend on how the Rescorla and Wagner (1972) conceptualizes associative competition, they carry over to A-learning (see Associative competition) and we will not consider them here. Rather, we turn to the extinction of behavioral chains. This topic has received relatively little attention (Thrailkill & Bouton 2015, 2016), and early notes by Skinner (1934) are still of interest. He considered chains of the form:

$$s_1 \rightarrow b_1 \rightarrow s_2 \rightarrow b_2 \rightarrow \text{reward} \tag{21}$$

and argued that extinguishing the first link of the chain, through experiences of $s_1 \rightarrow b_1 \rightarrow$ no reward, would leave the second link unaffected. He supported this claim by observing that the sound of a food magazine (s_2) continues to elicit approach to the magazine (b_2) even after extinguishing the behavior that used to cause the sound (b_1; for example, lever pressing). Skinner also observed that an extinguished first link can be reacquired by experiences of $s_1 \rightarrow b_1 \rightarrow s_2$ (without reward), which he attributed to s_2 having become a conditioned reinforcer during chain acquisition. In the absence of further reward, both the reinforcing value of s_2 and the response $s_1 \rightarrow b_1$ would eventually extinguish. From these observations, Skinner (1934) formulated the principle that “in a chain of reflexes not ultimately reinforced only the members actually elicited undergo extinction.” This principle is consistent with A-learning, in which S-R values $v(s \rightarrow b)$ are modified only when $s \rightarrow b$ is performed, while stimulus values are updated at every experience. Figure 9 shows that A-learning readily reproduces Skinner’s observations (panels a and b), driven by the dynamics of $w(s_1)$ and $w(s_2)$ (panels c and d).

In seeming violation of Skinner’s principle, Thrailkill and Bouton (2015, 2016) have recently reported that extinction of one link of the chain can depress responding in the other, which they interpreted in terms of response-response associations. A-learning suggests an interpretation compatible with Skinner’s analysis. In these experiments, rats were trained to perform a two-link chain that can be described as

$$s_1, m_1, m_2 \rightarrow b_1 \rightarrow s_2, m_1, m_2 \rightarrow b_2 \rightarrow \text{reward}$$

where stimuli m_1 and m_2 represent the instrumental manipulanda (a chain and a lever), while s_1 and s_2 were two

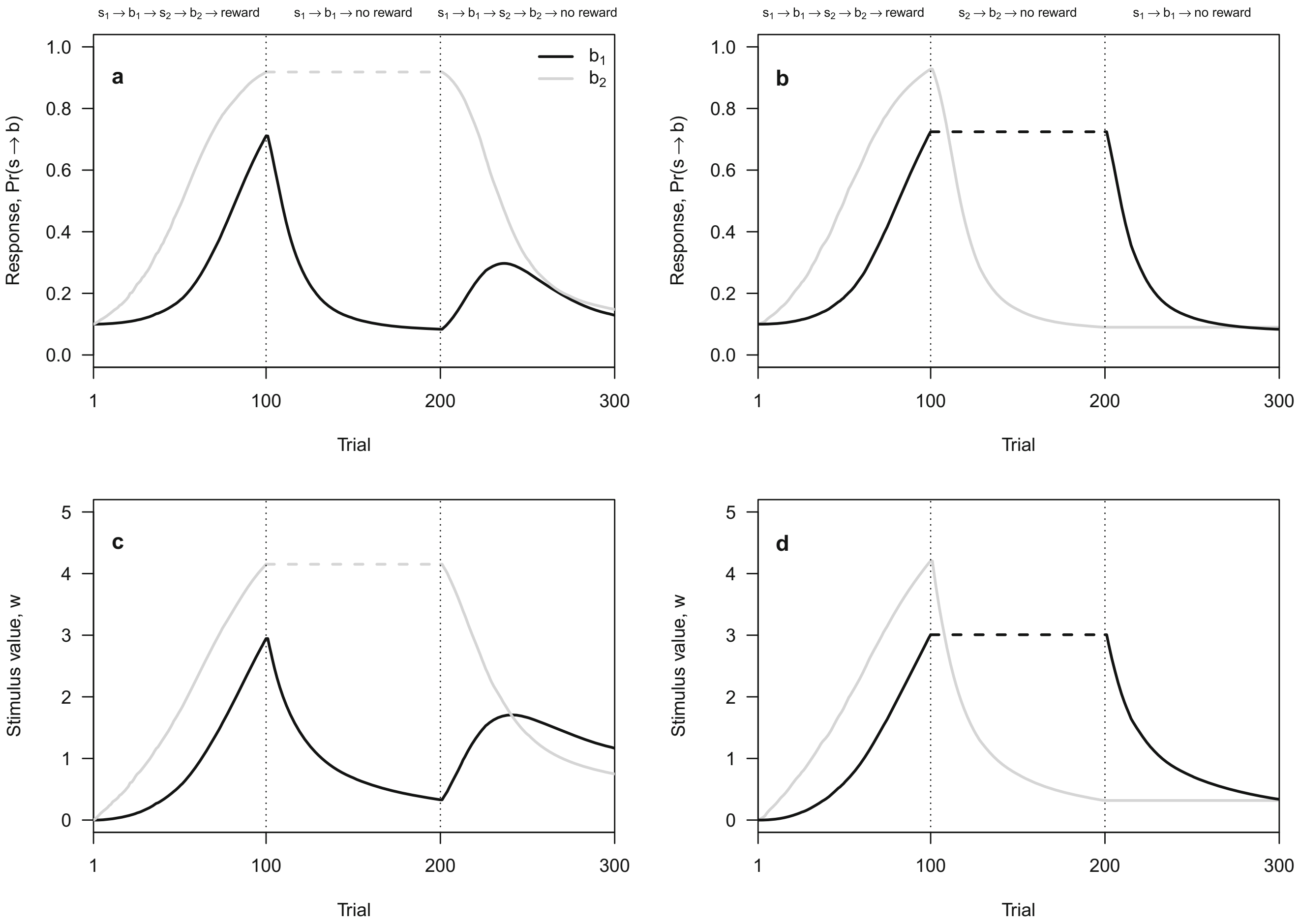

Fig. 9 Simulation of Skinner's (1934) observations about the extinction of behavioral chains. **a** Chain acquisition followed by extinction of the first link ($s_1 \rightarrow b_1 \rightarrow$ no reward), and then by temporary re-acquisition through experiences of the form $s_1 \rightarrow b_1 \rightarrow s_2$. **b** Chain acquisition followed by extinction of the second link ($s_2 \rightarrow b_2 \rightarrow$no reward) and then by extinction of the first link. **c**, **d** Corresponding stimulus values. *Dotted lines* indicate absent stimuli. Simulation scripts and model parameters are available online

lights differentiating the two links. Discrimination between s_1, m_1, m_2 and s_2, m_1, m_2 was not perfect, so that animals would occasionally perform b_2 during extinction of the first link, and b_1 during extinction of the second link. According to A-learning, these experiences would lower $v(m_2 \rightarrow b_2)$ and $v(m_1 \rightarrow b_1)$, respectively, and thus depress responding also in the link that, nominally, did not undergo extinction. This interpretation is supported by the fact that extinguishing one link had no effect on the other when only the manipulandum for the extinguished link was present during extinction, which would prevent the extinction of $v(m_2 \rightarrow b_2)$ when extinguishing the first link, and of $v(b_1 \rightarrow m_1)$ when extinguishing the second.

Extinction of Pavlovian higher-order conditioning

Classic results by Rizley and Rescorla (1972), Holland and Rescorla (1975b), and Holland and Rescorla (1975a) indicate that the extinction of Pavlovian second-order conditioning is similar to that of intermediate responses in instrumental chains. That is, extinguishing the first-order CS has typically little effect on responding to the second-order CS. A-learning reproduces this result in the same way as in the instrumental case. Consider a Pavlovian response CR_2 that is established by CS_1 →US pairings followed by CS_2 →CS_1 pairings. Extinction of the first-order CS through CS_1→no-US pairings modifies the S-R value $v(CS_1 \rightarrow CR)$ and the stimulus value $w(CS_1)$, but not the S-R value $v(CS_2 \rightarrow CR_2)$ that determines CR_2. Hence CR_2 should be unaffected by extinction of CS_1. Some studies, however, have reported a change in responding to CS_2 following the extinction of CS_1 (Rashotte et al., 1977). In A-learning, this variability has similar causes as variability in outcome revaluation studies (extinction can be considered a mild revaluation treatment). For this reason, we discuss variability in results at the end of the next section.

Table 2 Design of a typical revaluation experiment

Phase	Experiences								
	s_1	→	b_1	→	s_2	→	b_2	→	s_{outcome}
1. Acquisition	lever	→	press	→	pellet	→	eat	→	nutrients
2. Revaluation					pellet	→	eat	→	nutrients+illness
3. Extinction	lever	→	press	→					nothing
4. Reacquisition	lever	→	press	→	pellet	→	eat	→	nutrients

Outcome revaluation

Observations that animals can be sensitive to changes in the value of outcomes are central to contemporary understanding of associative learning, as they firmly reject simple S-R theories of both instrumental and Pavlovian learning (Rescorla, 1980; Hall, 2002; Holland, 2008; Balleine, 2011). Table 2 shows the typical design of an outcome revaluation experiment featuring an instrumental response (Pavlovian revaluation experiments are similar and not discussed here for brevity; see online materials for a simulation of Pavlovian revaluation). The first phase (acquisition) features a short sequence of actions, such as:

$$\text{lever} \rightarrow \text{press} \rightarrow \text{pellet} \rightarrow \text{eat} \rightarrow \text{nutrients} \qquad (22)$$

or, more generally:

$$s_1 \rightarrow b_1 \rightarrow s_2 \rightarrow b_2 \rightarrow \text{outcome} \qquad (23)$$

Different parts of this sequence enter the successive phases as indicated in Table 2. The revaluation phase (phase two) exposes the animals to a changed value of the outcome. It is typical to induce illness following food consumption, although other arrangements are possible.[2] In phase three (extinction), the response is tested again without yielding the outcome. Simple S-R theories predict that the response would be unchanged (compared to control animals that did not undergo revaluation), because the changed value of s_2 cannot affect the s_1-b_1 association if s_2 is not experienced. Thus, a change in responding would show that the animals learned something different, or something more than a S-R association (Miller, 1935; Rozeboom, 1958). Lastly, phase four (reacquisition) tests whether the response can be reacquired after having been extinguished in phase three. Intriguingly, the results of revaluation experiments are sometimes compatible with S-R theories and sometimes not (Rescorla, 1980; Mackintosh, 1983; Holland, 2008; Balleine, 2011). We use simulations with A-learning to discuss these results.

No revaluation in extinction

Figure 10 shows results from a first simulation, in which we find no difference during extinction between the experimental and control groups. During re-acquisition, however, the control group re-acquires lever pressing fully, while responding in the experimental group continues to decline. These results are as suggested by S-R theories, and mirror closely those of Adams (1980). To understand them, we consider how lever pressing is affected by the stimulus value of the pellet, w(pellet). Figure 10b shows that the revaluation treatment (a single pairing of the pellet with illness) causes w(pellet) to drop sharply, in the experimental group only. This drop does not influence lever pressing directly, because it does not change $v(\text{lever} \rightarrow \text{press})$. As the pellet is also absent during the extinction phase, the experimental and control groups respond identically throughout this phase. When the pellet is reintroduced during reacquisition, w(pellet) reinforces lever pressing in the control group, but punishes it in the experimental group, leading $v(\text{lever} \rightarrow \text{press})$ to increase in the former and further decrease in the latter (Fig. 10c and d).

Revaluation in extinction

Contrary to what found by Adams (1980), other revaluations experiments found differences between experimental and control animals already in the extinction phase (Holland, 2008), in both instrumental (Adams & Dickinson, 1981) and Pavlovian procedures (Rescorla, 1973; Holland & Rescorla, 1975a; Rescorla, 1980). This is a crucial result of revaluation experiments because it is incompatible with a simple S-R theory. These differences, however, are typically smaller and more variable than those seen during reacquisition (Dickinson et al., 2002; Holland, 2008). In A-learning, differences in extinction can arise if the changed stimulus value of the outcome has the opportunity to influence the S-R value for the response that produces the outcome. This may happen in several ways. A very direct way consists of adding experiences of the form lever→press→pellet between the revaluation and reacquisition phases, thus allowing the changed stimulus value w(pellet) to influence responding (Fig. 10b, e). This

[2] It is also common to devalue the food temporarily by sating the animal. Currently, this treatment must be hand-coded in simulations, but could be included in A-learning in several ways. For example, Zhang et al. (2009) propose to encode motivational state in the β parameter in Eq. 6, so that responses that are more (less) likely in a given motivational state are given a higher (lower) β value.

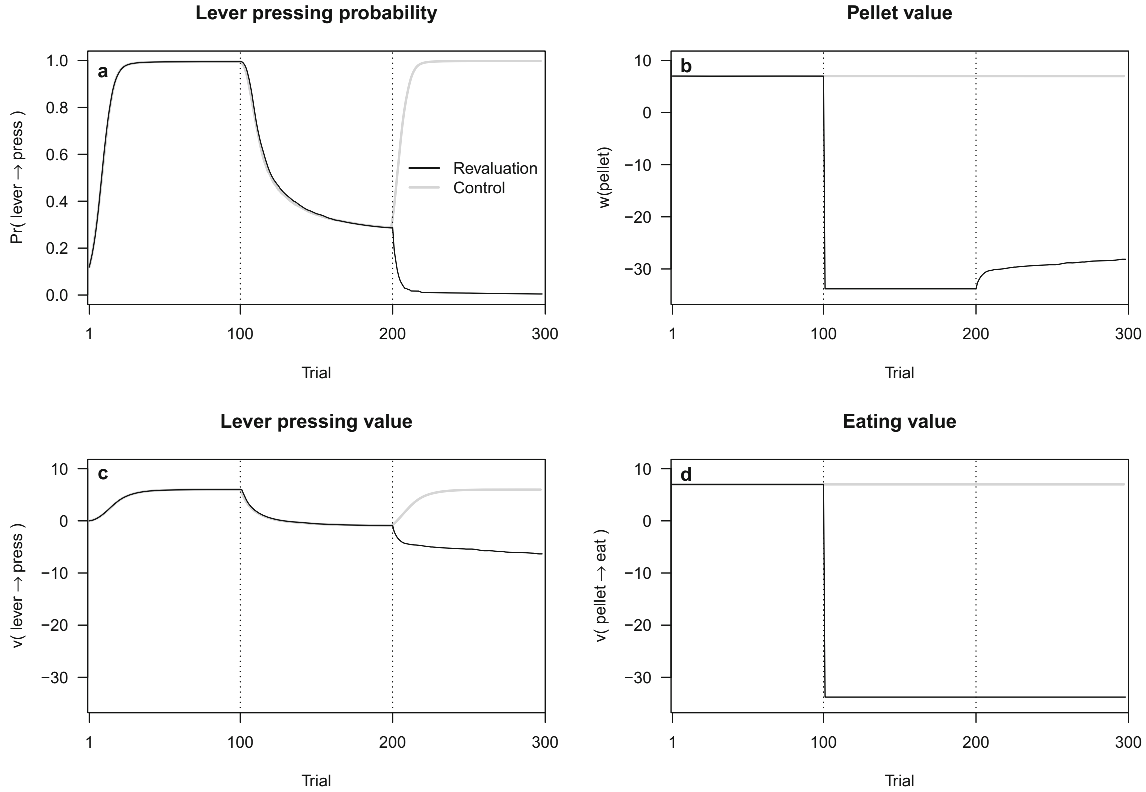

Fig. 10 Simulations of the revaluation experiment described in Table 2. **a** Acquisition, extinction, and reacquisition of lever pressing. **b** Pellet stimulus value throughout the experiment. **c** S-R value for lever→press. **d** S-R value for pellet→eat. The pellet is assumed to be familiar food, such that w(pellet) and v(pellet → eat) are both initially high. Simulation scripts and model parameters are available online

is seldom done, however, because it hardly distinguishes between alternative theories.

Another way to obtain extinction effects in A-learning is to include stimulus elements that appear in both the revaluation and extinction phases. If X is such a stimulus, the revaluation and extinction phases become:

$$\begin{array}{ll} \text{pellet} + \text{X} \to \text{eat} \to \text{illness} & \text{(Revaluation phase)} \\ \text{lever} \;\to \text{press} \to \text{X} & \text{(Extinction phase)} \end{array}$$

Under these circumstances, $w(\text{X})$ would become negative during revaluation, and thus capable of decreasing $v(\text{lever} \to \text{press})$ during extinction. Simulation results for this scenario are in Fig. 10c, f. There is no difference between the experimental and control groups on the first extinction experience, because $v(\text{lever} \to \text{press})$ is updated only after X is experienced. A robust difference, however, appears rapidly starting with the second extinction experience. Whether this account is viable depends on whether stimuli common to revaluation and extinction can be credibly identified. We leave a comprehensive evaluation to future research, but we offer two remarks.

First, common stimuli can readily be identified in some studies. For example, Adams and Dickinson (1981) conducted the revaluation experience in the same Skinner box used for the other phases (although the lever required for the instrumental response was temporarily removed), as did Rashotte et al. (1977) when demonstrating the revaluation of a second-order Pavlovian CS by the extinction of the first-order CS. Common stimuli, however, are harder to identify in other cases. For example, Chen and Amsel (1980) found that a devaluation treatment in the animal's home cage led to slower running toward the goal box of a runway, which had previously contained the devalued outcome.

Second, even in the presence of common stimuli, A-learning predicts that revaluation should not affect responding in extinction on the very first experience. This prediction is difficult to test using published data, in which each data point commonly includes many responses. When response rates have been measured with high temporal resolution, however, the pattern in Fig. 11 has been observed, such as in Balleine and Dickinson (2005) and Dezfouli et al. (2014), both employing one-minute bins.

Conclusions about revaluation

Data and theory highlight many factors that can contribute to revaluation, including the precise sequence of

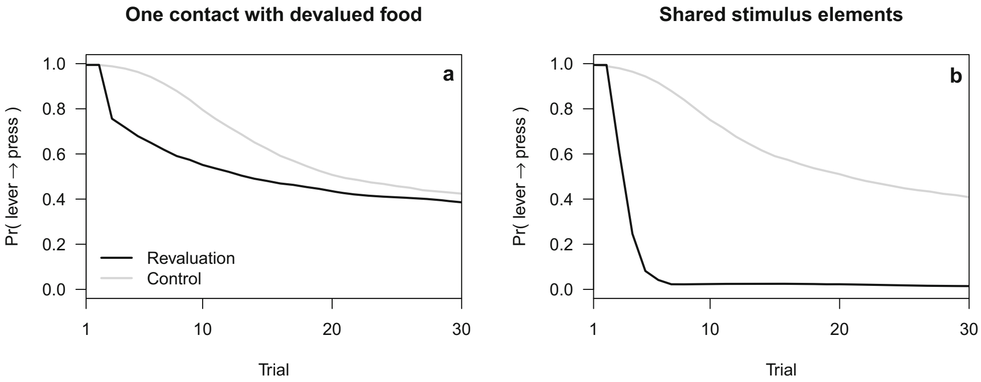

Fig. 11 Replication of the extinction phase of the simulation in Fig. 10, with modifications that enhance revaluation. **a** One experience of pressing the lever and receiving the devalued pellet was inserted between the revaluation and extinction phases. **b** The revaluation experience is assumed to have a shared stimulus element with extinction experiences. See text for details. Simulation scripts and model parameters are available online

experiences provided to the animals, which stimulus elements appear in each experimental phase, and which reinforcers are used (Rescorla, 1980; Holland, 2008; Balleine, 2011). While it remains to be determined whether A-learning can explain revaluation data satisfactorily, we hope to have shown that the joint effect of S-R values and stimulus values can produce a variety of outcomes, which are broadly consistent with available data.

Evaluative Conditioning and Incentive Learning A-learning's account of revaluation can be considered a mathematical formalization of the ideas pioneered by Garcia (1989) and further developed by Balleine and Dickinson (1991). In this account, outcome revaluation would be a two-step process. In the first step, termed *evaluative conditioning,* the pairing of food and illness changes the perceived value of the food, but does not change responding. The latter changes in the second step, termed *incentive learning,* when the animal contacts the outcome again and experiences its changed value. Only after this second step the new value of the outcome can affect responding. These ideas are remarkably close to how revaluation operations in A-learning. Evaluative conditioning corresponds to a change in the stimulus value of the food, $w(\text{pellet})$, while incentive learning to a change in the S-R value of the action that yields the food, $v(\text{lever} \to \text{press})$, which in turn is based on the changed stimulus value of the food.

Distance From the Reinforcer Using a Pavlovian procedure, Holland and Straub (1979) found that behaviors that are more proximal to the outcome, such as contacting and picking up a food pellet, are more sensitive to devaluation than more distal behaviors, such as general activity and approach to the food cup. Similarly, Balleine and Dickinson (2005) found that the second link of a two-link instrumental chain is more easily affected by devaluation than the first link. These findings are directly compatible with A-learning. Consider, for example, the sequence:

$$s_1 \to b_1 \to s_2 \to b_2 \to s_3 \to b_3 \to \text{outcome}$$

If the outcome changes value, the experience $s_3 \to b_3 \to$ new outcome can immediately change $v(s_3 \to b_3)$, but a minimum of two experiences would be necessary to change $v(s_2 \to b_2)$, because the latter change is mediated by $w(s_3)$ rather than directly by the value of the outcome. Likewise, it would take a minimum of three experiences to alter $v(s_1 \to b_1)$.

This reasoning may explain why Adams (1980) did not observe revaluation in extinction while Adams and Dickinson (1981) did, despite the two studies being performed in the same laboratory and with nearly identical procedures. The latter study, in fact, employed three pellet→toxin pairings, allowing the negative value of the toxin to influence more behaviors, whereas Adams (1980) employed a single pairing. If, for example, the act of eating is decomposed as follows:

$$\text{pellet} \to \text{seize} \to \text{taste} \to \text{swallow} \to \text{toxin}$$

then a single devaluation experience would change the value of the taste of the pellet, but not of its sight or of other external stimuli that may be shared between the revaluation and extinction phases. This is in line with Garcia (1989) suggestion that a single experience can revalue only the taste of food, while at least a second experience is necessary to affect consumption. The same considerations provide a formal justification to observations, at first puzzling, that adding or removing a movable panel in front of the food magazine can influence the results of revaluation experiments. As observed by Balleine and Dickinson (2005), operating the panel introduces an additional response, and thus increases the distance between previous responses and the outcome. According to A-learning, the back-propagation of value along stimulus-response sequences is a general principle that underlies many phenomena. It is not specific to the case of taste aversion learning in which it was recognized by Garcia (1989).

Pavlovian-to-instrumental transfer

In Pavlovian-to-instrumental transfer (PIT), a Pavlovian CS associated with a reinforcer is found to facilitate an instrumental response that procures the same reinforcer (specific PIT) or similar ones (general PIT). Here we consider the following simple PIT paradigm ("single-lever" paradigm in Cartoni et al. (2016)). First, a Pavlovian CS is established, for example by repeated CS→food experiences. Then, an instrumental response such as lever→press→food is established. Lastly, the response is tested in extinction, in the presence and in the absence of the CS. PIT occurs if the response rate is higher during the CS. We consider two potential contributions to PIT: that the CS can influence instrumental responding by means of the stimulus value it acquires during Pavlovian training, and that Pavlovian and instrumental learning affect some of the same behaviors. We consider these contributions separately for clarity, but they are not mutually exclusive.

The first mechanism is simple if perhaps surprising. Pavlovian training has two effects: it establishes CRs to the CS and it endows the CS with stimulus value. The latter can influence instrumental behavior during the PIT test as follows. Suppose that the animal presses the lever during the CS presentation, resulting in the following $S \to B \to S'$ sequence:

$$\ldots \to \text{CS} + \text{lever} \to \text{press} \to \text{CS} + \text{lever} \to \ldots$$

According to A-learning, the animal would perceive the outcome $S' = \text{CS}+\text{lever}$, and use its reinforcement value of $w(\text{CS}) + w(\text{lever})$ to update the S-R values $v(\text{CS} \to \text{press})$ and $v(\text{lever} \to \text{press})$. On the other hand, if the animal presses the lever in the absence of the CS the relevant $S \to B \to S'$ sequence is

$$\ldots \to \text{lever} \to \text{press} \to \text{lever} \to \ldots$$

which rewards pressing the lever with the value $w(\text{lever})$, which is smaller than $w(\text{CS})+w(\text{lever})$. During the PIT test, which is conducted in extinction, both $w(\text{CS})$ and $w(\text{lever})$ are predicted to decrease, but a high enough $w(\text{CS})$ can result in slower extinction of lever pressing in the presence of the CS than in its absence. This effect is demonstrated in Fig. 12, which also shows the underlying dynamics of S-R values and stimulus values. Notably, $v(\text{CS} \to \text{press})$ is zero at the beginning of the test because lever pressing has never been rewarded to the CS, but it transiently increases because lever pressing in the presence of the CS is rewarded by the stimulus value of the CS itself. Thus, a PIT effect can emerge during the test phase because of the temporary increase of $v(\text{CS} \to \text{press})$. This leads to the prediction that no PIT should be apparent at the very beginning of the test, but we have not found studies in which PIT is reported with sufficient granularity to test this hypothesis. A second prediction is that no PIT would ensue from this mechanism if pressing the lever terminates the CS.

A second potential contribution to PIT is that the CS can interfere with or facilitate instrumental learning. For example, the CS may evoke increased exploration of the environment (Holland & Rescorla, 1975b), which may lead the animal in proximity of the lever. At the same time, the CS may elicit approach to the food magazine, which would interfere with instrumental responding by drawing the animal away from the lever (Holmes et al., 2010; Cartoni et al., 2016). A PIT effect would arise if Pavlovian facilitation is, overall, stronger than interference. We study the interplay between different behaviors in PIT as follows. We assume that the food magazine is located through a

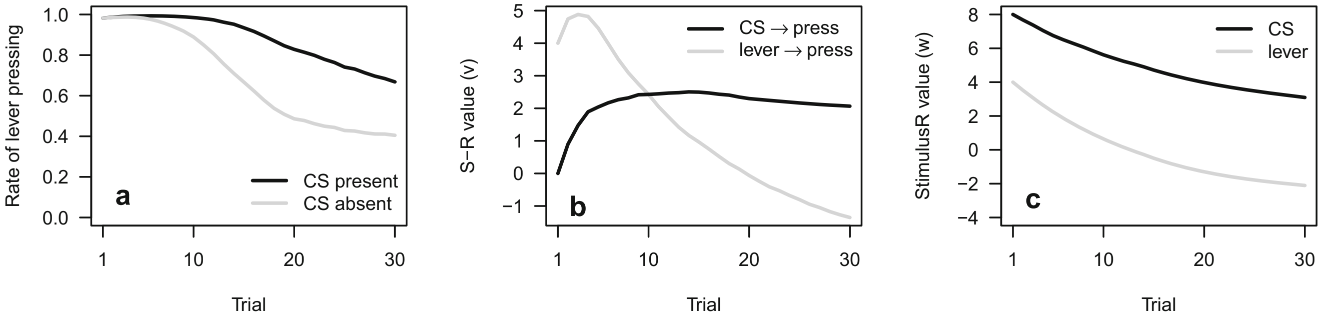

Fig. 12 Contribution of conditioned CS value to Pavlovian-to-instrumental transfer. **a** Probability of lever pressing during the PIT test in extinction, in the presence and absence of the CS. A PIT effect is evident in the slower extinction of lever pressing in the presence of the CS. **b** Underlying S-R values. **c** Underlying stimulus values. Similar results are obtained for many parameter values, provided the CS has high enough value and long enough duration to effectively reinforce lever pressing in the CS presence. Simulation script and model parameters are available online

Table 3 Design for the simulations of Pavlovian-to-instrumental transfer reported in Fig. 13. Pavlovian training establishes a search response to the compound stimulus of the conditioning box plus a CS. Instrumental training establishes a behavioral chain that includes the search response and lever pressing. The transfer test compares performance of the instrumental chain in the presence vs. absence of the CS. See text for further details. The notation $\neg b$ indicates behavior other than b

Phase	Experiences
Pavlovian training	box,CS → search → food
	box,CS → ¬search → no food
	box → ¬search → no food
Instrumental training	box → search → lever → press → food
	box → ¬search → no food
	box → search → lever → ¬press → no food
Transfer test	box → search → lever → press → no food
	box,CS → search → lever, CS → press → no food

"search" response that, following Pavlovian conditioning, is potentiated during the CS. We also assume that the same search behavior is performed during instrumental learning to locate the lever. Thus the instrumental response is actually the chain box→search→lever→press→pellet, where "box" represents the Skinner box in which both Pavlovian and instrumental training are conducted. When the lever is present, searching results in locating either the magazine or the lever, with equal probability. Finally, during the transfer test, periods with and without the CS alternate, but lever presses are not reinforced. The simulation setup is summarized in Table 3. The results presented in Fig. 13a demonstrate a PIT effect, in that more lever presses occur when the CS is present than when it is absent. Furthermore, Fig. 13b and c show that lever pressing is, in fact, depressed during the CS, but that this effect is more than compensated by the increased exploration of the environment, which leads to the lever being encountered more often.

PIT tests are often more detailed than our simple simulation, comprising multiple CSs, responses, and outcomes. In such tests, training with one CS-outcome pair facilitates instrumental behavior for other pairs, too (general PIT), but it facilitates the instrumental response that yields the same outcome the most (specific PIT; Cartoni et al. (2016)). We have not modeled such outcome specificity above, but it is conceivable that specific CSs can get associated with either internal states or external stimuli that favor certain responses over others. We do not claim to resolve all issues around PIT with these exploratory arguments, but rather to point out that an analysis that considers conditioned reinforcement and behavioral chain analysis could help understand PIT. Our computational framework can be used to model different scenarios in great detail, and can suggests new tests. For example, one could place the CS that signals outcome 1 (e.g., a light) near the lever for outcome 2, and vice-versa. According to our analysis, this arrangement should reduce the PIT effect.

Comparison with other TD models

As mentioned in the Introduction, TD algorithms are partly inspired by learning psychology and are having a growing impact in the behavioral neuroscience of learning (Dayan & Niv, 2008; Balleine et al., 2009). The latter has primarily considered three algorithms: Q-learning, SARSA, and the actor-critic model (Table 4). A-learning, on the other hand, adopts the lesser-known QV-learning (Wiering, 2005). Is any of these algorithms a better starting point to understand animal learning? Neural data suggest that animals may use several algorithms concurrently (Balleine et al., 2009; Dezfouli & Balleine, 2013), but behavioral data are also relevant because different TD algorithms can make strikingly different predictions when conditions change, such as when contingencies are altered or when stimuli are added or removed. For example, Fig. 14 shows

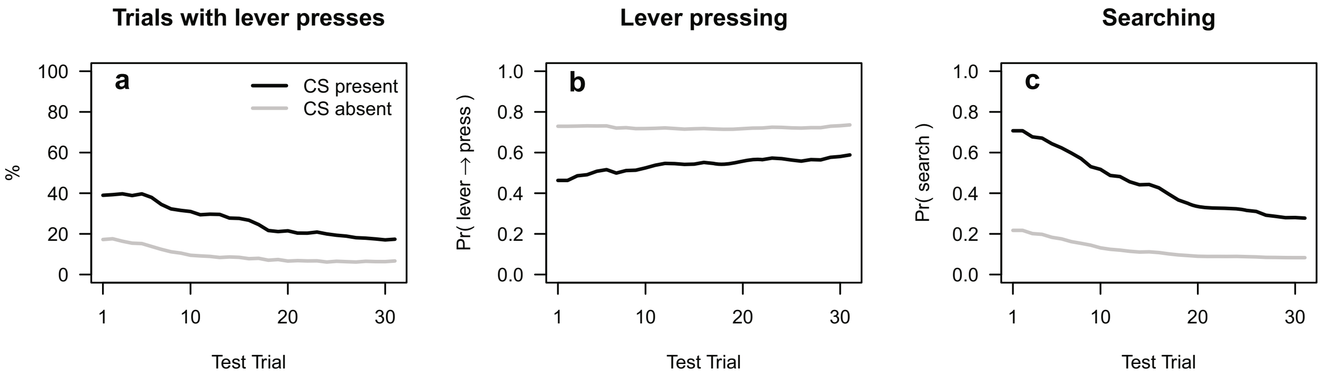

Fig. 13 Simulation of the single-lever Pavlovian-to-instrumental transfer experiment summarized in Table 3. **a** Average number of completed lever presses per trial during the transfer test, depending on whether the CS is present or absent. **b** Probability of a lever press, once the lever is found. **c** Probability of a search response, which can lead to finding the lever or the food magazine with equal probability. See Table 3 and the text for details. Simulation script and model parameters are available online

Table 4 A comparison of temporal-difference (TD) learning algorithms with reference to the QV-learning model adopted by A-learning (first row). A * indicates the same equation as in A-learning. Q-learning, SARSA, and Expected SARSA do not learn stimulus values; they compute them based on S-R values. Q-learning uses as stimulus value for s the maximum S-R value known for any behavior possible in s. SARSA uses simply the S-R value for the behavior that is actually performed. Expected SARSA uses the value expected from behaving according to its current knowledge. The actor-critic model learns stimulus values as A-learning, and uses the same error term to drive learning of both stimulus values and v variables (in this algorithm, v variables do not represent values). The A-learning equations are known as QV-learning in machine learning (Wiering, 2005). The other algorithms are described in Sutton and Barto (2018). Note: Machine learning uses Q for S-R values and V for stimulus values; here we use v and w, respectively, for closer adherence to psychological tradition

Model	S-R values	Stimulus values
QV-learning	$\Delta v(s \to b) = \alpha_v \left[u(s') + w(s') - v(s \to b)\right]$	$\Delta w(s) = \alpha_w \left[u(s') + w(s') - w(s)\right]$
Q-learning	*	$w(s') = \max_{b'} v(s \to b')$
SARSA	*	$w(s) = v(s' \to b')$
Expected SARSA	*	$w(s') = \sum_{b'} \Pr(s' \to b') v(s' \to b')$
Actor-critic	$\Delta v(s \to b) = \frac{\alpha_v}{\alpha_w} \beta \left[1 - \Pr(s \to b)\right] \Delta w(s)$	*

predictions from actor-critic, Expected SARSA, and Q-learning, about the partial reinforcement-extinction effect (PREE, results from A-learning are in Fig. 1b). Expected SARSA is equivalent to computing stimulus values as the average of current S-R values (Table 4, line 4). This model (and SARSA, not shown) does not reproduce the PREE because S-R values do not maintain a memory of whether they were attained by partial or continuous reinforcement. In A-learning, this memory is maintained by stimulus values (see The Sequential Nature of Learning). Interestingly, the actor-critic model ((4), line 5) does not reproduce the PREE despite learning stimulus values in the same way as A-learning. In this case, the reason is that its v variables are driven by *changes* in stimulus values, which are not, in general, smaller following partial rather than continuous reinforcement. In A-learning, S-R value changes depend on the magnitude of stimulus values, rather than of their changes. Lastly, the behavior of Q-learning depends on model parameters. The model in Table 4, line 2, predicts that a partially reinforced response should never extinguish, because it continues to be reinforced by the previously established maximum v value. This prediction can be corrected by multiplying the $\max_{b'} v(s' \to b')$ term (see Table 4) by a factor $\gamma < 1$ ("discounting," see Sutton & Barto 2018), as shown in Fig. 14c.

A comprehensive evaluation of TD model is outside the present scope, but myriad tests are possible using data from experimental psychology. We can also devise new tests based on the algorithmic characteristics of each model. For example, A-learning and the actor-critic model can be differentiated based on the fact that, in the latter, v variables stop changing once stimulus values are accurately predicted ($\Delta w(s') = 0$ in Table 4). In A-learning, however, this condition does not guarantee that S-R values do not change. Suppose, for example, that an animal learns to press a lever for food. After this behavior is acquired, a second lever is added that yields the same food. A-learning predicts that the animal would switch to use both levers equally, while the actor-critic model predicts that the animal would learn little about the second lever because it does not change the value of the situation (Fig. 15). We may call this phenomenon "response blocking" in analogy to blocking between stimuli that predict the same outcome (Kamin, 1969; Pearce, 2008). A-learning's prediction appears more

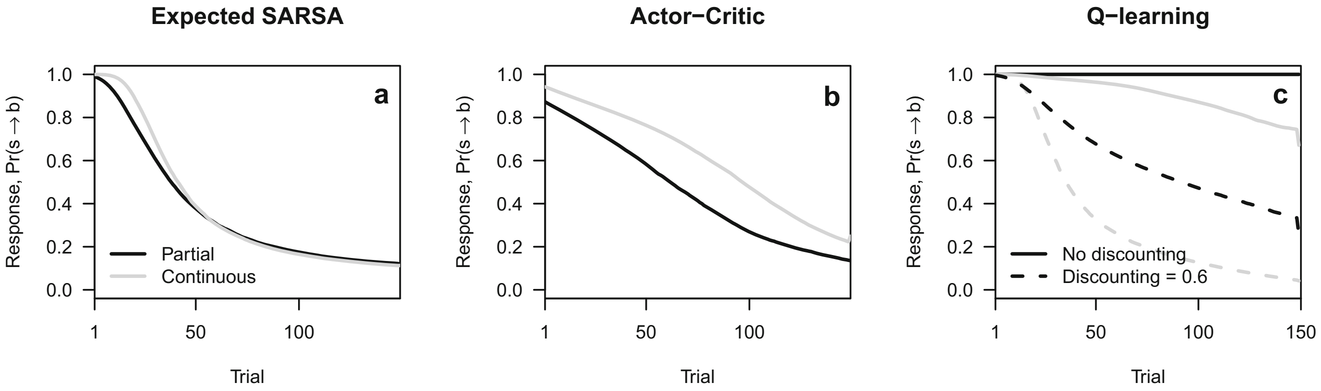

Fig. 14 Comparison of three temporal-difference learning algorithms during extinction of a partially vs. continuously reinforced response, see Fig. 1 for A-learning. Simulation scripts and model parameters are available online

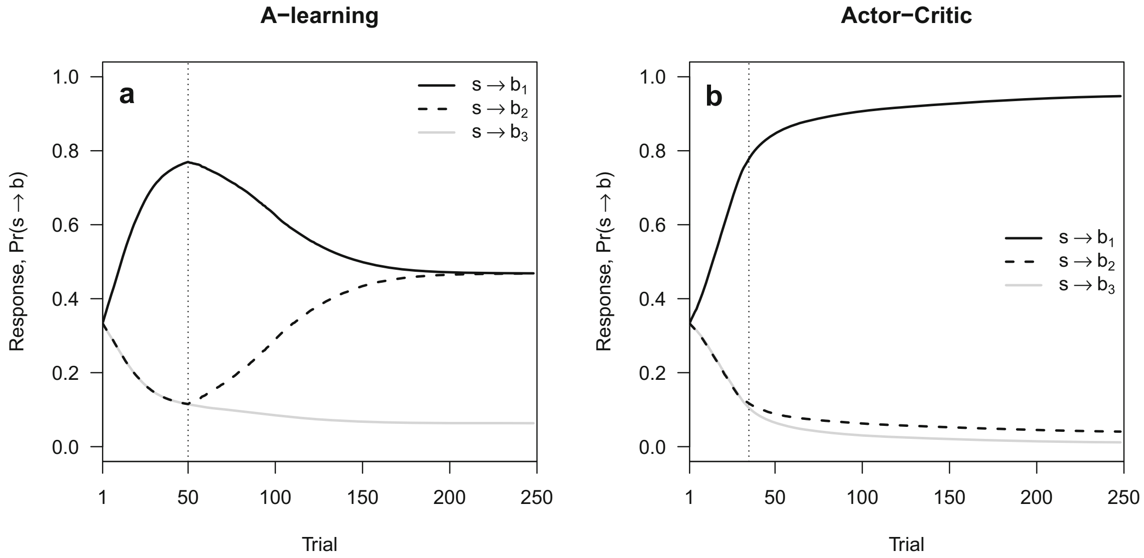

Fig. 15 A comparison of A-learning **a** and the actor-critic model **b**, see rows 1 and 5 in Table 4. The simulations assume that three responses are available to a stimulus, which are equally likely initially. In a first phase, one response is reinforced and the other two extinguished. Both models learn to perform the reinforced response with high probability. In a second phase (marked by the *dotted line*), a second response is reinforced in addition to the first. A-learning switches to using the two reinforced responses equally, while actor-critic continues to use the first response almost exclusively. The duration of the first phase has been set separately for the two algorithms to yield similar response levels for the three behaviors. Simulation script and model parameters are available online

realistic than the actor-critic's prediction, as an equalization of response rates is predicted based on the matching law (Herrnstein, 1974; Baum, 1974).

Discussion

A-learning retains key features of current psychological theory, such as error-correcting learning rules (Eqs. 4 and 5) and sum rules (Eqs. 10 and 11) from the Rescorla and Wagner (1972) model. To these, it adds two elements: conditioned reinforcement, implemented as stimulus values, and a decision rule based on S-R values. The resulting model remains conceptually simple, yet exhibits complex learning dynamics that often require detailed analysis. Above, we have seen that A-learning appears consistent with data on many phenomena, including the acquisition of instrumental, Pavlovian, and avoidance responses, omission training, autoshaping, matching, behavioral contrast, the acquisition and extinction of behavioral chains and higher-order Pavlovian CRs, the partial reinforcement extinction effect, the effect of free reinforcement on instrumental responding, Pavlovian-to-instrumental transfer, and outcome revaluation. In Enquist et al. (2016), we replicated “misbehavior” (Breland & Breland, 1961), and violation of expectation (Roper, 1984; Haskell et al., 2000), in addition to results from studies of planning in Lind (2018b) and of social learning in Lind et al. (2019). All of these results stem from Eqs. 4, 5, and 6. A-learning takes some unconventional stances (see Comparison with Current Theory), and some of the accounts it offers need to be validated empirically (see Pavlovian Acquisition and Outcome Revaluation). Furthermore, A-learning is still incomplete. We highlight some missing elements in the concluding part of this Discussion. Before that, we consider some remaining conceptual points.

A-learning models both Pavlovian and instrumental learning

Whether the procedural distinction between Pavlovian (S-S) and instrumental (R-S) contingencies reflects a separation of underlying learning processes has been hotly debated, to the point of trying to explain instrumental learning in Pavlovian terms, and Pavlovian learning in instrumental terms (Mackintosh & Dickinson, 1979; Mackintosh, 1983). These attempts were ultimately rejected because many instances of learning conform better to one of the two conceptions. That is, some learning appears rigidly determined by S-S contingencies, in which case we deem it Pavlovian, while other learning is mainly sensitive to R-S contingencies and is flexible in form, in which case we deem it instrumental. This resolution is appealing and has spurred productive theorizing over many decades, yet it leaves some open issues. As recalled in the Introduction, a typical learning experiment includes both Pavlovian and instrumental contingencies, and most responses are sensitive to both (Mackintosh, 1983). It is thus desirable to develop mathematical models that are sensitive to both Pavlovian and instrumental contingencies.

In A-learning, both instrumental and Pavlovian learning depend on stimulus value learning (w values) and S-R value learning (v values). With generic parameter settings (the same α_v, α_w, and β values for all combinations of stimuli and responses), A-learning behaves purely "instrumentally." That is, it learns by trial and error to maximize reinforcement. "Pavlovian" learning is achieved through parameter settings that enable only certain modifications of behavior (see Eqs. 14 and 15, and surrounding text). As discussed in Pavlovian Acquisition, these settings model the biological organization of behavior systems (Timberlake, 1983, 1994; Domjan, 1993, 2008; Hogan, 2017). The effects of biological predispositions on instrumental learning (Hinde & Stevenson-Hinde, 1973; Shettleworth, 1975, 1978; Poper, 1983) are implemented in the same way. To our knowledge, A-learning is the only mathematical model that can exhibit not only purely Pavlovian and purely instrumental learning, which appears rare (Mackintosh, 1983), but also the more realistic intermediate cases.

A-learning reinterprets Pavlovian learning

We discussed above several differences between current accounts of Pavlovian conditioning in terms of S-S associations, and A-learning's account in terms of S-R values and stimulus values. These differences lead to specific predictions about the course of conditioning (see Pavlovian Acquisition) and how outcome revaluation treatments affect learned behavior (see Outcome Revaluation). Studies of outcome revaluation have also led to the conclusion that first-order conditioning results primarily in CS-US associations, and higher-order conditioning in CS-CR associations, because first-order CRs are more easily affected by revaluation than higher-order CRs (see Distance From the Reinforcer). A difficulty with this conclusion is that many responses that are customarily considered first-order CRs might as well be considered higher-order ones. For example, characteristics of food such as flavor and appearance are often learned (Pavlov, 1927; Ewer, 1968), hence food revaluation treatments may be argued to affect second-order CRs. A-learning side-steps this difficulty because it posits that all responses have the same associative structure (S-R values plus stimulus values), yet it allows first-order CRs to be more sensitive to revaluation because they are more proximal to the outcome whose value is manipulated (see Distance From the Reinforcer).

A-learning includes habits and goals

That instrumental behavior is not always sensitive to outcome revaluation has been interpreted in terms of complementary "habitual" and "goal-directed" learning systems (Dickinson & Weiskrantz, 1985; Balleine et al., 2009). The habitual system is similar to our S-R values in that it responds to stimuli without taking into account outcome value. The goal-directed system is defined by being sensitive to outcome (goal) value, and has been further subdivided into two processes (Balleine & Dickinson, 1998): one that estimates the contingency between responses and outcomes, and one that learns about the value of the outcome (see Evaluative Conditioning and Incentive Learning). A-learning contains the same elements, albeit somewhat rearranged. S-R values are insensitive to outcome revaluation (habitual), but are sensitive to instrumental contingencies (see Instrumental Acquisition). In addition, stimulus values track outcome value and can influence S-R values, resulting in a variety of revaluation effects. This influence is not immediate as in some other models of goal-directed action (Balleine et al., 2009; Dezfouli & Balleine, 2013), but it may suffice to account for many revaluation findings (see Outcome Revaluation).

The distinction between habitual and goal-directed behavior has a further root in the finding that sensitivity to outcome devaluation can decrease with prolonged training (Adams (1982) and Dickinson and Weiskrantz (1985), but see Garr and Delamater (2019)), suggesting a shift of control from the goal-directed to the habitual system (Dickinson & Weiskrantz, 1985; Dezfouli & Balleine, 2013). In A-learning, this finding can occur because response probability depends non-linearly on S-R values. A moderate S-R value may be sufficient to achieve high response probability, after which prolonged training may continue to increase S-R value with a negligible increase in responding. When S-R values are very high, however, outcome revaluation may have a smaller impact, because reducing responding requires a large decrease in S-R values.

In summary, A-learning is not goal-directed in the sense of explicitly planning a course of action to reach a desired outcome, but it is in the sense that stimulus values can orient behavior toward high-value outcomes. At the same time, A-learning can behave habitually if stimulus values cannot readily influence S-R values. More work is necessary to determine whether this can be a satisfactory account of observed shifts between habitual and goal-directed behavior.

Outlook

Several important domains are currently outside of the scope of A-learning, such as motivational processes (Dickinson & Balleine, 1994; Balleine, 2011), perceptual learning effects like sensory preconditioning (Brogden, 1939; Rizley & Rescorla, 1972; Hall, 1991), the role of working memory in associative learning (Capaldi, 1994), attentional processes that may change both learning rates and stimulus salience (Mackintosh, 1975; Pearce & Hall, 1980; Le Pelley, 2004; George & Pearce, 2012), and how A-learning may map onto neural structures. Some of these issues are not specific to A-learning, and can be approached with standard methods. For example, in Enquist et al. (2016), we presented a connectionist implementation of A-learning showing that S-R and stimulus values may be encoded as synaptic strengths. Based on related models, this implementation is expected to reproduce stimulus generalization accurately (Van Roy, 2002; Enquist & Ghirlanda, 2005), and to be helpful in the study of perceptual learning effects, such as sensory preconditioning (Hebb, 1966; Enquist & Ghirlanda, 2005). Other phenomena, however, require novel research whose outcome is difficult to anticipate, such as how A-learning may interface with motivational and attentional processes. It remains to be seen whether A-learning can be extended satisfactorily to these domains.

Open practices statement

Simulation scripts are available at https://osf.io/b8mez.

References

Adams, C. D. (1980). Post-conditioning devaluation of an instrumental reinforcer has no effect on extinction performance. *The Quarterly Journal of Experimental Psychology*, *32*(3), 447–458.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*(2b), 77–98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *33*(2), 109–121.

Atkinson, R. R., & Estes, W. K. Stimulus sampling theory. In *Handbook of mathematical psychology*, (p. 1963). New York: Wiley.

Atnip, G. W. (1977). Stimulus-and response-reinforcer contingencies in autoshaping, operant, classical, and omission training procedures in rats. *Journal of the Experimental Analysis of Behavior*, *28*(1), 59–69.

Balleine, B. W. (2011). Sensation, incentive learning, and the motivational control of goal-directed action. In Gottfried, J. A. (Ed.) *Neurobiology of Sensation and Reward, Taylor & Francis Group 1 edn*.

Balleine, B. W., & Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology*, *43*(3), 279–296.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407–419.

Balleine, B. W., & Dickinson, A. (2005). Effects of outcome devaluation on the performance of a heterogeneous instrumental chain. International Journal of Comparative Psychology, 18(4).

Balleine, B. W., Daw, N. D., & O'Doherty, J. P. (2009). Chapter 24 - multiple forms of value learning and the function of dopamine. In Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (Eds.) *Neuroeconomics*. ISBN 978-0-12-374176-9, (pp. 367–387). London: Academic Press.

Baum, W. M. (1973). The correlation-based law of effect 1. *Journal of the Experimental Analysis of Behavior*, *20*(1), 137–153.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching 1. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242.

Baum, W. M. (2017). Understanding behaviorism: Behavior, culture, and evolution. John Wiley & Sons.

Bernstein, I. L. (1999). Taste aversion learning: a contemporary perspective. *Nutrition*, *15*(3), 229–234.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology:, Animal Behavior Processes*, *104*(1), 3–21.

Boakes, R. A. (1977). Performance on learning to associate a stimulus with positive reinforcement. In Davis, H., & Hurwitz, H. (Eds.) *Operant-Pavlovian interactions*, (pp. 67–97). Hillsdale: Lawrence Erlbaum Associates.

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological Review*, *77*(1), 32.

Bouton, M. E. (2016). Learning and behavior: A contemporary synthesis. Sinauer 2nd edn.

Bouton, M. E., Winterbauer, N. E., & Todd, T. P. (2012). Relapse processes after the extinction of instrumental learning: renewal, resurgence, and reacquisition. *Behavioural Processes*, *90*(1), 130–141.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *61*, 681–684.

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*(4), 323–332.

Brown, P. L., & Jenkins, He. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior*, *11*(1), 1–8.

Bush, . R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*(5), 313.

Capaldi, E.J. (1971). Memory and learning: A sequential viewpoint. Animal Memory, pp. 111–154.

Capaldi, E. J. (1994). The sequential view: from rapidly fading stimulus traces to the organization of memory and the abstract concept of number. *Psychonomic Bulletin & Review*, *1*(2), 156–181.

Cartoni, E., Balleine, B. W., & Baldassarre, G. (2016). Appetitive Pavlovian-instrumental transfer: A review. *Neuroscience & Biobehavioral Reviews*, *71*, 829–848.

Chen, J., & Amsel, A. (1980). Recall (versus recognition) of taste and immunization against aversive taste anticipations based on illness. *Science*, *209*(4458), 831–833.

Coleman, S. R., & Gormezano, I. (1979). Classical conditioning and the law of effect: Historical and empirical assessment. *Behaviorism*, *7*(2), 1–33.

Curio, E., Ernst, U., & Vieth, W. (1978). Cultural transmission of enemy recognition. One function of mobbing. *Science*, *202*, 899–901.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Delamater, A. R., & Westbrook, R. F. (2014). Psychological and neural mechanisms of experimental extinction: a selective review. *Neurobiology of Learning and Memory*, *108*, 38–51.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLos Computational Biology*, *9*(12), e1003364.

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130482.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge: Cambridge University Press.

Dickinson, A., & Weiskrantz, L. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. W. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Dickinson, A., & Charnock, D. J. (1985). Contingency effects with maintained instrumental reinforcement. *The Quarterly Journal of Experimental Psychology B*, *37*(4b), 397–416.

Dickinson, A., Wood, N., & Smith, J. W. (2002). Alcohol seeking by rats: action or habit. *The Quarterly Journal of Experimental Psychology:, Section B*, *55*(4), 331–348.

Dollard, J., Miller, N. E., & Doob, L. W. (1939). Orval Hobart Mowrer, and Robert R Sears Frustration and aggression.

Domjan, M. (1993). Biological constraints on instrumental and classical conditioning: Implications for general process theory. In Bower, G. H. (Ed.) *The psychology of learning and motivation*, (Vol. 17, pp. 215–277). New York: Academic Press.

Domjan, M. (2008). Adaptive specializations and generality of the laws of classical and instrumental conditioning. In Byrne, J. (Ed.) *Learning theory and behavior*, (Vol. 1, pp. 327–340). Oxford: Elsevier.

Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248.

Eldridge, G. D., & Pear, J. J. (1987). Topographical variations in behavior during autoshaping, automaintenance, and omission training. *Journal of the Experimental Analysis of Behavior*, *47*(3), 319–333.

Enquist, M., & Ghirlanda, S. (2005). *Neural networks and animal behavior*. Princeton: Princeton University Press.

Enquist, M., Lind, J., & Ghirlanda, S. (2016). The power of associative learning and the ontogeny of optimal behaviour. *Royal Society Open Science*, *3*(11), 160734.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Ewer, R. F. (1968). *Ethology of mammals*. London: Logos Press Limited.

Fanselow, M. S. (1989). The adaptive function of conditioned defensive behavior: An ecological approach to Pavlovian stimulus-substitution theory.

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, *1*(4), 429–438.

Frieman, J., & Reilly, S. (2015). Learning: A behavioral, cognitive, and evolutionary synthesis. Sage Publications.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131.

Garcia, J. (1989). Food for Tolman: Cognition and cathexis in concert. In Arche, T., & Nilsson, L.-G. (Eds.) *Aversion, avoidance and anxiety*, (pp. 45–85): Erlbaum.

Garr, E., & Delamater, A. R. (2019). Exploring the relationship between actions, habits, and automaticity in an action sequence task. *Learning & Memory*, *26*(4), 128–132.

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning & Behavior*, *40*(3r), 241–254.

Ghirlanda, S. (2018). ECCO: An error-correcting comparator theory. *Behavioural processes*, *154*, 36–44.

Ghirlanda, S., & Enquist, M. (2019). On the role of responses in Pavlovian acquisition. *Journal of Experimental Psychology:, Animal Learning and Cognition*, *45*(1), 59.

Ghirlanda, S., & Ibadullayev, I. (2015). Solution of the comparator theory of associative learning. *Psychological Review*, *122*(2), 242.

Gormezano, I., & Hiller, G. W. (1972). Omission training of the jaw-movement response of the rabbit to a water us. *Psychonomic Science*, *29*(5), 276–278.

Guthrie, E.R. (1942). Conditioning: a theory of learning in terms of stimulus, response and association.

Hall, G. (1991). *Perceptual and associative learning*. Cambridge: Cambridge University Press.

Hall, G. (1994). *Pavlovian conditioning: Laws of association*. New York and London: Academic Press.

Hall, G. (2002). Associative structures in Pavlovian and instrumental conditioning. In *Stevens' handbook of experimental psychology. Wiley online library*.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior*, *34*(3), 297–304.

Harris, J. A. (2011). The acquisition of conditioned responding. *Journal of Experimental Psychology:, Animal Behavior Processes*, *37*(2), 151.

Haselgrove, M., & Hogarth, L. (2013). Clinical applications of learning theory. Psychology Press.

Haskell, M., Coerse, N. C. A., & Forkman, B. (2000). Frustration-induced aggression in the domestic hen: the effect of thwarting access to food and water on aggressive responses and subsequent approach tendencies. *Behaviour*, *137*(4), 531–546.

Hauser, H., & Gandelman, R. (1985). Lever pressing for pups: evidence for hormonal influence upon maternal behavior of mice. *Hormones and Behavior*, *19*(4), 454–468.

Hearst, E. (1975). Pavlovian conditioning and directed movements. In *Psychology of learning and motivation*, (Vol. 9, pp. 215–262): Elsevier.

Hebb, D. O. (1966). *A textbook of psychology*, (2nd ed.). Philadelphia: W. B. Saunders Company.

Herrnstein, R. J. (1969). Method and theory in the study of avoidance. *Psychological Review*, *76*(1), 49–69.

Herrnstein, R. J. (1974). Formal properties of the matching law. *Journal of the Experimental Analysis of Behavior*, *21*(1), 159.

Hinde, R. A., & Stevenson-Hinde, J. (1973). *Constraints on learning*. New York: Appleton-Century-Crofts.

Hogan, J. A. (2017). The Study of Behavior: Organization, Methods, and Principles. Cambridge University Press.

Holland, P. C. (1977). Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. *Journal of Experimental Psychology:, Animal Behavior Processes*, *3*(1), 77.

Holland, P. C. (2008). Cognitive versus stimulus-response theories of learning. *Learning & Behavior*, *36*(3), 227–241.

Holland, P. C., & Rescorla, R. A. (1975a). The effect of two ways of devaluing the unconditioned stimulus after first-and second-order appetitive conditioning. *Journal of Experimental Psychology:, Animal Behavior Processes*, *1*(4), 355.

Holland, P. C., & Rescorla, R. A. (1975b). Second-order conditioning with food unconditioned stimulus. *Journal of Comparative and Physiological Psychology*, *88*(1), 459.

Holland, P. C., & Straub, J. J. (1979). Differential effects of two ways of devaluing the unconditioned stimulus after Pavlovian appetitive conditioning. *Journal of Experimental Psychology:, Animal Behavior Processes*, *5*(1), 65.

Holmes, N. M., Marchand, A. R., & Coutureau, E. (2010). Pavlovian to instrumental transfer: a neurobehavioural perspective. *Neuroscience & Biobehavioral Reviews*, *34*(8), 1277–1295.

Honey, R. C., Dwyer, D. M., & Iliescu, A. F. (2019). HeiDI: A model for Pavlovian learning and performance with reciprocal associations. bioRxiv.

Inoue-Nakamura, N., & Matsuzawa, T. (1997). Development of Stone Tool Use by Wild Chimpanzees (Pan troglodytes). *Journal of Comparative Psychology*, *11*(2), 159–173.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A., & Church, M. R. (Eds.) *Punishment and aversive behavior*, (pp. 279–296). New York: Appleton-Century-Crofts.

Konorski, J. (1967). *Integrative activity of the brain*. Chicago: University of Chicago Press.

Konorski, J., & Miller, S. (1937). On two types of conditioned reflex. *The Journal of General Psychology*, *16*(1), 264–272.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology*, *B57*, 193–243.

Lind, J. (2018a). What can associative learning do for planning?. *Royal Society Open Science*, *5*(11), 180778.

Lind, J. (2018b). What can associative learning do for planning?. *Royal Society Open Science*, *5*(11), 180778.

Lind, J., Ghirlanda, S., & Enquist, M. (2018). Social learning through associative processes: A computational theory. bioRxiv pp. 446906.

Lind, J., Ghirlanda, S., & Enquist, M. (2019). Social learning through associative processes: a computational theory. *Royal Society Open Science*, *6*, 181777.

Locurto, C., Terrace, H. S., & Gibbon, J. (1976). Autoshaping, random control, and omission training in the rat. *Journal of the Experimental Analysis of Behavior*, *26*(3), 451–462.

Lucas, G. A. (1975). The control of keypecks during automaintenance by prekeypeck omission training. *Animal Learning & Behavior*, *3*(1), 33–36.

Ludvig, E. A., Sutton, R. S., & James Kehoe, E. (2012). Evaluating the TD model of classical conditioning. *Learning & Behavior*, *40*(3), 305–319.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

Mackintosh, NJ. (1994). Animal learning and cognition, Academic Press.

Mackintosh, N. J., & Dickinson, A. (1979). Instrumental (Type II) conditioning. In Dickinson, A., & Boakes, R. A. (Eds.) *Mechanisms of learning and motivation. A memorial volume to Jerzy Konorsky*, (pp. 143–170). Hillsdale: Lawrence Erlbaum Associates.

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, *38*(1), 50–67.

McFarland, D. J. (1971). *Feedback mechanisms in animal behaviour*. London: Academic Press.

McGreevy, P., & Boakes, R. (2011). Carrots and sticks: Principles of animal training. Darlington Press.

Miller, N. E. (1935). A reply to sign-gestalt or conditioned reflex?. *Psychological Review*, *42*, 280–292.

Miller, N. E., & Carmona, A. (1967). Modification of a visceral response, salivation in thirsty dogs, by instrumental training with water reward. *Journal of Comparative and Physiological Psychology*, *63*(1), 1.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, *117*(3), 363.

Mineka, S., & Cook, M. (1988). Social learning and the acquisition of snake fear in monkeys. In *Social learning: Psychological and biological perspectives*. Hillsdale: Lawrence Erlbaum.

Mowrer, O. (1960). Two-factor learning theory: Versions one and two.

Oley, N. N., & Slotnick, B. M. (1970). Nesting material as a reinforcement for operant behavior in the rat. *Psychonomic Science*, *21*(1), 41–43.

Patten, R. L., & Rudy, J. W. (1967). The Sheffield omission training procedure applied to the conditioning of the licking response in rats. *Psychonomic Science*, *8*(11), 463–464.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Pearce, J. M. (2008). *Animal learning and cognition*. Hove: Psychology Press. East Sussex 3 edition.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*, 111–139.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning:, Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552.

Pearce, J. M., Redhead, E. S., & Aydin, A. (1997). Partial reinforcement in appetitive Pavlovian conditioning with rats. *The Quarterly Journal of Experimental Psychology: Section B*, *50*(4), 273–294.

Pierce, D. W., & Cheney, C. D. (2013). Behavior analysis and learning. Psychology Press.

Poling, A., & Poling, T. (1978). Automaintenance in Guinea pigs: Effects of feeding regimen and omission training 1. *Journal of the Experimental Analysis of Behavior*, *30*(1), 37–46.

Rashotte, M. E., Griffin, R. W., & Sisk, C. L. (1977). Second-order conditioning of the pigeon's keypeck. *Animal Learning & Behavior*, *5*(1), 25–38.

Rescorla, R. A. (1980). *Pavlovian Second Order Conditioning*. Hillsdale: Lawrence Erlbaum Ass. Reprinted in 2014 by Psychology Press, Hove, East Sussex.

Rescorla, R. A. (2002). Extinction. In Bäckman, L., & von Hofsten, C. (Eds.) *Psychology at the turn of the millennium. vol. 1: Cognitive, biological, and health perspectives*, (pp. 217–244). Hove: Taylor & Francis.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: current research and theory. Appleton-century-crofts*.

Rescorla, R. A. (1973). Effects of us habituation following conditioning. *Journal of Comparative and Physiological Psychology*, *82*(1), 137.

Reynolds, G. S. (1961). Behavioral contrast. *Journal of the Experimental Analysis of Behavior*, *4*(1), 57–71.

Rizley, R.oss.C., & Rescorla, R.obert.A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, *81*(1), 1.

Roper, T. J. (1983). Learning as a biological phenomenon. In Halliday, T. R., & Slater, P. J. (Eds.) *Genes, Development and Learning, no. 6*, (pp. 178–221). Oxford: Blackwell Scientific Publications.

Roper, T. J. (1984). Response of thirsty rats to absence of water: frustration, disinhibition or compensation? *Animal Behaviour*, *32*(4), 1225–1235.

Rozeboom, W. W. (1958). What is learned?—an empirical enigma. *Psychological Review*, *65*(1), 22.

Sanabria, F., Sitomer, M. T., & Killeen, P. R. (2006). Negative automaintenance omission training is effective. *Journal of the experimental analysis of behavior*, *86*(1), 1–10.

Schachtman, T. R., & Reilly, S. S. (2011). Associative learning and conditioning theory: Human and non-human applications. OUP USA.

Shapiro, M. M., & Herendeen, D. L. (1975). Food-reinforced inhibition of conditioned salivation in dogs. *Journal of comparative and physiological psychology*, *88*(2), 628.

Shettleworth, S. J. (1975). Reinforcement and the organisation of behavior in golden hamsters: hunger, environment and food reinforcement. *Journal of Experimental Psychology:, Animal Behavior Processes*, *1*(1), 56–87.

Shettleworth, S. J. (1978). Reinforcement and the organisation of behavior in golden hamsters: sunflower seed and nest paper reinforcers. *Animal Learning and Behavior*, *6*, 352–362.

Shettleworth, S. J. (1994). *Biological Approaches to Learning* Vol. 7. San Diego: Academic Press.

Sidman, M. (1953). Avoidance conditioning with brief shock and no exteroceptive warning signal. *Science*, *118*(3058), 157–158.

Skinner, B. F. (1934). The extinction of chained reflexes. *Proceedings of the National Academy of Sciences*, *20*(4), 234–237.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis*. Massachusetts: Copley Publishing Group Acton.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Konorski and Miller. *The Journal of General Psychology*, *16*(1), 272–279.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review*, *114*(3), 759–783.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–140. ftp://ftp.cs.umass.edu/pub/anw/pub/sutton/sutton-barto-81-PsychRev.pdf.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Thorndike, E. L. (1911). Animal intelligence: Experimental studies. Macmillan.

Thrailkill, E. A., & Bouton, M. E. (2015). Extinction of chained instrumental behaviors:, Effects of procurement extinction on consumption responding. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(3), 232.

Thrailkill, E. A., & Bouton, M. E. (2016). Extinction of chained instrumental behaviors: Effects of consumption extinction on procurement responding. *Learning & Behavior*, *44*(1), 85–96.

Timberlake, W. (1983). The functional organization of appetitive behavior: Behavior systems and learning. *Advances in analysis of behavior*, *3*, 177–221.

Timberlake, W. (1994). Behavior systems, associationism, and Pavlovian conditioning. *Psychonomic Bulletin & Review*, *1*(4), 405–420.

Trapold, M. A., & Overmier, J. B. (1972). The second learning process in instrumental learning. In Black, A. H., & Prokasy, W. F. (Eds.) *Classical Conditioning II: Current Research and Theory*, (pp. 427–452). New York: Appleton-Century-Crofts.

Van Roy, B. (2002). Neuro-dynamic programming: Overview and recent trends. In Feinberg, E. A., & Shwartz, A. (Eds.) *Handbook of Markov decision processes*, (pp. 431–459): Springer.

Wagner, A. R. (1981). Sop: A model of automatic memory processing in animal behavior. In Spear, N. E., & Miller, R. R. (Eds.) *Information Processing in Animals*, (pp. 15–58). Hillsdale: Erlbum.

Wiering, M. A. (2005). Qv (lambda)-learning: A new on-policy reinforcement learning algorithm. In *Proceedings of the 7th European Workshop on Reinforcement Learning*, (pp. 17–18).

Williams, B. A. (1994a). Conditioned reinforcement: Experimental and theoretical issues. *Behavior Analyst*, *2*, 261–285.

Williams, B. A. (1994b). *Reinforcement and choice*. New York and London: Academic Press.

Williams, B. A. (2002). Behavioral contrast redux. *Animal Learning & Behavior*, *30*(1), 1–20.

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Wayne Aldridge, J. (2009). A neural computational model of incentive salience. *PLOS Computational Biology*, *5*, 1–14.

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