A theory of actions and habits in free-operant behavior: The interaction of rate correlation and contiguity systems.

Omar D. Perez[1, 3] & Anthony Dickinson[2]

[1] Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA.

[2] Department of Psychology and Behavioural and Clinical Neurosciences Institute, University of Cambridge, Cambridge, UK.

[3] Nuffield College CESS-Santiago, Facultad de Administracion y Economia, Universidad de Santiago de Chile, Santiago, Chile.

Author Note

Correspondence concerning this article should be addressed to Omar D. Perez, 1200 East California Boulevard, Pasadena, California, CA91107. E-mail: odperez@caltech.edu

Abstract

Theories of instrumental actions assume the existence of multiple behavioral systems, one goal-directed which takes into account the consequences of actions, and one habitual that depends on previous reward history, both of which are predicated upon the notion of prediction-error to learn which actions should be performed. We present a model of free-operant instrumental actions in which goal-directed control is determined by the rate correlation between actions and outcomes whereas habitual responding is under the control of contiguous reward probability of an outcome, with these two systems interacting cooperatively and summating to control actions. The model anticipates the difference in performance between ratio and interval schedules and accounts for a number of additional phenomena such as the transition from goal-directed to habitual control with extended training and the persistence of goal-directed control under choice procedures and extinction. These results make the model unique in its joint predictions of behavioral control and performance for free-operant conditions.

Keywords: actions, habits, dual-system theory, reward schedules, instrumental conditioning, reinforcement learning

A theory of actions and habits in free-operant behavior: The interaction of rate correlation and contiguity systems.

Introduction

Instrumental action instantiates a unique reciprocal relationship between the mind and the world. Through instrumental learning we bring our representations of the consequences or outcomes of our actions into correspondence with the causal relationships in the world, whereas through instrumental action we bring the world into correspondence with the representations of our desires. However, this reciprocity assumes that instrumental behavior is goal-directed in the sense that it is based upon an interaction between a belief about the causal relation between an action and its outcome and a desire for that outcome (Dickinson and Balleine, 1994; Heyes and Dawson, 1990). Over the last forty years a wealth of evidence has accumulated that not only are humans capable of goal-directed action in this sense but so are other animals.

The canonical assay for the goal-directed status of instrumental behavior is the outcome revaluation procedure, which we shall illustrate with an early study by Adams and Dickinson (1981). They initially trained hungry rats to press a lever to receive either sugar or grain pellets with the alternative reward or outcome being delivered freely or non-contingently. The lever was then withdrawn and a flavor aversion was conditioned to one type of pellet by pairing its consumption with the induction of gastric malaise until the rat would no longer eat this type of pellet when freely presented. The purpose of this outcome devaluation was to remove the rat's desire for this type of pellet, while maintaining the desirability of the other type. If lever-pressing was mediated by knowledge of the causal relationship with the pellet outcome, devaluing this outcome should have reduced the rat's propensity to press when the lever was once again presented relative to the level of responding observed when the non-contingent pellet was devalued. This is exactly the result they observed (Adams and Dickinson, 1981). More recently, the finding has also been documented in both humans (Valentin *et al.*, 2007) and monkeys (Rhodes and Murray, 2013). It is important to note that this test is conducted under an extinction procedure where the delivery of the outcome is suspended; any devaluation effect should therefore reflect knowledge acquired during training rather than during the test itself.

Although research on the brain systems supporting goal-directed behavior has advanced during the last 20 years (for a review, see Balleine and O'Doherty, 2010), the nature of the psychological processes underlying the acquisition of action- or response-outcome knowledge remains relatively under-studied. This is in part because the psychology of learning has focused on the Pavlovian paradigm for the last 50 years or so given the greater experimental control afforded by such procedures. This research has generated a rich corpus of associative learning theories, all of which assume that learning is driven, in one way or another, by prediction errors (for a review, see Vogel *et al.*, 2004). In the case of Pavlovian learning, these errors reflect the extent to which the conditioned stimulus fails to predict to the occurrence (or non-occurrence) of the outcome. In the most straightforward of these theories, the larger the prediction error on a learning episode the less predicted is the outcome and the greater is the change in associative strength of the stimulus. As a consequence, the prediction error is reduced appropriately on subsequent, congruent learning episodes (Rescorla and Wagner, 1972). Based on the idea that Pavlovian associative learning is controlled by prediction errors and the multiple phenomena that paralleled those found in instrumental learning, Mackintosh and Dickinson (1979) suggested such errors play an analogous role in both types of learning processes.

Over the last decade or so, goal-directed learning has become increasingly couched in terms of computational reinforcement learning (RL). According to this approach (Daw *et al.*, 2005; Maia, 2009; Sutton and Barto, 1998), goal-directed behavior is controlled by model-based (MB) computations in which the agent learns a model of the state transitions produced by the instrumental contingencies and the value of each of the experienced states. At the time of performance, the agent searches the model to estimate the value of each of the actions available, and chooses the one that maximizes the outcome rate obtained over a number of episodes acting on the environment. Critically, what determines the value of each action in each state (or, alternatively, the probability of choosing each of the available actions in each state) is the probability that a rewarding outcome will be received given that the action is performed in each one of the states.

Whatever the difference between the associative and RL theory accounts of goal-directed action, both of these approaches share the assumption that the probability of a rewarding outcome is a primary determinant of instrumental goal-directed action. The reward probability directly determines the strength of the response-outcome association according to associative theory (Mackintosh and Dickinson, 1979) and the estimated value of an action in the case of RL theory. For both approaches, instrumental performance should be directly related to these variables. However, ever since the initial studies of instrumental outcome revaluation using free-operant schedules we have known that reward probability is unlikely to be the primary determinant of goal-directed control.

Ratio and interval contingencies

The initial investigations of goal-directed free-operant behavior using outcome devaluation with rats were uniformly unsuccessful (Adams, 1980; Holman, 1975; Morrison and Collyer, 1974). In contrast to the successful demonstration of devaluation reported by Adams and Dickinson (1981), prior studies had all trained rats to press the lever on a variable interval (VI) contingency between the response and the outcome. This class of schedule models a resource, such as nectar, that depletes when taken and regenerates with time. In practice, a VI schedule specifies the average time interval that has to elapse before the next outcome becomes available. In contrast, Adams and Dickinson (1981) used a variable ratio (VR) schedule, which models foraging in a non-depleting source so that each action has a fixed probability of yielding an outcome independently of the time elapsed since the last outcome obtained.

In an experimental analysis of the ratio-interval contrast, Dickinson et al. (1983) used a yoking procedure to match the outcome probability on the two schedules. In one pair of groups, the master rats were trained in an interval schedule, whereas the yoked animals were trained on ratio schedules with outcome probabilities that matched those generated by the master rats. In spite of the fact that the outcome probability per response was matched between the groups, outcome devaluation reduced performance of the ratio-trained but not the interval-trained group, suggesting that ratio training more readily establishes goal-directed control than interval training. This conclusion was reinforced when the outcome rate was matched by yoking the rates of the interval-trained rats to those generated by master ratio-trained animals. Again, ratio-, but not interval-trained animals, were sensitive to outcome devaluation. As the interval-trained rats pressed at a lower rate than the ratio-trained animals, goal-directed control was observed in the ratio-trained group even under a lower outcome probability experienced by those rats. The impact of the training schedule on the outcome devaluation effect has now received extensive replication (see Gremel and Costa, 2013; Hilario *et al.*, 2012; Wiltgen *et al.*, 2012).

The claim that ratio schedules more readily establish goal-directed control than does interval training finds further support by a study of the acquisition of beliefs about the effectiveness of an action in causing an outcome. Reed (2001) trained human participants on a fictional investment task in which pressing the space-bar on the keyboard acted as the instrumental response. Ratio training uniformly yielded higher judgments of the causal effectiveness of the key-press in producing the outcome than did interval training both when the probability and rates of the outcome were matched by within-participant matching.

Two properties of reward schedules

This brief review raises the issue of the critical feature that determines the relative sensitivity of ratio and interval performance to outcome revaluation. There are two properties that distinguish the contingencies. The first is that interval contingencies differentially reward pausing between responses or, in the operant conditioning jargon, long inter-response times (IRTs). Having performed a response, and collected the outcome if available, the longer that the agent waits before performing the next response, the more likely it is that the resources will have regenerated so that the next response will be rewarded with an outcome. Figure 1a illustrates the relationship between the seconds elapsed since the last response has been performed and the probability of the next response being rewarded for different parameters of a random interval (RI) schedule under which there is fixed probability of an outcome becoming available in each second. As can be appreciated, the probability of reinforcement increases monotonically with the time between responses, with faster increases with shorter programmed intervals between rewards. In contrast, since the ratio between responses and outcomes required under a ratio contingency establishes a fixed probability of reward which is independent of the time elapsed since the last response, this probability is independent of the pause to the next response[1].

It is unlikely, however, that this feature of interval contingencies reduces sensitivity to outcome revaluation because when an animal is trained with a choice between with two interval sources yielding different outcomes as opposed to a single interval source, performance is highly sensitive to outcome devaluation. Kosaki and Dickinson trained their hungry rats with a choice between pressing two levers (group *choice*), one yielding grain pellets and the other a sugar solution, both on interval schedules (Kosaki and Dickinson,

[1] Although it can be argued that some patterns of responding under ratio training can differentially reinforce short IRTs—for example because of the development of response bursting—our assumption in this paper will be that responding

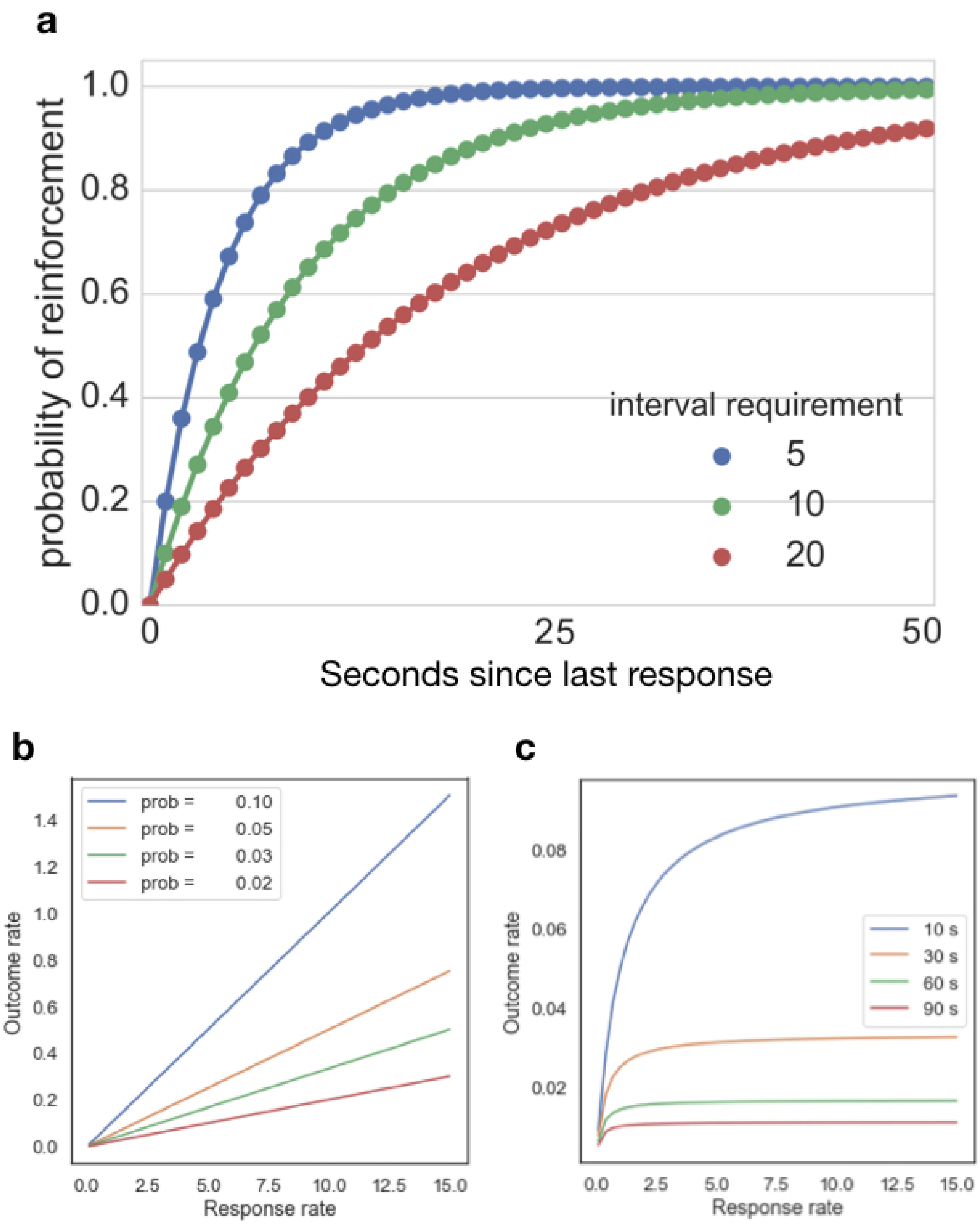

Figure 1. Different properties of response-outcome reward schedules. (a) Probability of obtaining and outcome after a pause between responses for different programmed inter-reinforcement intervals under an interval schedule. (b) Functional relationship between response rates and outcome rates for ratio schedules with different outcome probabilities $(1/ratio)$. (c) Functional relationship between response rate and outcome rates for interval schedules under different interval parameters (or inter-reinforcement intervals).

2010). In spite of this interval training, devaluing one of the outcomes reduced performance of the corresponding response on test even when there was only a single lever present during the test so that no choice was available at that time. This goal-directed control contrasted with the insensitivity to outcome devaluation following matched training with a single response. The second, *non-contingent* group of rats was trained with only a single lever present so that pressing yielded one of the outcomes on the interval schedule with the other being delivered at the same rate but independently, or non-contingently of the instrumental response. In contrast to the goal-directed control observed following choice training, lever pressing during the test was unaffected by whether the contingent or non-contingent outcome had been devalued. As the target responses were both trained under identical interval schedules, both of which should differentially reinforce long IRTs, it not clear why choice versus single response training should affect the degree of goal-directed control if IRT reinforcement is the critical factor affecting sensitivity to outcome revaluation under interval schedules.

The second distinction between ratio and interval contingencies relates to their response-outcome rate feedback functions, which are mathematical descriptions of the empirical relationship between response rates and outcome rates (Baum, 1973; Baum, 1992; Soto *et al.*, 2006). Figure 1b presents the feedback functions for typical ratio and interval schedules. Under a ratio contingency, the outcome rate rises linearly with increasing response rate, with the slope of the function decreasing systematically as the ratio parameter increases. The feedback function for ratio schedules can be described by a linear function of the form $Y = nB$, where Y is the outcome rate and B the response rate performed by the agent. The parameter n represents the inverse of the ratio requirement, or, equivalently, the outcome probability per response that the particular ratio schedule programs. By contrast, the feedback function for an interval schedule is nonlinear, with the outcome rate rising rapidly with increases in response rates when the baseline response rate is low and reaching an asymptote as soon as the response rate is higher than the rate at which the outcomes become available (Baum, 1992; Prelec, 1982). At this point, variations in response rates do not have an effect in the outcome rate [2].

In his correlational version of the Law of Effect, Baum (1973) suggested that the difference between the ratio and interval feedback functions can be captured by the linear correlation between the response and outcome rates established by the schedules, which in turn led Dickinson (1985; see also Dickinson and Perez, 2018) to argue that response-outcome learning is driven by the rate correlation experienced by the agent: the greater the experienced rate correlation, the stronger is the response-outcome learning.

Rate Correlation Theory

Baum (1973) illustrated the empirical application of his approach to the Law of Effect by dividing the time-line in an experimental session into a number of successive time samples and displayed the rate correlation by plotting the number of responses in each sample against the number of outcomes in that sample. In the present approach, however, we develop rate correlation theory in terms of psychological processing and assume that the agent computes the rate correlation at a given point in time by reference to the contents of a number of immediately prior samples of responses and outcomes held in memory.

Figure 2a illustrates a schematic representation of the time-line divided into different samples in memory of our model. At the end of each cycle, the number of responses and outcomes in that sample is registered in memory and the content of the memory is recycled. Given that the memory has a limited capacity, for simplicity we assume that this recycling

[2] Although the exact analytic form of the feedback function for interval schedules is still a matter of debate (see Baum, 1992), it is well accepted that this function needs to flatten once response rates attain a sufficiently high level, which depends on the outcome rate programmed by the schedule. A widely-accepted form of this function is $Y = \frac{B}{tB+a}$, where t is the interval parameter and a is a parameter that depends on the conditions of the experiment, and which affect the animal's pattern of responding independently of the outcome rate generated by the schedule.

involves not only the registration of the contents of the next sample but also the erasure of the oldest sample in memory. Figure 2a displays a memory of four samples. The initial memory cycle involves the first four samples, the second memory cycle involves the second to fifth samples, and so on. In general, cycle k involves the deployment of the contents of memory from samples $S_k, S_{k+1}, ..., S_{k+(n-1)}$, where n is the memory size deployed by the agent. In the following simulations, we assume that the memory size is the same for all subjects.

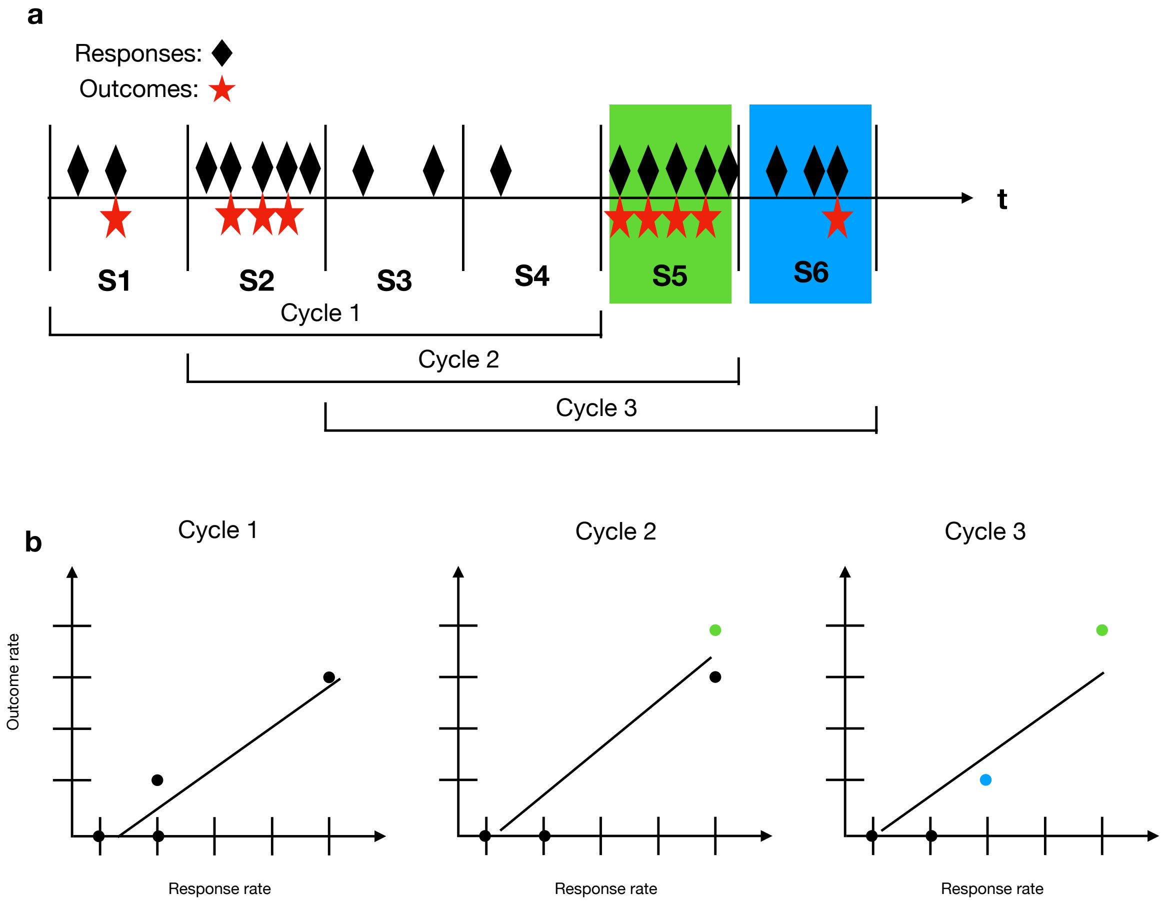

Figure 2. Memory model for a rate-correlation approach to instrumental actions. (a) In this simplified illustration, each memory cycle is comprised by four time-samples. The romboids represent response events and the outcomes are represented by red stars. (b) Different experienced rate correlations for each of the memory cycles exemplified in (a).

Following each mnemonic recycle, we assume that the agent estimates the response-outcome rate correlation based upon the current contents of the memory. For simplicity, we assume that the agent computes a standard correlation coefficient which, psychologically speaking, accounts for the agent's experienced linear relationship between the action and outcome rates. More formally, if b_i and r_i represent, respectively, the number of responses and outcomes in the $i-th$ sample in memory, then each sample can be understood as an ordered-pair $(b_i, r_i), i = 1, ..., n$, from which the agent computes the rate correlation by the following expression:

$$r_{br} = \sum_{i=1}^{m} \frac{(b_i - \bar{b})(r_i - \bar{r})}{m s_b s_r} = \frac{cov(b, r)}{s_b s_r} \tag{1}$$

where $cov(b, r)$ is the covariance between b and r, $\bar{b}$ and $\bar{r}$ the average responses and outcomes per sample, and s_b and s_r the standard deviations of b and r, respectively.

Let k be the current memory cycle and let g_k the strength of the rate correlation system in each cycle. The simplest model would assume that response strength during the following cycle $k + 1$ is determined in this system by the rate correlation computed on the basis of the memory contents at the last cycle, that is, $g_{k+1} = f(r_k) = r_k$. However, there are two concerns about this simple algorithm. First, the algorithm is sensitive solely to the currently experienced rate correlation and so gives no weight to prior experience.

Second, and most importantly, if the memory contains no events, either outcomes or responses at a cycle, the rate correlation is undefined. Under these circumstances, it would seem reasonable to assume that the agent needs to rely on its prior experience to determine responding in the current cycle. To determine the rate of responding, we assume that each cycle in the past has an effect on the current level of responding, with the effect being discounted with time. A typical function representing the discounting for previous cycles is given by $\theta = \lambda e^{-\lambda d}$, which assigns the importance to the cycles according to how far back they are in time (d). For a given value for d, different values of λ will yield different weights to the cycles. We use the discrete version of this function (Killeen, 1994). According to an exponential weighted moving average (EWMA) model, the agents compute the rate-correlation for each cycle and uses this value to generate responding according to

$$g_{k+1} = \theta r_k + (1 - \theta)\bar{r_k} \qquad (2)$$

where $\bar{r}$ is the average experienced rate correlation across the previous $k - 1$ cycles, computed in each memory cycle k as

$$\bar{r}_k = \bar{r}_{k-1} + \beta(r_k - \bar{r}_{k-1}) \qquad (k > 3) \qquad (3)$$

where $\beta = 1/k$ is the learning rate in each cycle and $\bar{r}_1 = r_1$, by definition. The parameter θ is a weighting parameter that represents the importance of the current rate correlation on responding for the next cycle. If $\theta = 1$, all the weight is put on the current cycle; if $\theta = 0$, responding is driven only by the average experienced rate correlation; other values of θ will give different degrees of importance to the history of rate correlation on current performance.

Simulations of a rate-correlation theory

We first investigated the robustness of a correlation coefficient in this model with respect to variations in the sample duration parameter. To this end, we probed the effect of varying the sample duration between 10 and 120 s on the rate correlation generated by random ratio (RR) 5-to-50 and RI 5-to-90 s schedules with response rates varying between 30 and 150 responses per minute. These two types of schedules assign, respectively, a probability of an outcome being delivered for each response and a probability of the outcome becoming available in each second. Once the outcome was available, it remained so until collected by a response. Those simulations showed that, for the range of response rates that we tested, the experienced rate correlation is not significantly affected by the size of the sample deployed by subjects. So we choose a value of 20 s for the time samples in a memory cycle, primarily to limit the total duration of the agents' memory to a few minutes. But it is important to note that using different sample lengths will not affect the results reported in this paper. As the simulations were run with a memory size of 20, the total memory duration was 400 s. For simplicity, we also limited the agent to perform a maximum of 60 responses per min (i.e. a maximum of 1 response per second) by arranging for the probability of a response in each second to be g. In what follows, we show the results for the EWMA model with theta set at .5, but note that the same results hold for the other values of θ tested in our simulations (see Supplemental Material).

Ratio-interval effects

We investigated the rate correlation model by running simulations under variations in outcome probability using RR schedules and variations of outcome rate using RI schedules. Our initial reason for investigating the role of rate correlation in goal-directed learning arose from the fact that ratio schedules establish responding that is more sensitive to outcome devaluation than does interval training even when the outcome probability is matched by yoking (Dickinson *et al.*, 1983). Within our rate correlation theory, g is the agent's learned representation of the strength of the causal relationship between action and outcome. However, as g also determines the probability of responding, the theory predicts concordance between judgments of the strength of the response-outcome relationship and the rate of responding.

The most direct evidence for such concordance comes from a study by Reed (2001), who reported the performance of human participants on ratio and interval schedules with matched outcome probabilities. Not only did he find that ratio training yielded higher causal judgments of the effectiveness of the action but also higher response rates, but also that the performance under ratio training was higher than under interval training. To investigate whether a rate-correlation model could reproduce these data, we simulated training on a master RI 20-s schedule, which was the temporal parameter employed by Reed (2001), and then used the outcome probability generated by each master subject to determine the parameter for a yoked subject trained under a ratio schedule. The initial response rate during the first cycle was 10 per min, and we trained the simulations across 3 sessions, each of which terminated after 13 outcomes, in an attempt to match the training received by participants in Reed's (2001) experiment. Figure 3 shows the data obtained by Reed (left panel) and the simulations produced by the rate-correlation model of the response strength, g, during the last 50 cycles and averaged across 100 replications of each simulation. As can be appreciated in the right panel of Figure 3, the model generated lower response-outcome rate correlation values following interval rather than ratio training with matched outcome probabilities.

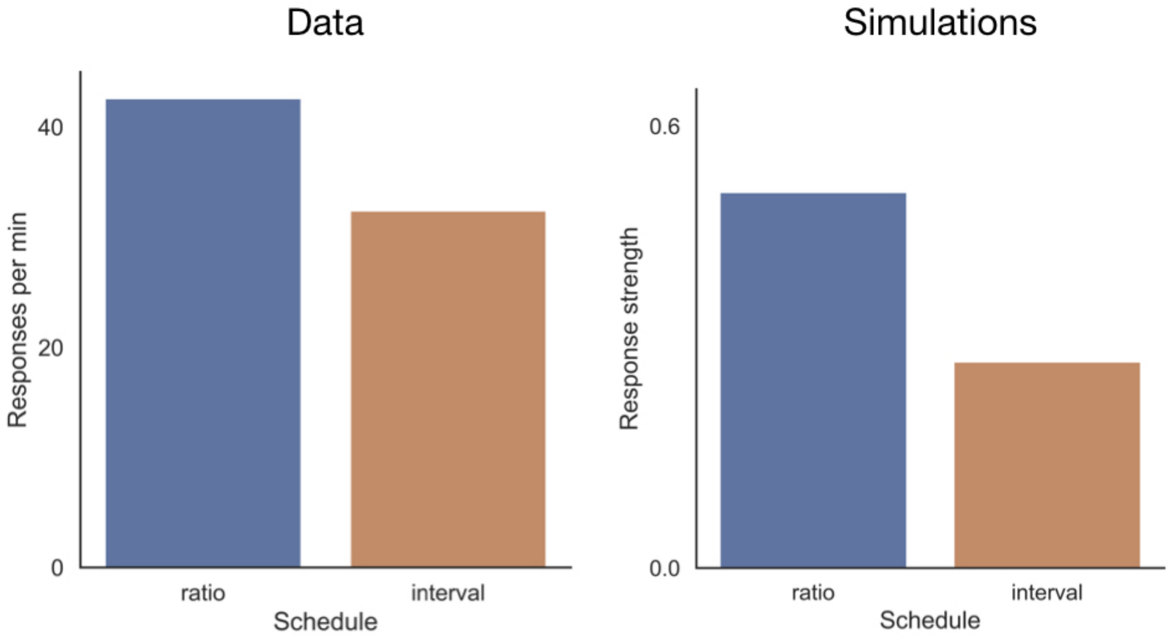

Figure 3. Simulations of a rate correlation model for ratio and interval schedules with matched reward probabilities. The left panel shows the data obtained by Reed (2001) in a human causal judgment experiment. The right hand panel show the simulations produced by a rate correlation model.

Outcome probability

Having established that rate correlation theory can reproduce the ratio-interval difference, we investigated whether the theory could simulate the general effects of the major variables determining free-operant performance. From these simulations we report the response strength, g, during the last 50 cycles from the 2000 cycles of each simulation averaged across 100 replications of each simulation.

We have already noted that both associative and model-based RL theories of goal-directed behavior predict that instrumental performance should be determined—either because an outcome follows from its execution or because its value is determined by reward prediction-error—by the outcome probability (Mackintosh and Dickinson, 1979; Sutton and Barto, 1998). This prediction was confirmed empirically by Mazur (1983), who trained hungry rats to press a lever on a RR schedule under different ratio requirements. To ensure that the motivational state was kept relatively constant, Mazur scheduled a limited number of food outcomes per session in an open economy [3]. To assess performance only during periods of engagement in the instrumental action, he also removed the outcome handling time by assessing the rate following the first lever press after an outcome delivery. The left panel of Figure 4 shows a relevant selection of the response rates obtain by Mazur.

To investigate the response rates generated by a rate correlation model when the outcome probability was varied, we replicated a similar design by simulating performance on RR schedules with ratio requirements varying between 10 and 30. Figure 4 shows that the likelihood of responding decreased systematically when outcome probability was reduced by increasing the ratio parameter, correctly predicting the pattern of results obtained by Mazur in his parametric investigation of ratio performance in rats.

[3] In an open economy, the animal is also fed in the home cage with a different food to the one earned by the instrumental response during training, so that its weight remains constant throughout the experiment.

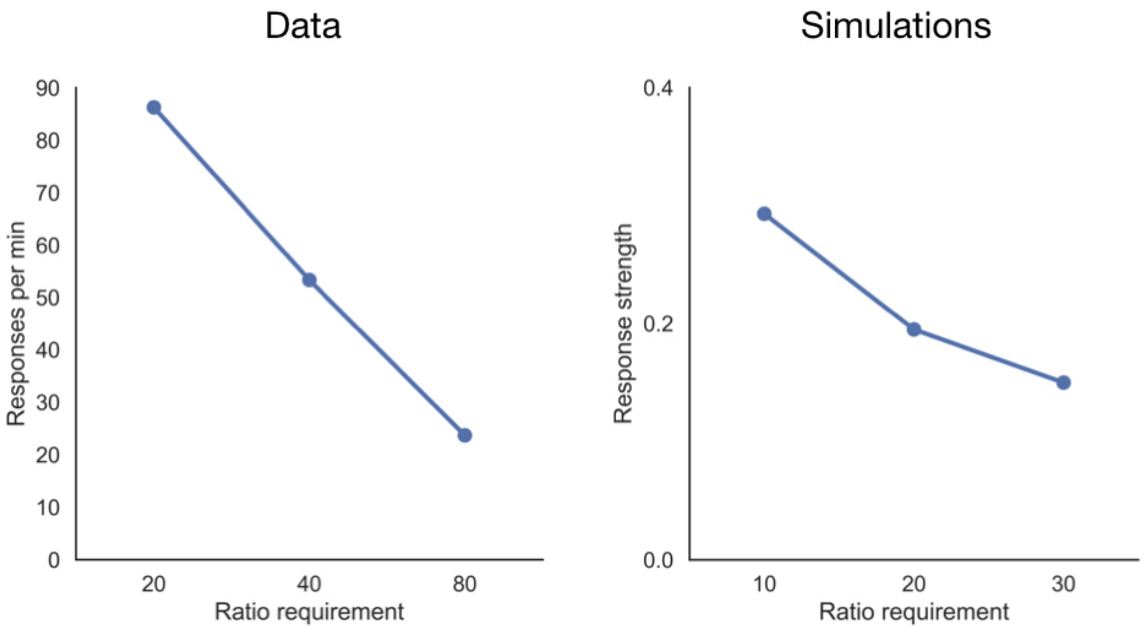

Figure 4. Simulations of rate correlation models for ratio training with different outcome probabilities (the inverse of the ratio requirement). The left panel shows the results obtained by Mazur (1983) in a within-subject study in rats. The right panel shows the simulations of a rate correlation model.

Outcome rate

Herrnstein and his colleagues have argued that instrumental performance on interval schedules is systematically related to the outcome rate, such that longer intervals between reinforcers should bring about lower performance than shorter ones (Herrnstein, 1969; Herrnstein, 1970). This prediction has been confirmed multiple times in different species. One example, shown in the left panel of Figure 5, was provided by Bradshaw et al. (1981), who trained hungry rats to lever press for milk and reported that there was a systematic decrease in the response rates as the interval was increased except at high rates of rewards when outcome handling time may well have interfered with lever pressing. A selection of their results for intermediate intervals are shown in the left panel of Figure 5. To match the conditions of this experiment, we repeated the simulation procedure used for outcome probability but with RI schedules and interval parameters varying between 30 and 90 s. As

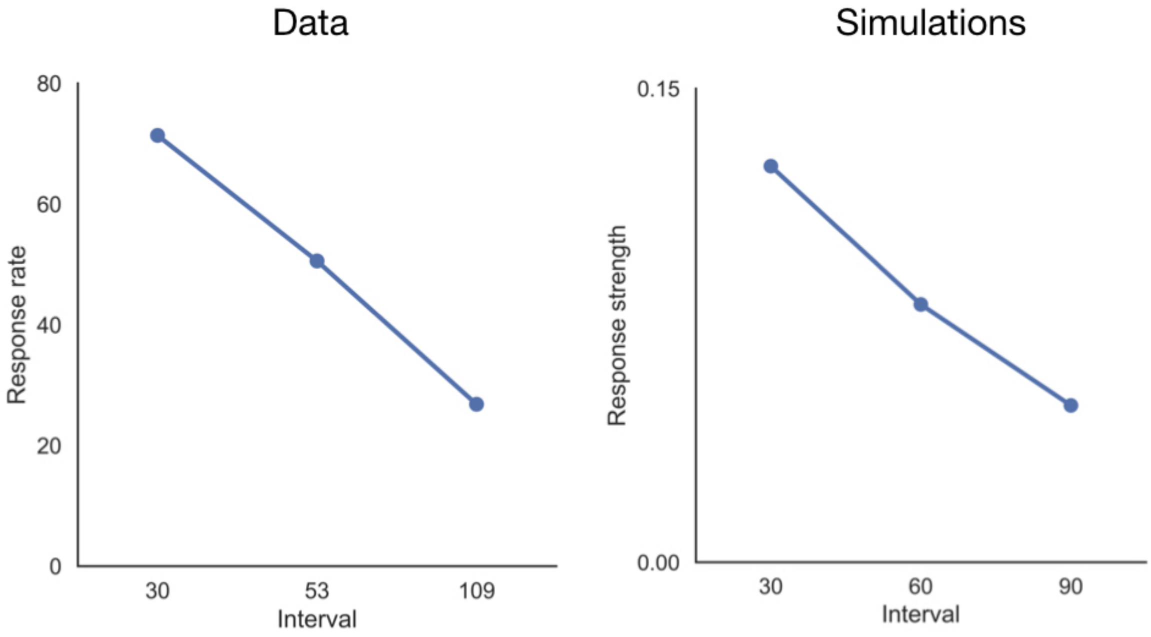

Figure 5. Simulations of a rate correlation model for interval schedules with different interval parameters. The left panel shows the results obtained by Bradshaw et al. (1981) in rats. The right panel shows simulations of a rate correlation model using parameters similar to the ones used by these authors.

the right panel of Figure 5 shows, all values produced a systematic decrease in responding as the outcome rate was reduced by increasing the temporal parameter of the interval schedule, replicating the pattern of results obtained by these authors.

Outcome delay

Baum (1973) noted that his correlational Law of Effect anticipated the fact that delaying the outcome following the response that generated it will have a deleterious impact on the acquisition of instrumental responding. For example, the left panel of Figure 6 illustrates the terminal rates of lever pressing by hungry rats obtained by Dickinson et al. when each lever press produced a food outcome after a delay of 16, 32, or 64 s (Dickinson *et al.*, 1992). With the 16-s delay and a 20-s memory sample used in our model, only outcomes generated by responses during the first 4 s of a sample occur in the same sample as their responses, whereas with the 32-s and 64-s delays all the outcomes occur in a different sample, thereby reducing the experienced rate correlation. The simulations displayed in the right panel of Figure 6 confirm this intuitive prediction. Following a similar reasoning to that of causal judgments for ratio- and interval-trained responses, these simulations anticipate a similar result for the acquisition of a causal belief when outcomes are delayed. This prediction has been confirmed by Shanks and Dickinson (1991) using fictitious credits as the outcome and key presses as the instrumental response in human participants. Moreover, the impact of outcome delay on goal-directed behavior has been more recently confirmed by Urcelay and Jonkman (2019), who reported that delaying the food outcome by 20 s abolished sensitivity to outcome devaluation compared to a group that underwent training with no delay between the response and the outcome.

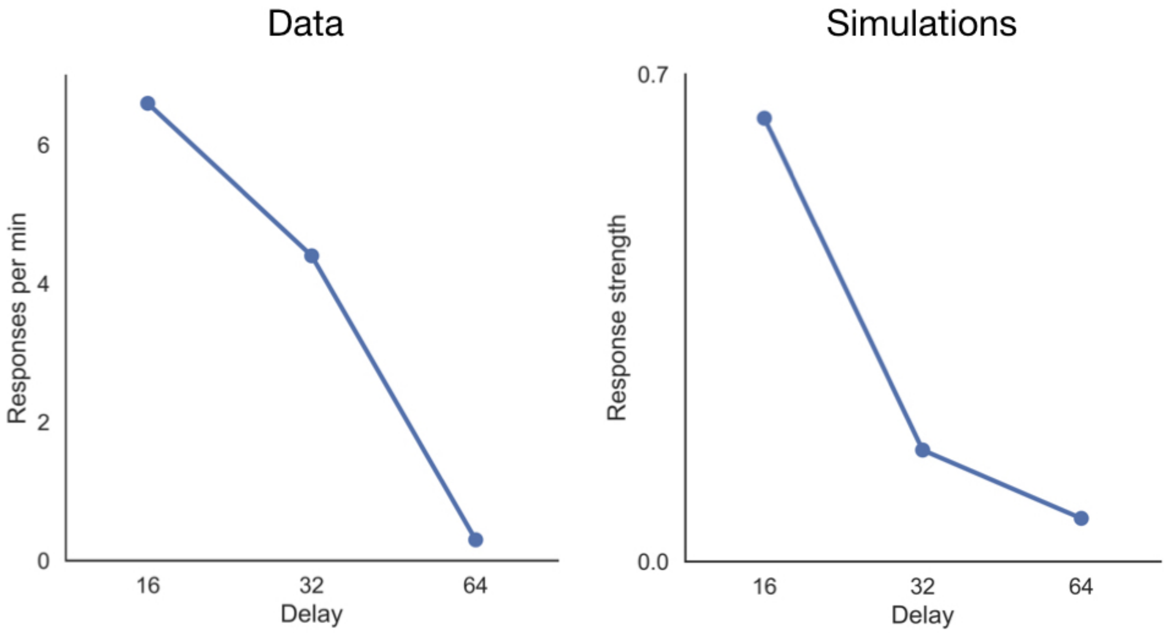

Figure 6. Simulations of rate correlation models for delayed rewards. The left panel shows the data obtained by Dickinson et al. (1992) in rats. The right panel shows simulations of a rate correlation model for the same delay parameters used in the original paper.

Contingency degradation

At first sight, the most direct evidence for a rate correlation approach to instrumental learning is the sensitivity of free-operant performance to the action-outcome contingency, in that a correlation provides a measure of this contingency. However, the strength of the causal relationship between action and outcome can be varied not only by changing the probability of a contiguous outcome as in Mazur's (1983) experiment, but also by varying the likelihood that the outcome will occur in the absence of the action or, in other words, the probability of non-contiguous outcomes. When the contiguous and non-contiguous probabilities are the same, the agent has no control over the number of outcomes received in any given time period. Hammond (1980) was the first to study the effect of such manipulation in a free-operant procedure. Using rats, Hammond fixed the probability of a contiguous outcome for the first lever press in each second while varying the probability of delivering a non-contiguous outcome at the end of any second without a lever press. Non-contingent schedules, in which the contiguous and non-contiguous outcomes probabilities were the same, failed to sustain lever pressing initially established without the non-contiguous outcomes.

We cannot be certain, however, that the low rate of lever pressing under the non-contingent schedules was due to the absence of a causal relationship between this action and the outcome. Inevitably, the non-contingent schedule greatly increases the frequency of the outcome and therefore the time required to handle and process the outcome with the result that the depression of responding under a non-contingent outcome may have been due to interference with lever pressing by the enhanced outcome handling and processing. One way of addressing this issue is to use a non-contingent schedule while varying the identity of the contiguous and non-contiguous outcome. When the contiguous and non-contiguous outcomes are the same, the agent has control over neither the outcome frequency nor its identity. However, when the outcomes are different, the agent can control the type of outcomes they received. By responding, the agent can increase the relative frequency of the contiguous outcome.

To illustrate the simulation of contingency degradation by the rate correlation model, we followed an experiment reported by Balleine and Dickinson (1998). Hungry rats were initially trained to lever-press for one of two different food outcomes on an RR 20 schedule so that the probability of the contiguous outcome was .05. The instrumental contingency was then degraded by delivering a non-contiguous outcome with a probability of .05 in each second without a lever press. As the left panel of Figure 7 shows, the rats pressed at a higher rate if the non-contiguous and contiguous outcomes were different rather than the same. The right panel illustrates that the rate correlation model can replicate this effect on the assumption that different outcomes receive distinct representations in memory with separate response strengths being calculated for each outcome type. Numerous studies have shown that human causal judgments of the response-outcome association and the rate of responding are lower when the contingency between the response and the outcome is degraded by increasing the probability of non-contiguous outcomes (Shanks, 1991).

Interim summary

In summary, this set of simulations demonstrate that a rate-correlation model can in principle provide an account of primary determinants of instrumental performance: the impact of outcome probability, rate and delay on instrumental performance. In addition, the model correctly anticipates the ratio-interval schedule effect when the outcome probabilities are matched, and the effect of degrading the causal contingency between the response and the outcome, both of which are prerequisites for any theory of goal-directed control.

It is equally clear, however, that a further learning system is required for a complete account of instrumental behavior. To the extent that goal-direct learning is assigned to a rate correlation system, we are left with no account of sustained responding on an interval

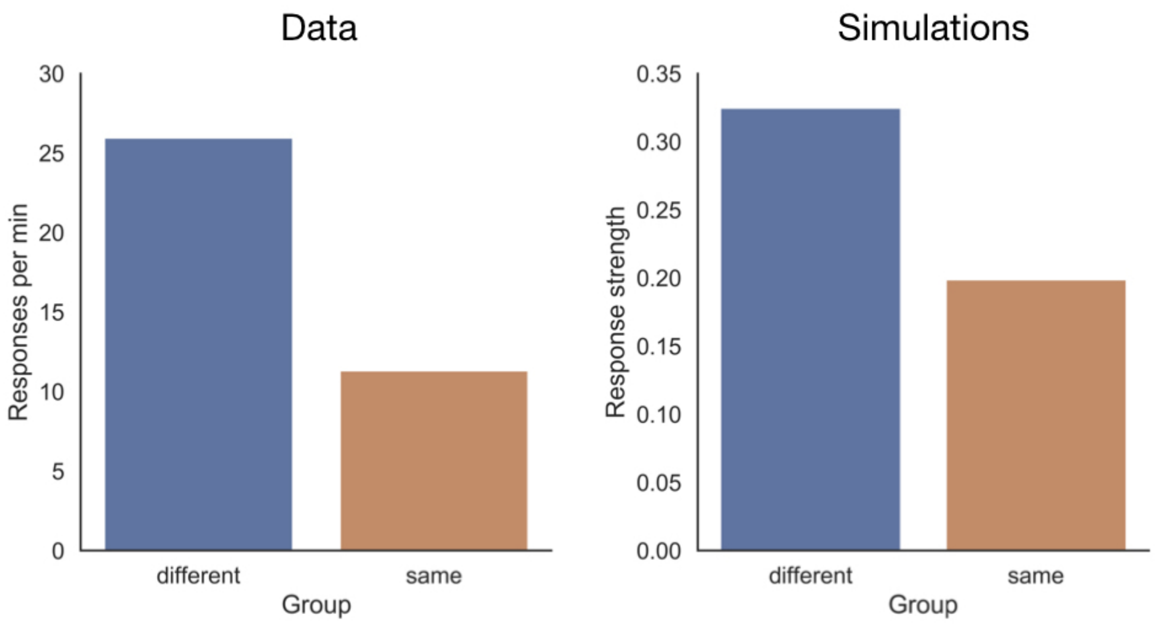

Figure 7. Simulations of a contingency degradation experiment. Left panel. Data obtained by Balleine and Dickinson (1998) in rats. Group *diff* was given freely an alternative outcome with the same probability as the outcome produced by the target action. Group *same* was given freely the same outcome as that produced by the target action. Right panel. Simulations of a rate-correlation model for a similar procedure.

schedule, given the low rate correlation experienced under this schedule. Furthermore, rate-correlation theory on its own provides no principled explanation of why responding extinguishes when outcomes are withheld. As we have noted, the rate correlation cannot be calculated at a recycle if no outcomes are represented in memory (as would be the case during extinction), and under this circumstance the response strength remains at the value computed at the last recycle in which the memory contained at least one outcome representation. A comprehensive account of instrumental action therefore requires an additional learning system.

Dual-System Theories

When Dickinson (1985) first argued that a rate correlation account of instrumental action could explain goal-directed learning, he embedded it within a dual-system theory to explain instrumental responding that is autonomous of the current value of the outcome, as assessed by the outcome revaluation paradigm. He envisaged this second system as a form of habit learning that involved the acquisition of an association between the stimuli present during training and the instrumental response. This is, of course, the form of stimulus-response (S-R) learning envisaged by Thorndike in his original Law of Effect (Thorndike, 1911) more than a century ago. According to Thorndike, the occurrence of a contiguous attractive outcome following a response simply serves to strengthen or reinforce the S-R association so that the re-presentation of the training stimuli are more likely to elicit the response. However, because all information about the outcome is discarded once it has served its reinforcing function, any subsequent change in the value of the outcome cannot impact on instrumental performance without re-presenting the revalued outcome contingent upon responding. For this reason, to test whether an outcome representation exerts goal-directed control over responding, the outcome devaluation paradigm tests responding in the same training context but in the absence of the now-devalued outcome. Any decrease in responding under these conditions indicates that a representation of the outcome controls an action in accord with the current value of the outcome, thereby demonstrating its goal-directed status (Balleine and Dickinson, 1998; Dickinson and Balleine, 1993; Dickinson and Perez, 2018).

To date, only RL theory has attempted to offer a computational account using a similar dual-system view of instrumental control. RL theory recognizes two types of systems that closely resemble the two psychological processes described by Dickinson (1985) in his original dual-system framework. Both RL systems aim to maximize the number of rewards obtained by the agent during a task (Daw *et al.*, 2005; Dolan and Dayan, 2013; Keramati *et al.*, 2011). Model-based (MB) computations learn a model of the environment by estimating the probability that an action in the current state will lead to each following state, and the probability of each action leading to a reward in each state. This "forward-looking" control is based on the online estimation of different state trajectories and is therefore highly sensitive to abrupt changes in either the response-outcome contingencies or the motivational value of the outcome, and therefore resembles a goal-directed system as proposed by Dickinson and his colleagues.

RL theory also recognizes another system that is relatively impervious to outcome revaluation. Model-free (MF) computations estimate the value of each action in each state ($Q(action|state)$) by simply caching the running average rate of rewards obtained by each action in a given state adjusting their value by reward-prediction error. Because all the history of rewards is collapsed in $Q(action|state)$, the agent maximizes the outcome rate by simply selecting the actions with a higher $Q-$value. For this reason, MF computations are less computationally expensive and faster than MB computations. When an outcome is revalued, however, the MF computations can only adjust to outcome revaluation by re-experiencing the outcome as a contingent consequence of an action so that, in this important respect, the behavioral control exerted by a MF RL system is similar to habitual behavior.

Because the estimations in MB and MF computations are updated by state and reward prediction-errors, respectively, the value of actions, and hence the probability of performing an action are ultimately determined by outcome probability. To capture the distinction between different reward schedules, RL needs significant modifications.

To our knowledge, only a model proposed by Niv et al. (Niv *et al.*, 2006) explicitly addresses free-operant performance within a RL framework. The normative approach proposed by Niv and colleagues (2005; 2007) distinguishes between the ratio and interval contingencies by deploying an economic argument that determines the rate of responding on the basis of the trade-off between the utility of obtaining more outcomes by responding faster and the cost of emitting those responses. This aim is achieved by choosing a behavioral strategy that obtains the most outcomes with the least effort. Such a point is reached when the marginal utility of increasing responding equals the marginal cost of such increase (i.e., waiting, or performing other behavior), a point that is reached at a lower response rate on an interval as opposed to a ratio schedule. Critically, however, this account is a form of MF RL and therefore provides no explanation of the differential sensitivity of ratio and interval responding to outcome revaluation, which is the focus of our analysis[4].

In the following sections we formalize a dual-system model in which a goal-directed system based on the response-outcome rate-correlation interacts with a habit MF algorithm based on reward prediction-error. We show how this model can explain all the phenomena we have already noted, along with additional phenomena from the literature that are not currently fully captured by RL or associative models of instrumental learning.

The Dual-System Model

Having demonstrated that a goal-directed system based on rate correlation can capture the primary determinants of free-operant behavior, we now specify a habit algorithm that will integrate with the goal-directed system to explain both behavioral performance and control in free-operant training. To this end, we employ an algorithm similar to those employed in the RL literature to account for MF strategies, but modified so that it can account for free-operant data (Bush and Mosteller, 1951). The algorithm deploys a reward prediction-error to increase or decrease the likelihood of performing the response in a similar situation or context. Let h_t denote habit strength at each time-step t. In our habit system, the acquisition and extinction of habit strength in cycle k follows the following equation:

$$h_{t+1} = \begin{cases} h_t + \alpha^+ PE_t & \text{if } PE_t > 0 \\ h_t - \alpha^- PE_t & \text{if } PE_t < 0 \end{cases} \tag{4}$$

[4] An exception to this is a recent model by Miller et al. (2019). Although their model can predict different sensitivities to devaluation for ratio and interval training, this is only achieved by importing arbitrary assumptions rather than providing an account embedded within an integrated model.

where α^+ and α^- are parameters between 0 and 1 and represent the learning rates for excitation and inhibition of the S-R connection, respectively [5] and PE_t is the reward prediction-error at time-step t, defined as:

$$PE_t = \begin{cases} 1 - (h_t + g_k) & \text{if response is reinforced} \\ (h_t + g_k) & \text{if response is not reinforced} \end{cases} \tag{5}$$

Following on evidence showing that learning rates for rewarded and non-rewarded episodes are asymmetric (Behrens *et al.*, 2007; Gershman, 2015; Lefebvre *et al.*, 2017; Palminteri *et al.*, 2017), we assume that the learning rate of a reinforced response is higher than the learning rate for a non-reinforced response ($\alpha^+ > \alpha^-$). This assumption is also necessary from a practical perspective: in a partial reinforcement schedule as the ones we have been simulating, the reinforced connection must be counteracting the effect of a much greater proportion of non-reinforced responses to sustain positive levels of responding. Under this algorithm, every reinforced episode strengthens the connection between the context and the instrumental response when the reward prediction-error, given by $PE_t = \alpha^+[1 - (h_t + g_k)]$ is positive. Likewise, every non-reinforced episode weakens the strength by $PE_t = \alpha^-(h_t + g_k)$ [6].

Similar to MF algorithms which assign the value $Q(action|state)$ to a specific action in a given state, Equation 4 explains the change of response strength according to the value of

[5] Previous versions of this algorithm deployed only one connection for increasing and decreasing the probability of responding. The original RL algorithm postulated by Bush & Mosteller had the form $h_{t+1} = h_t + \alpha^+[1 - (h_t)] - \alpha^-(h_t)$ and assumed that $\alpha = 0$ when a response was not reinforced. The term $-\alpha^-(h_t)$ can thus be regarded as reflecting an inhibitory potential present both in reinforced and non-reinforced responses.

[6] It should be noted that the PEs employ a summed prediction term by combining the current response strengths generated by the goal-directed (g_k) and habit systems. The rationale for this summed prediction term lies with the fact that a PE is intended to capture the extent to which an outcome (or its omission) is surprising or unexpected with respect to the predictions from both systems. In this respect, the rationale for the summed PE is the same as that in the Rescorla-Wagner rule (1972) for determining associative strength in Pavlovian Learning.

h_t, which completely summarizes the history of reinforcement in that particular state or context. Given that this algorithm is only driven by PE_t, it does not explicitly model the information regarding the relationship between the response and the outcome or its current motivational value, making it insensitive to both outcome revaluation and manipulations of the causal relationship between and response and outcome. Such behavioral autonomy is the cardinal feature of habitual behavior (Dickinson, 1985; Heyes and Dawson, 1990).

Given the above specifications for the habit and goal-directed systems, the next step is to specify the type of interaction between these systems that would explain total performance and behavioral control for different experimental conditions. To this end, we define a response function that jointly deploys both processes to explain total response strength for each memory cycle k. We denote this total response strength by p_k.

Our assumption regarding the interaction between the systems will be based on the data reported by Dickinson et al. (1983). As noted above, after having trained two groups of rats under interval and ratio schedules with matched outcome probabilities, Dickinson and colleagues devalued the outcome in half of the rats of each group by pairing it with toxicosis. After this devaluation manipulation, only the ratio-trained rats decreased responding (i.e., were under goal-directed control); the performance of the interval-trained rats at test was unaffected by outcome devaluation. An interesting feature of these data is that the level of responding after devaluation in the ratio-trained group did not differ from that of the interval-trained group. Because the outcome probability was matched between the groups, the habit system's contribution to responding should have been equal in both groups. Likewise, because by definition responding that is sensitive to devaluation must be attributed to the goal-directed component, the residual responding that was not affected by devaluation in the ratio-trained group must, by necessity, be attributed to the habitual component. Therefore, this study suggests that both systems were summing their relative strengths to determine the response probability p.

To transform response strengths into probabilities of responding, we assume that response probability in cycle $k+1$, p_{k+1}, is governed by a sigmoid function:

$$p_{k+1} = s(Ig_k + h_k) = \frac{1}{1 + e^{-\tau(Ig_k + h_k - C)}} \tag{6}$$

where g_k is the goal-directed strength in cycle k as defined above, h_k is the habit strength accumulated by Equation 4 during the experiment, up to cycle k, and I is a variable representing the current incentive value of the outcome by taking the value 1 if the outcome is valued and 0 if the outcome is devalued (that is, we assume that the devaluation procedure successfully decreases the value of the outcome to zero). The parameter τ is an inverse temperature parameter that reflects how sensitive the agent is to increases in total response strength $(Ig_k + h_k)$) and C is a parameter that determines the midpoint value of the function. Under this response function, the two systems sum to determine total responding, so that the response probability in the next cycle p_{k+1} reflects the relative contribution of each system (see Figure 8). In the following sections, we will discuss the implications of such an assumption for the type of behavioral control that should be expected after we present simulations for different experimental procedures.

Ratio and interval training

Initially we simulated goal-directed and habitual learning under interval and ratio contingencies using a RI 15-s master schedule. The outcome probability generated by each master interval simulation was then used to generate a yoked simulation on a ratio schedule with a parameter that yielded to same outcome probability. The initial response probability for the first session of training reflected one session of pretraining under RR-5 and each session terminated after 30 outcomes had been received. Panels a and b of Figure 9 display the mean values generated by 200 simulations under the master interval and yoked ratio

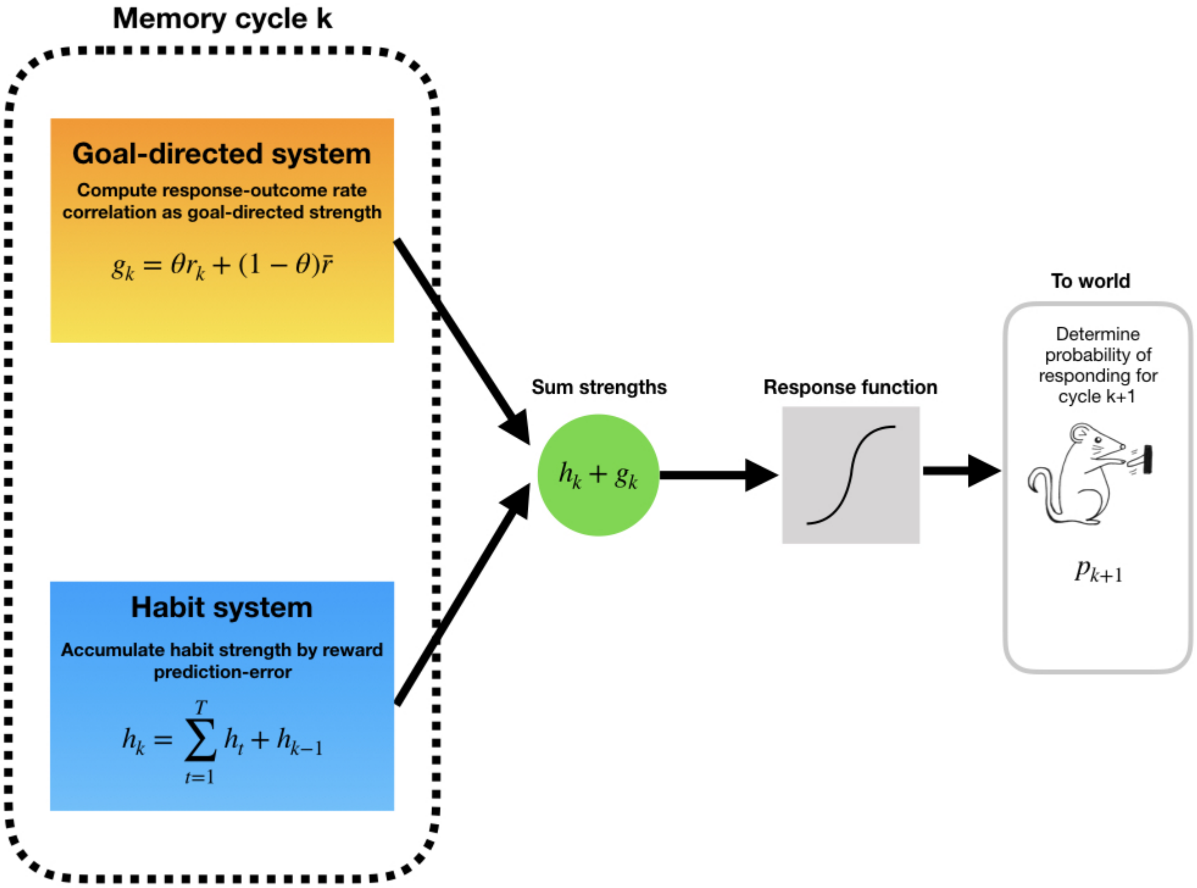
Memory cycle k
Goal-directed system
Compute response-outcome rate correlation as goal-directed strength
$g_k = \theta r_k + (1-\theta)\bar{r}$
Habit system
Accumulate habit strength by reward prediction-error
$h_k = \sum_{t=1}^{T} h_t + h_{k-1}$
Sum strengths
$h_k + g_k$
Response function
To world
Determine probability of responding for cycle k+1
p_{k+1}

Figure 8. Schematic representation of the dual-system model. For each cycle, the agents concurrently computes the response-outcome rate correlation and habit strength is accumulated. The strength of both systems is then summed and a response function produces the probability of responding for the following cycle. The rate correlation on the goal-directed system is only computed when both responses and outcomes are held in memory. (Illustrations courtesy of Loreto Contreras.)

schedules, respectively. Shown separately are the response strengths generated by goal-directed and habit systems, g and h respectively, and the resultant probability of responding per 1-s time sample, p, produced by the interaction of these response strengths.

The first point to note is that the model reproduces the differential sensitivity of ratio and interval performance to outcome devaluation early in training. For example, ratio training generates a goal-directed response strength, g, of about 0.4 by the third session, whereas the interval response strength is close to zero for equivalent training. As the model assumes that outcome devaluation, if complete, abolishes the contribution of g to overall responding, the model naturally explains why devaluation has a greater impact on ratio than on interval responding early in training (Dickinson *et al.*, 1983). This finding is summarized in Figure 9c in terms of a devaluation ratio, DR, defined as $DR = \frac{s(Ig)}{s(Ig+h)}$, where s is the sigmoid function defined in Equation 6.

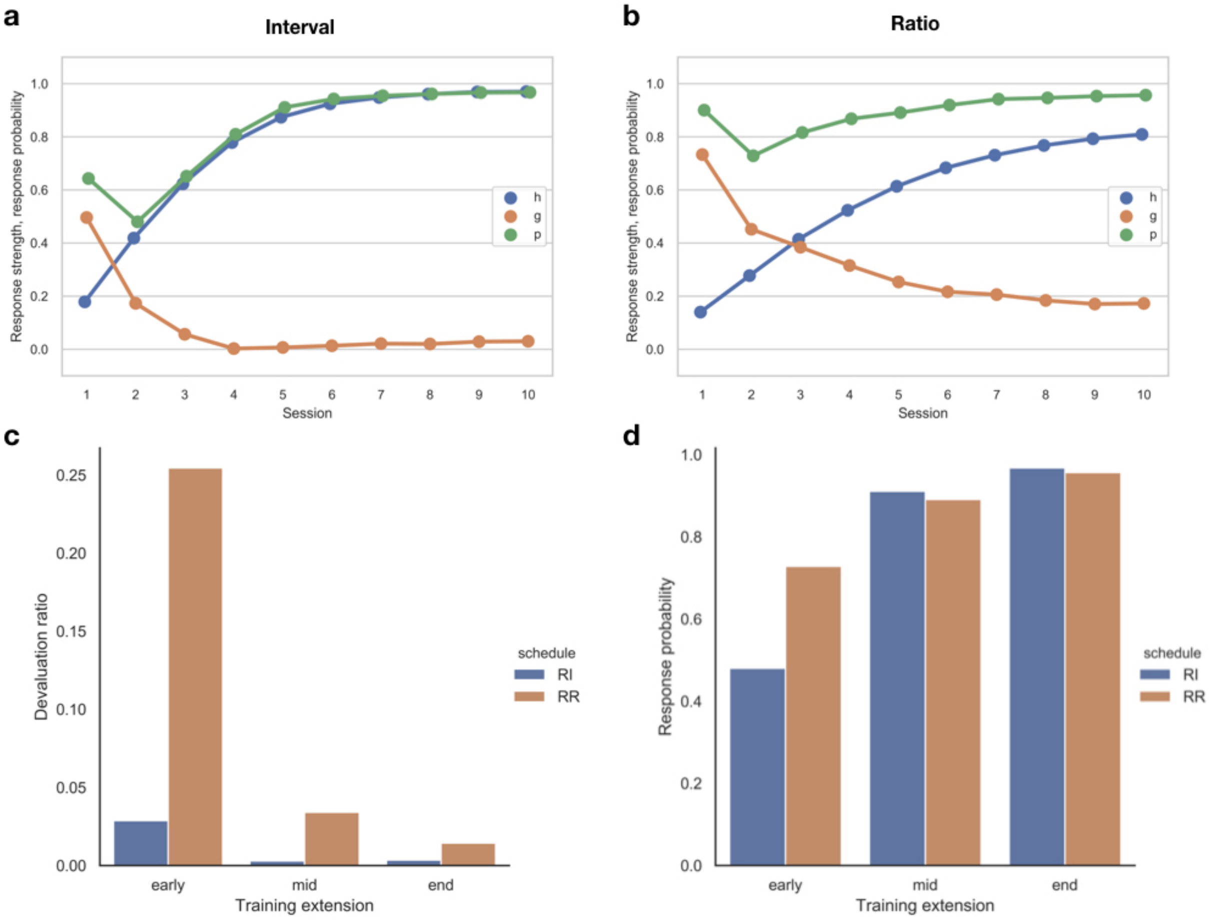

Figure 9. Simulations of the dual-system model for the experiment reported by Dickinson et al. (1983). (a) Strength from each system and response probability across 10 sessions of interval training. (b) Strength from each system and response probability across 10 sessions of training under yoked ratio training, matching outcome probabilities with the interval-trained subjects. (c) Sensitivity to outcome devaluation for ratio and interval training as assessed by a devaluation ratio early in training (Session 2); at mid-training (Session 5) and at the end of training (Session 10). (d) Response probability per second for ratio and interval training across different extensions of training.

Development of behavioral autonomy

Perhaps, however, the most notable feature of these simulations is the decline in the goal-directed response strength as the habit strength grows with training. This reduction in g reflects, at least in part, the reduction in the variance of the rate of responding across the time samples in memory as the overall response rate increases with the consequence that the experienced rate correlation, and therefore g, declines. Thus, according to our model, behavioral autonomy should develop as responding becomes stereotyped with more extended training. The reduction in sensitivity to outcome devaluation with training predicted by the simulations is documented in Figure 9c in terms of devaluation ratio.

Adams (1982) was the first to report that behavioral autonomy developed with training on a variety of ratio schedules. Although the development of autonomy with training has been independently replicated multiple times (e.g., Dickinson *et al.*, 1995; Holland, 2004; Killcross and Coutureau, 2003), a number of studies have reported goal-directed control after extended training. For example, de Wit et al. (2018) have documented two failures to replicate the development of behavioral autonomy observed by Tricomi et al. (2009) after training humans under an interval schedule (see Corbit *et al.*, 2014; Nelson and Killcross, 2006). Similarly, Jonkman et al. (2010) found that rats remained sensitive to outcome devaluation throughout 20 sessions of training on an interval schedule.

In interpreting these divergent results it is important to emphasize that it is not the type schedule (ratio versus interval) nor the amount of training per se that determines whether responding becomes behaviorally autonomous of the current outcome value, but rather whether the mechanism of memory recycling yields a low local rate correlation. For example, consider the case of extend training on fixed interval schedules (FI) in which an an outcome becomes available after a fixed interval between each obtained outcome. FI schedules have a similar overall functional relationship between response and outcome rates as variable interval schedules (see Figure 1c). However, two types of schedule generate very different local rate correlations as represented in a memory cycle of our model. The RI interval schedule establishes a steady rate of responding that, in conjunction with the temporal constraint on the outcome rate, ensures the rate correlation encoded in memory cycle is low. By contrast, a fixed schedule produces a sustained variation in the local rate of responding in the form of a "scalloped" pattern in which responding is low immediately after the receipt of an outcome before increasing as the availability of the next outcome approaches in time. As a consequent, the contrasting response rates within the interval ensures that the agent continues to experience a local rate correlation however much training is given. Importantly, this prediction accords with the report by DeRusso et al. (2010) who reported that extended training on a RI schedule established behavioral autonomy, whereas FI responding remained sensitive to outcome devaluation after equivalent training.

Choice training.

The analysis of extended makes clear that, according to rate correlational theory, the conditions for developing behavioral autonomy are not directly determined by the operant schedule or the amount of training but rather by whether or not the agent experiences a correlation between the rates of responding and outcomes as represented within the memory cycle. To recap, embedding rate correlational theory within a dual-system model predicts a reduction in the experienced rate correlation through the development of invariant stereotyped responding with the growth of habit strength, an effect enhanced in the case of interval schedules by the temporal control of outcome availability.

The cardinal importance of the experienced rate correlation is reinforced by the contrast between the single-response training, which has been our focus so far, and free-operant choice or concurrent training. It has long been known that responding remains sensitive to outcome devaluation when the training involves interleaved experience with two different response-outcome contingencies (Colwill and Rescorla, 1985; Colwill and Rescorla, 1988). However, of more directly relevance to the present analysis is the study by Kosaki and Dickinson (2010), which we have already discussed briefly with respect to the differential reinforcement of long IRTs, in that they directly compared behavioral autonomy after concurrent and single-response training.

To recap, Kosaki and Dickinson (2010) trained rats on two RI schedules that were concurrently active during each session of training. In one group, the *choice* group, responding on different levers produced different outcomes. Another group of rats, the *single-response* group, received the same two outcomes, only that in this group one of the outcomes was earned by responding on one lever, whereas the other outcome was delivered non-contingently after the same average period of time as the contingent outcome but independently of responding. After 20 sessions, a contingent reward was devalued in both groups by aversion conditioning and responding tested in a subsequent extinction session . Kosaki and Dickinson observed that responding in the single-response group was insensitive to devaluation, whereas the choice group markedly reduced the rate of the response whose outcome was devalued. There are two points to note about this finding. First, the devaluation effect was assessed against control conditions in which the other outcome was devalued. As a consequence, any effects of contextual conditioning on general performance was equated across conditions. Second, the same devaluation effects was found whether or not the choice was tested with both levers present or just a single lever. Thus, the devaluation effect exhibited by the choice group arose from the training rather testing conditions. In conclusion, these results demonstrated that responding in the choice group was still under goal-directed control even when similar training extension rendered responding habitual in the single-response group.

Recall that, according to the rate correlation component of our dual-system model, behavioral autonomy develops through extended training because responding becomes stereotyped with little variation across time-samples, thereby yielding a low rate correlation within a memory cycle, an effect compounded by the intrinsic low correlation engendered by an interval contingency. However, response rate variation across time-samples is an inevitable consequence when the agent is engaged with two interval sources of reward. When engaged in one of the sources, the memory samples will register neither responses nor outcomes from the non-engaged source. Consequently, any memory cycle containing a switch with have some samples with no response nor outcomes representations of the switched-to-source and other samples containing these representations. And, of course, the same will be true of the switched-from-source. As a consequence, the agent will experience a sustained rate correlation for both responses, each of which will therefore sustained goal directed control.

To substantiate this intuitive analysis, we simulated a concurrent choice procedure similar to that employed by Kosaki and Dickinson (2010) using our dual-system model. The simulations were run under the same conditions as the previous ones for interval training. It is well established that the probability of switching away from a source remains constant during responding to that source (Heyman, 1979) and so we programmed a fixed probability per 1-s time sample, p_{switch}, for a change-over between levers in the case of the *choice* group. Inspection of the authors' original data-set revealed that their rats switched between levers on average every 10 s; we therefore set $p_{switch} = .1$ for the following simulations.

Figure 10 shows the results of the simulation by the dual-system model for this choice experiment. As can be seen, similar amounts of training under a choice procedure yield significant contributions of the goal-directed system compared to single training. The result holds even when the amount of training is sufficient to drive the habit strength to asymptote, a factor that should reduced the experienced rate correlation, and hence goal directed control if only a single response was available. In summary, the model predicts that both systems should contribute to the control of responding under choice training, and therefore outcome devaluation should be effective in modulating responding under choice procedures, in line with the results reported by Kosaki and Dickinson (2010).

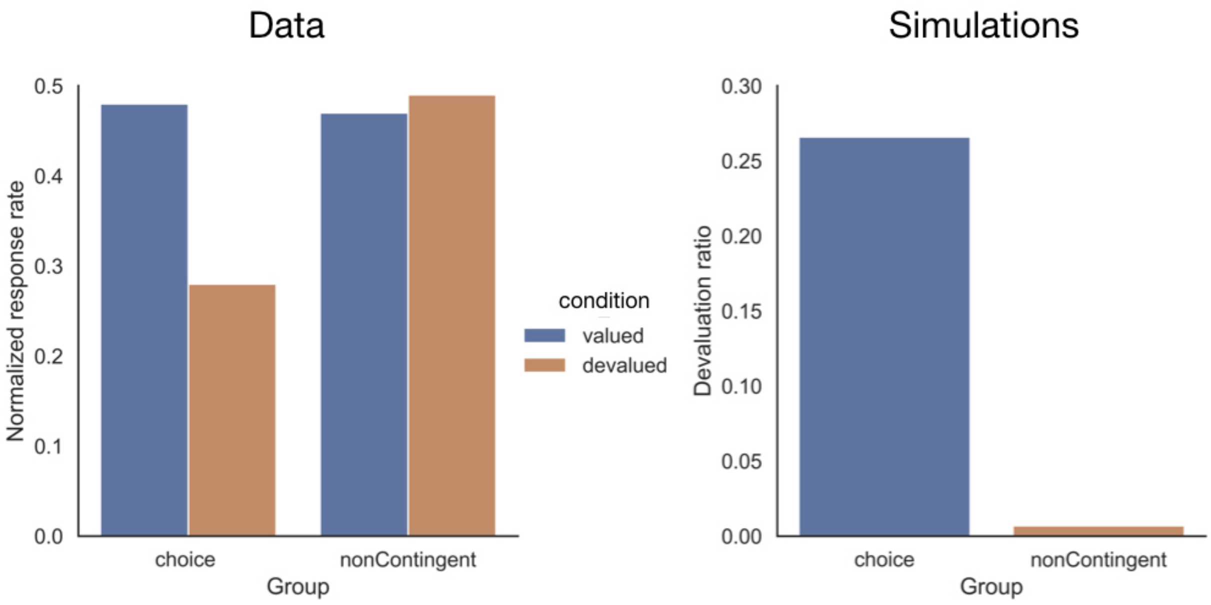

Figure 10. Simulations Kosaki and Dickinson (2010), investigating sensitivity to reward devaluation in a choice procedure. The choice group was trained with two responses concurrently available under an RI schedule.

One thing to note with regard to Kosaki & Dickinson's (2010) study is that the outcomes produced by each response differed in their sensory properties, which is critical if the dual-system model is to predict devaluation sensitivity after overtraining. Using the same outcome for each of the responses effectively changes the schedule into a non-contingent one for both responses because the outcome rate when the agent is responding to one source would be the same as that when response are not directed at that source. Hence, the rate correlation for this response should be close to zero with the consequence that responding under such a schedule should be purely habitual. Holland (2004, Experiment 2) conducted an experiment where the same training regime was given to two different groups of rats under interval schedules, with two different responses and outcomes available in one group, and with two responses producing the same outcome in another group. After extended training, only the rats in the group trained with multiple outcome was sensitive to devaluation; using a single outcome even when two responses were available made responding habitual, in line with the predictions of our model.

Extinction.

As it stands, the rate correlation system in our model makes what at first sight appears to be a highly problematic prediction: goal-direct control should never extinguish. Recall that the goal-directed system only computes the response-outcome rate correlation for memory cycles in which at least one response and one outcome are registered in memory. The consequence of this assumption is that goal-directed strength remains frozen throughout extinction at the the level attained during acquisition following the last memory cycle that contained an outcome representation.

Although not generally acknowledged by RL theory, this prediction accords with a series of studies conducted by Rescorla (1993), who reported that the impact of the outcome devaluation is not reduced by extinction. In one of his experiments, Rescorla trained two responses each with a different outcome, and then one of the responses was extinguished before a final devaluation test. Rescorla found that devaluation one of the original training rewards produced a comparable reduction in performance of the associated response in extinguished and non-extinguished conditions, thereby demonstrating that goal-directed learning survived the extinction phase. The left panel of Figure 11 presents the comparable outcome devaluation effect observed by Rescorla (1993) in the extinguished and non-extinguished conditions. It should be noted that the relatively high response rates in the extinguished condition reflects the fact that responding was reacquired with a third outcome prior to the devaluation test to ensure comparable response rates at test. If the goal-directed system remains relatively unaffected by extinction procedures, what would then explain the systematic decrease in responding observed across extinction sessions? One possibility, originally suggested by Colwill (1991), is that the habit system inhibits the goal-directed system during the extinction phase, masking the contribution that would otherwise be present during contingent response-outcome training.

Our dual-system model anticipates the acquisition of this inhibitory habit strength during extinction. In our model, the prediction error term in the habit algorithm includes the total prediction determined by the habit and goal-directed strengths g and h, respectively. Assume that cycle k is the last one containing response and outcome events in memory (the last cycle in training, in this example). If g retains a positive value during extinction, because the response-outcome rate correlation is not computed in a memory cycle that does not contain any outcomes, then $g_{k'} = g_k = g_0$ for all cycles k' in extinction. Then it follows that the prediction-error for h will be negative at each time step $(PE_t = -(h_t + g_0))$ and hence there will be a systematic decrease of h during extinction. The reductions in h will in turn decrease p with training, and responding will eventually extinguish. Indeed, for complete extinction to occur, the habit strength, h, will have to become negative or inhibitory because g remains constant and positive throughout the extinction phase. To simulate extinction, we initially trained our virtual rats as in previous simulations and then suspended outcome delivery for 2000 memory cycles. As can be appreciated in the right panel of Figure 11, the model correctly predicts a systematic decrease in total responding while maintaining a positive goal-directed strength, thereby providing an account of the retention of goal-directed control reported by Rescorla [7].

A different prediction in this regard can be made with respect to contingency degradation manipulations. Indeed, to reduce the goal-directed strength in our rate correlation system the agent would have to be transferred from a contingent to a non-contingent schedule in which outcomes occur independently of responding, and

[7] The decrease in p under the value of α^- chosen for previous simulations made p decrease at a low rate and remained at a positive and low value after 2000 memory cycles. Therefore, for illustrative purposes, in the simulations shown in Figure 11 we employed a higher value for α^- and kept everything else identical to previous simulations (see Supplemental Material for the parameters used in each of the simulations shown in the paper).

responding should become insensitive to devaluation under these conditions. As far as we know, the impact of non-contingent training on outcome devaluation has not been reported [8].

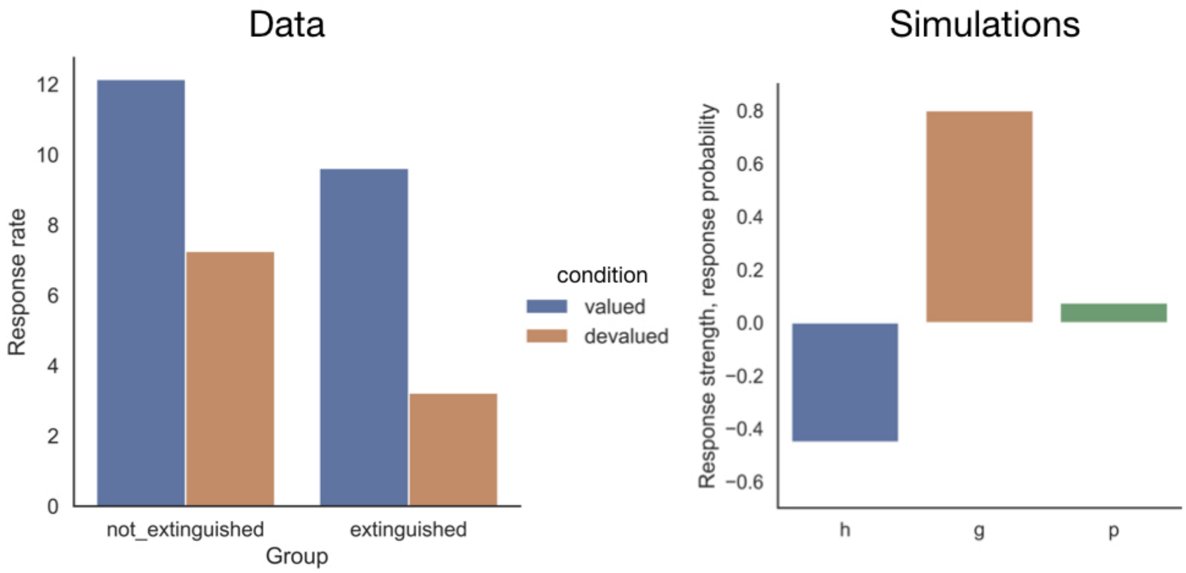

Figure 11. Simulations of a devaluation manipulation after extinction. The left panel shows the results reported by Rescorla (1993), which involved devaluation of one outcome for one response after an extinction phase compared with a response for which the outcome was not devalued. A control group had similar training but without undergoing an extinction phase. The right panel shows the final values after 2000 cycles of extinction for the dual-system model.

Additional phenomena and some outstanding issues

In spite of the wide range of phenomena that we have shown can be captured by this dual-system model, there still remain a number of outstanding issues that will need to be addressed in future refinements of the theory, and other phenomena that follow directly from the simulations presented in this paper. We discuss some of these below.

[8] Exposure to non-contingent outcomes does not reduce outcome-specific Pavlovian-instrumental transfer (Colwill, 2001; Rescorla, 1994), which is thought to be unaffected by outcome devaluation. However, transfer learning differs from that mediating goal-directed behavior (1994).

Performance after extended training

As Figure 9d clearly illustrates, the dual-system model predicts that a ratio schedule maintains a higher response rate than a comparable interval schedule early in training, as was originally observed by Dickinson et al. (1983). With more extended training, however, the difference in performance disappears as responding comes under habitual control. This prediction is clearly at variance with the sustained schedule effect on performance widely documented in the literature (e.g. Catania *et al.*, 1977). We have already noted that interval schedules differentially reinforce long IRTs—the longer an agent waits before responding again, the more likely it is that a further outcome has become available with the resultant increase the probability of reinforcement (see Figure 1a). To the extent that habit learning is conceived of as a form of stimulus-response learning, we should expect this form of learning to be sensitive to the temporal cues registering the time since the last response and to come under the control of these cues with the resulting impact on the rate of responding. By contrast, on a ratio schedule the probability of reinforcement is independent of the IRT and responding should be independent of the size of the emmited IRTs (for a discussion, see Chapter 4 in Mackintosh, 1974).

As the habit system does not incorporate a mechanism for the differential stimulus control of responding, we cannot use our model to assess impact of IRT reinforcement on responding. However, if this differential reinforcement could be removed while implementing the low rate correlation characteristic of interval contingencies, the model predicts that there should be no sustained ratio-interval performance effect. Kuch and Platt (1976) specified such a schedule, now referred to as a regulated-probability interval schedule (RPI). Without going into the implementation details, the RPI schedule sets the probability of reinforcement for the next response so that if the agent continues responding at the current rate, the rate of the outcome will match that specified by the scheduled interval parameter. As a consequence, variations in the rate of responding will have little impact on the obtained outcome rate so that the schedule maintains the low rate correlation characteristic of a standard interval schedule. However, as the outcome probability for the next response is fixed at the time of the preceding response, the RPI schedule, like a standard RR schedule, does not differentially reinforce any particular IRT. Consequently, our dual system model predicts that there should be no difference in the sustained responding on ratio and RPI schedules with matched outcome rates or probability.

The limited empirical evidence on this contrast is mixed. Neither Tanno and Sakagami (2008) nor Perez et al. (2018), who both trained hungry rats to lever-press for a food outcome, reported a sustained difference between responding on ratio and matched RPI schedules, while observing the reduced response on a standard matched interval schedule. In contrast, Dawson and Dickinson (1990) observed a higher response rate of chain pulling on a ratio schedule than on a yoked RPI schedule and, more recently, Perez and Soto (2019) have reported a similar result in humans. This remains an anomaly for our dual-system model.

Discriminative control

As it stands, our dual-system model offers no mechanism by which goal-directed responding can come under stimulus control as the goal-directed strength, g, is solely a product of the correlation between responses and outcomes. There is, however, extensive evidence such responding can come under discriminative control. The most compelling comes from an elegant biconditional discrimination studied by Colwill and Rescorla (1991). They trained rats with two different responses (R) and outcomes (O) and arranged for the different stimuli (S) to signal which outcome would be produced by each response. When S1 was present, R1 led to O1 and R2 to O2 whereas the opposite relation held when S2 was present (R1 led to O2 and R2 led to O1). When one of the outcomes was then devalued, rats responded more in the extinction test during the stimulus that during training signalled the non-devalued outcome for the target response. As this design equates the S-O associations across stimuli and the R-O association across responses, this devaluation effect requires the encoding of the triadic relationship between S, R and O, a representation that is not incorporated into our current formulation of rate correlation theory.

There is evidence, however, that goal-directed responding does not spontaneously come under the control of the stimulus context in which the response-outcome contingency is experienced. Thrailkill and Bouton (2015) found that after limited instrumental training the magnitude of the devaluation effect shown by their rats was unaffected by a shift from the training context to another familiar context between the end of instrumental training and testing. It is unlikely that their rats did not discriminate between the contexts because with more extended training, when responding had become autonomous of outcome value, this context shift reduced overall responding. This pattern of results accords with the idea that with limited training responding is predominantly under goal-directed control that encodes only the response-outcome relationship and, consequently, this control transfers spontaneously across contexts as anticipated by our current formulation of rate correlation theory. However, when responding has become under habitual control with more extended training, a context shift automatically produces a response decrement because such control reflects the development of context (stimulus)-response strength.

Motivational processes

Different processes are involved in the motivation of habits and goal-directed action and so we shall consider each in turn.

Motivating habits. Discriminative control, whereby a stimulus or context signals or "sets the occasion" for a response-outcome contingency, is not the only function by which stimuli and contexts impact upon free-operant responding. In accord with classic two-process theory (Rescorla and Solomon, 1967), it is well established that Pavlovian stimuli associated with appetitive reinforcers motivate the performance of free-operant behavior reinforced with an appetitive outcome. Estes (1948) was the first to demonstrate this effect using what has come to be called the Pavlovian-instrumental transfer (PIT) effect. He initially established a Pavlovian stimulus as a signal for food before training his hungry rats to press a lever for the food. When he then presented the stimulus for the first time while the rats were lever-pressing, he observed an increase in response rate during the stimulus. Given this transfer, two-process theory assumes that the Pavlovian conditioning to contextual cues occurs concurrently with instrumental learning during standard operant training so that the context comes to exert a motivational influence on free-operant performance.

The concordance between the impact of outcome rate on operant performance and Pavlovian responding accords with this two-process theory of instrumental motivation. It has long been recognized that an important variable in determining the rate of responding on interval schedules is the outcome rate rather than the outcome probability per response, and Killeen (1982; 1978) proposed that outcome rate has a direct motivational impact, so that higher outcome rates will have a general and sustained energizing effect on behavior. Indeed, this effect has been formalized by Herrnstein and collegues (Villiers and Herrnstein, 1976) in terms of a hyperbolic function between response and reinforcement rates and, more recently, Harris and Carpenter (2011) have reported that the same function applies to Pavlovian conditioning of magazine approach in rats, consistent with the idea that the sensitivity of instrumental responding outcome rate reflects the motivational influence of Pavlovian contextual conditioning.

This Pavlovian motivation modulates habitual rather than goal-directed behaviour. Holland (2004) reported that a larger PIT effect when behavioral autonomy had been induced by extended training, whereas Wiltgen et al. (2012) reported a similar association between the habitual status of responding and general PIT in mice by contrasting ratio and interval training. They observed greater PIT following interval training when performance was impervious to outcome devaluation than following ratio training when responding was sensitive to the current outcome value. Further evidence that the target of Pavlovian motivation is habitual comes from the fact that the magnitude of PIT was unaffected by whether the outcome associated with the Pavlovian stimulus was the same as or different from the instrumental outcome [9].

The most compelling demonstration of the generality of Pavlovian motivation comes from an irrelevant incentive study of PIT. Dickinson and Dawson (1987) trained hungry rats to lever-press for food pellet while also pairing one stimulus with the pellets and another with sugar water in the absence of the lever. When for the first time the rats were given the opportunity to press the lever during the stimuli while thirsty and in the absence of any outcomes, they did so more during the sugar-water stimulus than during the pellet signal. This finding establishes two important points. The first is the generality of the motivational influence which augments any prepotent habitual response even if that response was trained with a reinforcer that differs from that associated with the stimulus. Second, the Pavlovian motivational process can endow habitual responding with a veneer of goal-directedness. The shift of motivational state from training under hunger to PIT testing under thirst is an apparent outcome revaluation procedure in that the sugar-water reinforcer remained relevant to the test motivational state whereas the pellet reinforcer did not. However, this apparent outcome revaluation effect did not indicate goal-directed control because the revaluation did not operate through a representation of the action-outcome contingency in that lever pressing was trained with the food pellets (Corbit *et al.*, 2007), not the sugar water. In conclusion, the sensitivity of this Pavlovian motivation to an outcome revaluation procedure

[9] This motivational effect of Pavlovian stimuli on instrumental responding is called *general* PIT, as it increases the probability of responding for all the available responses and is thought to be mediated by a general energizing effect of a stimulus that is associated with the motivational properties of the outcome. This is in contrast with *specific* PIT, where responding is enhanced only to the response that predicts the same outcome as in training and is thought to be mediated by the association between the stimulus and the sensory properties of the outcome (see Cartoni *et al.*, 2016 for a review).

can easily lead to the erroneous attribution of goal-directed status. For example, Jonkman et al. (2010) reported that rate lever pressing remained sensitive to outcome revaluation even after extensive training on an interval schedule. It is very likely, however, that the apparent devaluation effect was mediated by Pavlovian contextual motivation of habitual responding. Extinguishing context conditioning prior to devaluation test significantly reduced the magnitude of the effect (see also Killcross and Coutureau, 2003).

Recall that the performance function, Equation 6, which transforms response strengths into response probability, includes a term I that represented the current incentive value of the outcome and in product with g determines the contribution of the goal-directed system to performance. By analogy, we also include an additional parameter that reflects the motivational effects of appetitive Pavlovian stimuli on habitual performance. Following Hull's (1943) classic nomenclature, we denote this parameter as D for drive, which multiplies the habit strength h to represent the contribution of the habit system to overall performance. Like the Hullian drive concept, D appears to exert a general motivational effect, at least within the appetitive domain, so that the complete response function has the form $p_{k+1} = s(Ig_k + Dh_k)$, where s is the sigmoid function as shown in Equation 6.

Incentive learning. In contrast to the Pavlovian motivational control of habits, animals have to learn about the incentives value I of ouctomes, such as foods and fluids, through consummatory experience with these commodities if they are to function as goals of an instrumental action, a process that Dickinson and Balleine (1994; 2002) refer to as incentive learning. Moreover, they also also have to learn how these incentive value vary with motivational state. Dickinson and Dawson (1988; 1989) first reported the role of incentive learning in the motivational control of goal-directed action using an irrelevant incentive procedure similar to the one they had employed to investigate the Pavlovian motivation of habits, namely a shift from training under hunger to testing under thirst. Their rats were initially trained to lever-press and chain-pull, one for food pellets and the other for sugar water, while hungry. Note that this training ensured that the contextual stimuli were equally associated with both outcomes whatever the action-outcome assignment, thereby equating any contextual motivation. During a subsequent extinction test, thirsty rats only preferentially performed the action trained with the sugar water if they had previously had the opportunity to drink the sugar water while thirsty, indicating that they had to learn about the incentive value of the sugar water when thirsty. Such incentive learning is required not only for shifts between motivational states but also variations with a motivational state, such as that between satiety and hunger (Balleine, 1992). Dickinson and Balleine (2019; 2009) have subsequenty argued that the assignment of incentive value to an outcome is based on the experienced hedonic reactions to, and evaluation of that outcome.

In summary, the motivation of habits and goal-directed actions is varied and complex, even in the case of basic biological commodities, such food and fluids. Habits are motivated by a general appetitive drive conditioned to contextual and eliciting stimuli, whereas the incentive value of the outcome, which is learned, motivates goal-directed action. Habitual motivation is directly sensitive to shifts in motivation state, whereas the agent has to learn about incentive values of outcomes in different motivational states before they can control goal-directed action.

Avoidance

So far we have developed rate correlation theory within a dual-system framework by reference to positive reinforcement of free-operant behavior using appetitive or attractive outcomes. However, Baum (1973) also analyzed free-operant avoidance in terms of his correlational law of effect. Under a typical free-operant avoidance contingency, a response causes the omission or postponent of a future scheduled outcome with the consequence that our recycling memory model yields a negative goal-directed strength ($g < 0$), at response rates that do not avoid all the schedule outcomes in a memory cycle. On the assumption that experience with the aversive outcome through incentive learning produce a negative incentive value, $(I < 0)$, the product of the negative goal-directed strength and incentive value, Ig, will be positive and thereby contribute to the probability of a response being performed, p. Moreover, once the response rate is sufficient to avoid all schedule outcomes within a memory cycle, the goal-direct strength will remain frozen at the established g value and thereby produce sustained avoidance in the absence of the aversive outcomes. This simple mechanism would explain the persistence of avoidance actions in the absence of an explicit reinforcing event, which has been the subject of multiple discussions in the literature (for a recent review, see Gillan *et al.*, 2016).

The most radical aspect of this account is its assumption of goal-directed control of avoidance responding. Although there are precedents for a goal-directed account of avoidance (e.g. Seligman and Johnson, 1973), contemporary RL theory follows traditional two-process theory in assuming that avoidance responding is purely habitual or MF (see Maia, 2009). Although human discrete-trial procedures have demonstrated a reduction in avoidance following revaluation of the aversive outcome (Gillan *et al.*, 2011), more critical for a rate correlation account of goal-directed avoidance is a demonstration by Fernando et al. (Fernando *et al.*, 2014a) of an outcome revaluation effect using a a free-operant schedule. They trained rats to lever-press to avoid foot-shocks that were programmed to be delivered at fixed intervals. Their revaluation procedure consisted of non-contingent presentations of the shock under morphine, so that pain would be reduced and the aversive status of the shock devalued. During an extinction test, their rats decreased responding compared to a non-revalued control group, demonstrating that the their rats were performing the avoidance action to reduce the rate of an unpleasant outcome.

In accord with our dual-system model, Fernando and colleagues (2014) also investigated the role habit learning in free-operant avoidance. An enduring problem for reinforcement theory is the absence of any event following an avoidance response that could act as a reinforcer. However, Konorski and Miller discovered that avoidance training established performance of the response itself, or more strictly speaking the feedback stimuli generated by responding, as a conditioned aversive inhibitor and, subsequently, Weisman and Litner (1969) reported that an explicit aversive inhibitor can function as a conditioned reinforcer of free-operant avoidance responding by rats. Taken together, these results suggest that habitual responding may be reinforced by the feedback stimuli generated by responding itself. In accord with this analysis, Fernando et al. (2014) found that avoidance responding by their rats was enhanced by the presence of an explicit feedback stimulus and, moreover, this enhancement appeared to be habitual. Although exposure to the feedback stimulus under morphine enhanced its reinforcing property, the enhancement was not evident in an outcome revaluation test. This finding led Fernando and colleagues to conclude that the responding generated by the presence of the explicit feedback stimulus was habitual.

In summary, free-operant avoidance, like its appetitive counterpart, is under joint control by goal-directed and habitual systems with the former reflecting rate correlation learning between the response and aversive outcome and the latter reinforcement by the aversive inhibitory property of response-generated feedback stimuli.

Conclusions

In this paper we have formalized a theory of instrumental actions and habits in free-operant conditions based on two different systems that concurrently control behavior. After discussing the multiple difficulties of theories based solely on outcome probability and reward prediction-error to explain instrumental control and performance, we presented an alternative theory of goal-directed control where agents compute a correlation between rates of responding and rate of outcomes in a fixed working memory to establish the casual association between their actions the outcomes and jointly determine the amount of responding and sensitivity to outcome revaluation under different reward schedules. We showed how such a theory can capture instrumental performance under ratio and interval schedules when reward probabilities or rates are matched, how goal-directed control transitions to habits with extended training and a faster development of habits under interval than under ratio schedules. The model also explains why responding under choice procedures tends to remain goal-directed control in spite of the amount of training when different outcomes are employed. These results make our model unique in its joint predictions with respect to instrumental control and performance in free-operant training.

Another aspect which is unique to the present model is that it provides a mechanism to explain the survival of goal-directed control across extinction. In our model, the reward prediction-error for the habit system includes the total prediction of both behavioral systems. This, together with the additional assumption that the goal-directed system can only compute a rate correlation when there are events in memory which can be processed, make it so that the habit system effectively inhibits the goal-directed system when the outcome is suspended in an extinction phase. The implication is that responding extinguishes because the sum of the strengths of the systems approaches zero, even though the goal-directed system remains active with the value of the last rate correlation experienced during instrumental training.

In summary, the main contribution of our theory is extending the widely-held view that outcome probability and reward-prediction error are the cardinal determinants of instrumental behavior, to one in which agents' computations are made simultaneously in correlational and contiguity systems to determine the decision to perform an instrumental action. Although there is some evidence suggesting that humans can compute a response-outcome rate correlation to inform their causal beliefs of a response-outcome association (Tanaka *et al.*, 2008), the exact neural processes underlying this computation, and the way in which these computations are transferred to performance remain unknown (Perez and Soto, 2019). This is an under-studied area for which the predictions of the present model might help our understanding of goal-directed and habitual processes, as clear evidence for arbitration between the systems in humans is still sparse.

Author contributions

OP and AD formalized the model. OP performed the simulations. OP and AD wrote the manuscript.

Acknowledgments

This work was part of OP's Ph.D. thesis work in the Department of Psychology and the Behavioural and Clinical Neuroscience Institute at the University of Cambridge, funded by a scholarship from CONICYT. OP thanks Amy Milton and Mike Aitken for supervision, and Gonzalo Urcelay and Fabian Soto for helpful discussions on this topic. We also thank Yutaka Kosaki for providing us with data to inform our simulations. OP would also like to thank the NHS for all the support provided during his Ph.D. studies.

References

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

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*: 37–41.

Adams, C. D. and Dickinson, A (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B : Comparative and Physiological Psychology*, **33**: 109–121.

Balleine, B. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **18**: 236.

Balleine, B. W. (2019). The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. *Neuron*, **104**: 47–62.

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

Balleine, B. and O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, **35**: 48–69.

Baum, W. (1973). The Correlation-Based Law of Effect. *Journal of the Experimental Analysis of Behavior*: 137–153.

Baum, W. (1992). IN SEARCH OF THE FEEDBACK FUNCTION FOR VARIABLE INTERVAL SCHEDULES. *Journal of the Experimental Analysis of Behavior*, **3**: 365–375.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, **10**: 1214–1221.

Bradshaw, C., Ruddle, H., and Szabadi, E (1981). Relationship between response rate and reinforcement frequency in variable-interval schedules: III. The effect of d-amphetamine. *Journal of the Experimental Analysis of Behavior*, **36**: 29–39.

Bush, R. R. and Mosteller, F (1951). A mathematical model for simple learning. *Psychological review*, **58**: 313–323.

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

Catania, A. C., Matrrhews, T. J., Silverman, P. J., Yohalem, R., Matthews, T. J., Silverman, P. J., and Yohalem, R. (1977). Yoked Variable-Ratio and Variable-Interval responding in pigeons. *Journal of the Experimental Analysis of Behavior*, **28**: 155–161.

Colwill, R. M. (1991). Negative discriminative stimuli provide information about the identity of omitted response-contingent outcomes. *Animal Learning & Behavior*, **19**: 326–336.

Colwill, R. M. and Rescorla, R. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**: 520–536.

Colwill, R. M. and Rescorla, R. (1988). Associations between the discriminative stimulus and the reinforcer in instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **14**: 155–164.

Colwill, R. M. (2001). The effect of noncontingent outcomes on extinction of the response-outcome association. *Animal Learning & Behavior*, **29**: 153–164.

Corbit, L. H., Janak, P. H., and Balleine, B. (2007). General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the ventral tegmental area. *The European journal of neuroscience*, **26**: 3141–3149.

Corbit, L. H., Chieng, B. C., and Balleine, B. (2014). Effects of Repeated Cocaine Exposure on Habit Learning and Reversal by N-Acetylcysteine. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, **39**: 1–9.

Daw, N., Niv, Y, and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, **8**: 1704–1711.

Dawson, G. R. and Dickinson, A (1990). Performance on ratio and interval schedules with matched reinforcement rates. *The Quarterly Journal of Experimental Psychology*, **42**: 37–41.

Derusso, A. L., Fan, D., Gupta, J., Shelest, O., Costa, R. M., and Yin, H. H. (2010). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in integrative neuroscience*, **4**: 1–8.

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

Dickinson, A (1994). Instrumental conditioning. *Animal Cognition and Learning.* Ed. by N. J. Mackintosh. London: Academic Press. Chap. 3: 45–78.

Dickinson, A and Balleine, B. (1993). Actions and responses: The dual psychology of behaviour. *Spatial representation.*

Dickinson, A and Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning {&} Behavior.* Ed. by N. J. Mackintosh. Vol. 22. 1. London: Academic Press. Chap. 3: 1–18.

Dickinson, A and Balleine, B. (2002). The Role of Learning in the Operation of Motivational Systems. **3**: 497–534.

Dickinson, A and Balleine, B. (2009). Hedonics : The Cognitive – Motivational Interface. *Pleasures of the Brain.* 74–84.

Dickinson, A and Dawson, G. R. (1987). Pavlovian processes in the motivational control of instrumental performance. *The Quarterly Journal of Experimental Psychology*, **39**: 201–213.

Dickinson, A and Perez, O. D. (2018). Actions and habits: Psychological issues in dual-system theory. *Goal-Directed Decision Making: Computations and Neural Circuits.* Ed. by R. W. Morris, A. M. Bornstein, and A. Shenhav. Elsevier: 1–37.

Dickinson, A, Watt, A, and Griffiths, W. (1992). Free-operant acquisition with delayed reinforcement. *The Quarterly Journal of Experimental Psychology Section B*, **45**: 241–258.

Dickinson, A, Balleine, B., Watt, A., Gonzalez, F., and Boakes, R. a. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, **23**: 197–206.

Dickinson, A, Squire, S, Varga, Z, and Smith, J. W. (1998). Om ission Learn in g after In stru m en tal Pretrain in g: 271–286.

Dickinson, A. and Dawson, G. (1988). Motivational control of instrumental performance: The role of prior experience of the reinforcer. *The Quarterly Journal of Experimental Psychology Section B*, **40**: 113–134.

Dickinson, A. and Dawson, G. (1989). Incentive learning and the motivational control of instrumental performance. *The Quarterly Journal of Experimental Psychology*, **41**: 99–112.

Dickinson, A., Nicholas, D., and Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, **35**: 35–51.

Dolan, R. and Dayan, P. (2013). Goals and habits in the brain. *Neuron*, **80**: 312–325.

Estes, W. K. (1948). Discriminative conditioning. II. Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *Journal of experimental psychology*, **38**: 173.

Fernando, A., Urcelay, G. P., Mar, A., Dickinson, A, and Robbins, T. W. (2014a). Free-operant avoidance behavior by rats after reinforcer revaluation using opioid agonists and D-amphetamine. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **34**: 6286–6293.

Fernando, A. B., Urcelay, G. P., Mar, A. C., Dickinson, A., and Robbins, T. W. (2014b). Safety signals as instrumental reinforcers during free-operant avoidance. *Learning & Memory*, **21**: 488–497.

Gershman, S. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*: 1–8.

Gillan, C. M., Papmeyer, M., Morein-zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., and Wit, S de (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, **168**: 718–726.

Gillan, C. M., Urcelay, G. P., and Robbins, T. W. (2016). An associative account of avoidance. *The Wiley Handbook on the Cognitive Neuroscience of Learning*. Wiley Online Library: 442.

Gremel, C. M. and Costa, R. M. (2013). Orbifrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature Communications*, **4**: 1–12.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the experimental analysis of behavior*, **34**: 297–304.

Harris, J. A. and Carpenter, J. S. (2011). Response rate and reinforcement rate in Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, **37**: 375.

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

Herrnstein, R. J.J.J. J. (1970). On the Law of Effect. *Journal of the experimental analysis of behavior*, **2**: 243–266.

Heyes, C. M. M. and Dawson, G. R. (1990). A demonstration of observational learning in rats using a bidirectional control. *The Quarterly Journal of Experimental Psychology.*

Heyman, G. M. (1979). Matching and Maximizing in Concurrent. *Psychological review*, **86**: 496–500.

Hilario, M., Holloway, T., Jin, X., and Costa, R. M. (2012). Different dorsal striatum circuits mediate action discrimination and action generalization. *European Journal of Neuroscience*, **35**: 1105–1114.

Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, **30**: 104.

Holman, E. W. (1975). Some conditions for the dissociation of consummatory and instrumental behavior in rats. *Learning and Motivation*, **6**: 358–366.

Hull, C. (1943). No Title. *Principles of behavior*. Appleton-century-crofts.

Jonkman, S., Kosaki, Y., Everitt, B. J., and Dickinson, A (2010). The role of contextual conditioning in the effect of reinforcer devaluation on instrumental performance by rats. *Behavioural processes*, **83**: 276–281.

Keramati, M., Dezfouli, A, and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, **7**.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex (New York, N.Y. : 1991)*, **13**: 400–408.

Killeen, P. R. (1982). Incentive theory: II. Models for choice. *Journal of the Experimental Analysis of Behavior*: 217–232.

Killeen, P. R. (1994). Mathematical principles of reinforcement. *Behavioral and Brain Sciences*, **17**: 105.

Killeen, P. R., Hanson, S. J., and Osborne, S. R. (1978). Arousal: Its genesis and manifestation as response rate. *Psychological Review*, **85**: 571–581.

Kosaki, Y. and Dickinson, A (2010). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of experimental psychology. Animal behavior processes*, **36**: 334–342.

Kuch, D. and Platt, J. R. (1976). Reinforcement rate and interresponse time differentiation. *Journal of the experimental analysis of behavior*, **3**: 471–486.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, **1**: 0067.

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

Mackintosh, N. J. and Dickinson, A (1979). Instrumental (Type II) Conditioning. *Mechanisms of learning and motivation*: 143–167.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, affective & behavioral neuroscience*, **9**: 343–364.

Mazur, J. E. (1983). STEADY STATE PERFORMANCE ON FIXED, MIXED, AND RANDOM RATIO SCHEDULES. *Journal of the Experimental Analysis of Behavior*, **2**: 293–307.

Miller, K. J., Shenhav, A., and Ludvig, E. A. (2019). Habits without values. *Psychological review*.

Morrison, G. R. and Collyer, R. (1974). Taste-mediated conditioned aversion to an exteroceptive stimulus following LiCl poisoning. *Journal of Comparative and Physiological Psychology*, **86**: 51.

Nelson, A. and Killcross, S. (2006). Amphetamine exposure enhances habit formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **26**: 3805–3812.

Niv, Y, Daw, N., and Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, **18**: 1019–1026.

Niv, Y, Dayan, P., and Joel, D. (2006). The Effects of Motivation on Extensively Trained Behavior. *Leibniz Technical Report, Hebrew University*: 1–25.

Niv, Y, Daw, N., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, **191**: 507–520.

Palminteri, S., Lefebvre, G., Kilford, E. J., and Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, **13**: e1005684.

Perez, O. D. and Soto, F. (2019). Evidence for a dissociation between causal beliefs and instrumental actions. *The Quarterly Journal of Experimental Psychology*.

Pérez, O. D., Aitken, M. R., Milton, A. L., and Dickinson, A. (2018). A re-examination of responding on ratio and regulated-probability interval schedules. *Learning and Motivation*, **64**: 1–8.

Prelec, D (1982). Matching, maximizing, and the hyperbolic reinforcement feedback function. *Psychological Review*, **89**.

Reed, P (2001). Schedules of reinforcement as determinants of human causality judgments and response rates. *Journal of Experimental Psychology: Animal Behavior Processes*, **27**: 187–195.

Rescorla, R. (1993). Inhibitory associations between S and R in extinction. *Animal Learning {&} Behavior*, **21**: 327–336.

Rescorla, R. (1994). Transfer of instrumental control mediated by a devalued outcome. *Animal Learning {&} Behavior*, **22**: 27–33.

Rescorla, R. and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, **2**: 64–99.

Rescorla, R. A. and Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological review*, **74**: 151.

Rhodes, S. E. and Murray, E. A. (2013). Differential effects of amygdala, orbital prefrontal cortex, and prelimbic cortex lesions on goal-directed behavior in rhesus macaques. *Journal of Neuroscience*, **33**: 3380–3389.

Seligman, M and Johnson, J (1973). A cognitive theory of avoidance learning. Washington.

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological science*, **2**: 341–350.

Soto, P. L., McDowell, J. J., and Dallery, J. (2006). Feedback Functions, Optimization, and the Relation of Response Rate to Reinforcer Rate. *Journal of the Experimental Analysis of Behavior*, **85**: 57–71.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.

Tanaka, S. C., Balleine, B., and O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **28**: 6750–6755.

Tanno, T and Sakagami, T (2008). On The Primacy of Molecular Processes in Determining Response Rates Under Variable-Ratio and Variable-interval Schedules. *Journal of the Experimental Analysis of Behavior*, **89**: 5–14.

Thorndike, E. L. (1911). Edward Lee Thorndike. *Animal Intelligence*, **1874**: 1949.

Thrailkill, E. A. and Bouton, M. E. (2015). Contextual control of instrumental actions and habits. *Journal of Experimental Psychology: Animal Learning and Cognition*, **41**: 69.

Tricomi, E., Balleine, B., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The European journal of neuroscience*, **29**: 2225–2232.

Urcelay, G. P. and Jonkman, S. (2019). Delayed rewards facilitate habit formation. *Journal of Experimental Psychology: Animal Learning and Cognition*.

Valentin, V. V., Dickinson, A, and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **27**: 4019–4026.

Villiers, P. a. de and Herrnstein, R. J. (1976). Toward a law of response strength. *Psychological Bulletin*, **83**: 1131–1153.

Vogel, E. H., Castro, M. E., and Saavedra, M. A. (2004). Quantitative models of Pavlovian conditioning. *Brain research bulletin*, **63**: 173–202.

Weisman, R. and Litner, J. (1969). Positive conditioned reinforcement of Sidman avoidance behavior in rats. *Journal of Comparative and Physiological Psychology*, **68**: 597.

Wiltgen, B. J., Sinclair, C., Lane, C., Barrows, F., Molina, M. M., and Chabanon-Hicks, C. (2012). The Effect of Ratio and Interval Training on Pavlovian-Instrumental Transfer in Mice. *PLoS ONE*, **7**: 1–5.

Wit, S. de, Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., and Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, **147**: 1043.