Dual-system avoidance: extension of a theory

Omar D. Perez

Department of Industrial Engineering, University of Chile, Santiago, Chile

Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

Instituto Sistemas Complejos de Ingeniería, Chile

Anthony Dickinson

Department of Psychology and Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge, UK

Address correspondence to:

Omar D. Perez

Department of Industrial Engineering

University of Chile

Santiago, Chile

omar.perez.r@uchile.cl

Abstract

Rewarded behavior is controlled by 2 systems during free-operant training. One system is sensitive to the correlation between response and reinforcement rate and controls goal-directed behavior, whereas a habitual system learns by reward prediction error. We present an extension of this theory to the aversive domain which explains why free-operant avoidance responding increases with both the experienced rate of the negative reinforcer and the decrease in this rate produced by responding. The theory also assumes that the habitual component is reinforced by the acquisition of aversive inhibitory properties by the feedback stimuli generated by responding, which then act as safety signals that reinforce habit performance. Our results show that the same distinction of habitual and goal-directed control of rewarded behavior can be applied to the aversive domain.

Keywords: avoidance, habits, free-operant, goal-directed, model-free, model-based

Sidman (1953) introduced free-operant avoidance by scheduling periodic shocks that rats could postpone or omit by pressing a lever. This form of avoidance has always been problematic for a contiguity-based reinforcement account because a rat will respond for many minutes in the absence of any obvious source of reinforcement. However, a contiguous source of reinforcement was identified by Konorski and Miller in their 1936 report of one of the initial experimental studies of avoidance (Konorski, 1948 p. 228-232; 1967 p. 380-383). They initially established a Pavlovian defensive or aversive salivary conditioned response to a noise by pairing this signal with the delivery of dilute distasteful hydrochloric acid into the dog's mouth. Once the salivary conditioned response was established, they occasionally passively flexed one of the dogs forelegs for 5 s during the noise and omitted the acid outcome. In the protocol illustrated by Konorski (1967 p. 381), after 7 noise presentations with the passive flexion, the dog for the first time spontaneously (and presumably voluntarily) flexed its leg during the noise, resulting in the omission of the acid delivery. Thereafter the dog flexed its leg many times during almost every noise presentation, thereby avoiding most of the impending acid deliveries.

Critically, Konorski and Miller also observed that the spontaneous leg flexions were accompanied by a marked reduction in the conditioned salivary response to the noise, which led them to suggest that feedback stimuli generated by the leg flexion had become conditioned aversive inhibitors through Pavlovian conditioning because these stimuli predicted the omission of an expected aversive outcome, the acid. Moreover, they argued that this property of the

feedback stimuli instrumentally reinforced avoidance responding. We shall refer to this account of avoidance as the safety signal theory, which was the first instantiation of what has come to be called a two-process theory (Rescorla & Solomon, 1967) and has received more contemporary endorsement (Dinsmoor, 2001).

Although Konorski and Miller's observations are compatible with the safety signal account, they did not experimentally manipulate the properties of the feedback stimuli to evaluate the theory. The first to do so was Rescorla as reported in two papers published shortly after the completion of his doctorate. The safety signal account makes two claims. The first is that the feedback stimulus becomes an aversive inhibitor on a free-operant schedule. To evaluate this claim, Rescorla trained dogs on a free-operant avoidance schedule under which each shuttle response produced a brief auditory feedback stimulus before presenting this stimulus independently of responding to assess its conditioned properties. Relative to a control condition in which the stimulus had been presented randomly while the dogs were responding, the feedback stimulus suppressed avoidance, thereby demonstrating its inhibitory properties (Rescorla, 1968). Rescorla and LoLordo had previously demonstrated that an independently established Pavlovan aversive inhibitor would suppress the free-operant shuttling avoidance response (Rescorla & Lolordo, 1965).

The second claim is that an aversive inhibitor acts as a positive reinforcer for the avoidance response. To address this issue Rescorla initially trained his dogs in a concurrent schedule of a free-operant avoidance in which pressing either of two panels postponed the next shock (Rescorla, 1969). In the second

stage the panels were removed and the dogs were given a Pavlovian inhibitory conditioning. The avoidance schedule was reinstated with the condition inhibitor from the prior stage being presented following each press of one of the panels. During this test stage the dogs showed a clear preference for the response producing the aversive inhibitor, thereby establishing the aversive inhibitor as a positive reinforcer. Weisman and Litner reached the same conclusion using bidirectional instrumental control assessment of free-operant avoidance in rats (Weisman & Litner, 1969).

The safety signal theory assumes that free-operant avoidance, although operationally an example of negative reinforcement with the shock as the reinforcer, is in fact functionally an example of positive reinforcement by the feedback stimuli generated by the avoidance response. This form of positive reinforcement was explained by Dickinson and Dearing in terms of an opponent process between appetitive and aversive motivational systems under which a conditioned aversive inhibitor activates the appetitive motivational system and thereby functions like a conditioned appetitive excitor (Dickinson & Dearing, 1979). A variety of evidence confirms this functional equivalence – for example, DeVito and Fowler reported that training a flashing light as aversive inhibitor facilitates subsequent appetitive conditioning to this stimulus (DeVito & Fowler, 1994), whereas appetitive conditioning is blocked when conducted in the presence of an aversive inhibitor (Dickinson & Dearing, 1979; Laurent et al., 2018).

Habitual Avoidance

It is now generally accepted that positively reinforced instrumental behavior comes in two forms (Daw & O'Doherty, 2013; Dickinson, 1985; Dickinson & Pérez, 2018; Dolan & Dayan, 2013): as a habitual response and as a goal-directed action. Experience with an instrumental contingency is assumed to strengthen habitual responding without encoding information about the reinforcer or outcome of the response. A classic example of such a mechanism is Thorndike's law of effect according to which an association between a current stimulus and a response is strengthened when the response is followed by an effective reward (Thorndike, 1911). By contrast, a goal-directed behavior is mediated by a rational interaction between knowledge of the causal relation between the action or response and the reinforcer and the current value of the reinforcer (Heyes & Dickinson, 1990). Such an interaction is goal-directed in the sense it is *directed* by knowledge of the action-reinforcer contingency and motivated by the representation of an outcome as a *goal*. Consequently, goal-directed learning involves encoding a representation of the outcome, in this case feedback stimulus, as a goal of the instrumental action

The canonical assay for distinguishing between habitual and goal-directed control is the reinforcer or outcome revaluation test (Adams & Dickinson, 1981). The rationale for this test can be illustrated by a revaluation test conducted by Fernando and colleagues to determine whether free-operant avoidance by rats is goal-directed or habitual with respect to a feedback stimulus that functioned as a safety signal (Fernando et al., 2014b). Their rats were trained on a free-operant variable cycle (VC) schedule, which consisted of a variable avoidance period followed by a shock period in which three foot-shocks were presented with a

short interval between them before the next component was presented. A lever press during either period terminated the current cycle so that any further programmed shocks in that cycle were omitted. Therefore, by pressing in each avoidance period the rat could avoid all of the shocks that were scheduled to occur after the variable avoidance period. Furthermore, each lever press produced a 5-s auditory feedback stimulus, which was assumed to function like the endogenous feedback stimuli produced by pressing the lever and thereby enhance the salience of the sensory feedback produced by the instrumental response. In accord with the safety signal theory, in separate experiments Fernando and colleagues not only replicated Rescorla's finding of a preference for an avoidance response that produced the feedback stimulus but also that the feedback stimulus contingency enhanced the rate of the avoidance response (Dinsmoor & Sears, 1973), as well inhibiting avoidance in its presence.

Following this training, the lever was withdrawn and the revaluation group received non-contingent exposures to the feedback stimulus under morphine. In a prior experiment, the same authors had demonstrated that this treatment enhanced the reinforcing effect of the feedback stimulus on avoidance responding when tested in the absence of the shock. In the critical experiment, however, they tested the effect of the revaluation treatment in the absence of both the feedback stimulus and the shock. If the revaluation treatment acts by enhancing the capacity of the feedback stimulus to reinforce habitual responding, such an enhancement should not be observed in the absence of this stimulus during the extinction test. By contrast, if the impact of the feedback stimulus is goal-directed in the sense of being mediated by a representation of the current

value of this stimulus, the revaluation should enhance responding even during the extinction test. Critically Fernando and colleagues failed to detect any effect of the morphine revaluation on extinction suggesting that the safety signal operates through habit-based reinforcement rather than enhancing the value of the signal as a goal. As it stands, this inference is based on a null result in the extinction test and therefore the habit-based interpretation also requires the demonstration of an interaction between the revaluation effect and the type of test - extinction versus a reinforced test with response-dependent feedback stimulus. A subsequent reinforced test replicated the effect of the revaluation and yielded a significant interaction. In conclusion, the Fernando et al. (2014) experiments suggest that the positive reinforcement generated by a safety signal does so by enhancing habit learning and not by establishing a representation of the causal relation between the action and its feedback stimulus that is necessary for goal-directed control.

Goal-directed Avoidance

As well as investigating the nature of the avoidance maintained by a safety signal, Fernando and colleagues also used the reinforcer revaluation procedure to determine whether the current value of the shock plays a direct role in controlling free-operant avoidance (Fernando et al., 2014a). Rather than revaluing the feedback stimulus as in their previous study, they sought to revalue the shock rather than the feedback stimulus. To this end, they trained the rats on their VC avoidance schedule and then, in the absence of the lever, exposed their rats to non-contingent presentations of the shock under morphine in an attempt

to reduce its aversiveness. If avoidance was motivated by the negative value of the shock interacting with knowledge of the negative causal relationship between responding and the shock, this treatment should have reduced responding. This reduction is exactly what they observed in an extinction test without the shock and during initial exposure to the avoidance schedule in a reinforced test, thereby demonstrating that lever pressing on their schedule was goal-directed with respect to avoiding the shock.

What is less clear, however, is the nature of the representations and processes underlying this goal-directed avoidance. Shortly after Rescorla reported evidence for a role of safety signals in avoidance, Seligman and Johnson (Seligman & Johnson, 1973) published a seminal chapter in which, having critically reviewed classic two-process or -factor accounts of avoidance (Rescorla & Solomon, 1967), they argued for a goal-directed account of avoidance in terms of Tolmanian expectations and preferences (Tolman, 1948). Specifically, they suggested responding is generated by the interaction of expectations of particular outcomes following different actions, the avoidance response and non-avoidance response, and the preferences among their outcomes, no shock and shock, respectively, an idea that has been endorsed by Lovibond (2006). Another contemporary framework for analyzing goal-directed behavior is that provided by computational reinforcement learning (RL) (Sutton & Barto, 2018). For example, Wang and colleagues (Wang et al., 2018) developed a model-based RL account of goal-directed avoidance, which assumes that the agent learns stimulus state-action-outcome state transitions using state prediction-errors which are then deployed in the selection of the action that yields the preferred

outcome state. Such state transitions learning map paradigmatically onto discrete-trial avoidance learning in which performing the avoidance response in the state generated by the presence of a warning signal for an aversive outcome leads to a state in which the outcome is omitted. Although expectancy and model-based RL theories may provided accounts of discrete-trial human goal-directed avoidance (e.g. Gillan et al, 2013), it is not clear that this theory can explain the impact of the two major interacting parameters that determine the rate of free-operant avoidance where a warning signal is not present and no explicit trial structure: the experienced shock rate and the reduction in shock rate produced by avoidance responding in the absence of any warning signal.

We shall illustrate the impact of these two variables by a data set reported by de Villiers for foot shock avoidance by lever pressing in rats. Like Fernando et al. (2014), de Villiers used a VC shock schedule under which the first lever press in an interval canceled the next scheduled shock[1]. Given the similarity with the schedule employed in Fernando et al's (2014) shock revaluation study, we have grounds for assuming that de Villiers' rats also performed goal-directed avoidance. When he varied the parameter of the VC schedule, the resulting avoidance rate increased systematically with both the experienced shock rate and the reduction in shock rate from the rate programmed by the schedule (see top panels in Figure 1C). In the next section we consider whether a dual-system theory that we have recently presented to explain positively reinforced

[1] de Villiers (1974) refers to his avoidance contingency as a variable interval (VI) schedule but we have chosen to follow Uhl and Eichbauer's (1975) terminology by referring to the schedule as a variable cycle (VC).

free-operant behavior (Perez & Dickinson, 2020) can also capture responding under free-operant avoidance.

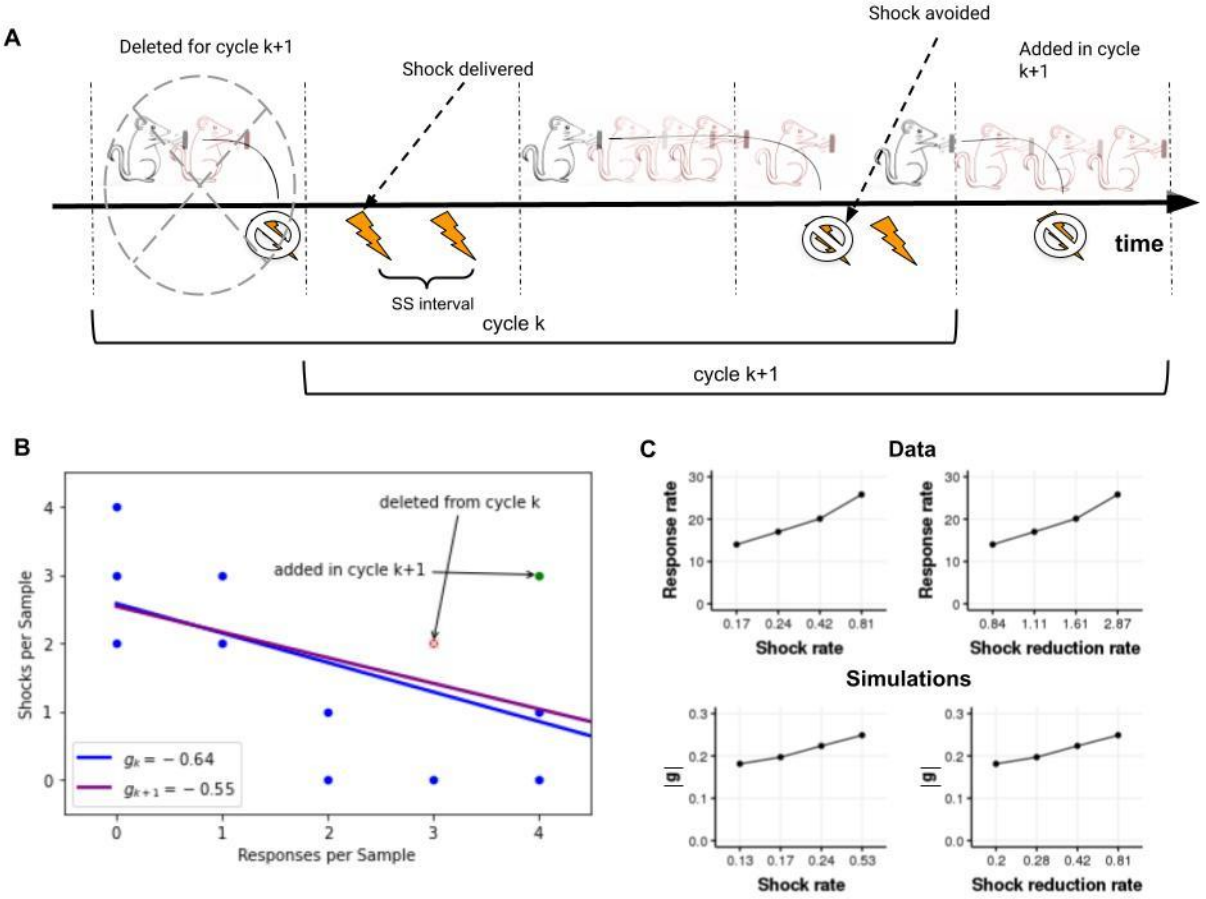

Figure 1. Avoidance free-operant training. A. In free-operant avoidance shocks (Ss) are predetermined to come on a variable cycle (VC) schedule where a mean shock-shock (SS) interval is programmed by the experimenter. Subjects can respond at any time during a cycle. When a response is performed (black rat in the figure), the next scheduled shock is canceled, but further responses within the cycle (before the shock is indeed canceled) have no consequences on future programmed (red rat in the figure). Each time-window represents a sample in memory. For each memory cycle k, the agent registers the responses per sample and received shocks per sample. After a memory recycle, one of the samples is randomly erased from memory (the first one, in this example) and a new memory sample is added in cycle $k + 1$. **B.** Events can be plotted as data points representing responses and shocks per memory sample. The blue line represents the best fitting line for cycle k and the red line represents the best fitting line for cycle $k + 1$; the slope of each line represents the computed rate correlation. One datapoint from memory cycle k (red dot) is erased from memory while another is added for the next cycle $k + 1$ (green dot). The negative value of g is multiplied by the negative incentive value of the aversive shock (I) to yield a positive response strength for the goal-directed system. **C.** Top panel. Results obtained by De Villiers

(1974) in an experimental design involving different programmed shock rates under VC training. Bottom Panel. Simulations of a goal-directed rate correlational system for the same schedules employed in De Villiers (1974).

Rate-correlation System

Within our theory (Perez & Dickinson, 2020), the represented strength of the action-reinforcer relationship acquired through instrumental training is referred to as *g* and the incentive value of the reinforcer as *I* with the propensity to perform the action being determined by the product of *g* and *I*. As both *g* and *I* are positive for a rewarded response so is their product, thereby engendering performance of this response. By contrast, in the case of an avoidance response both *g* and *I* are negative - the former because the action-reinforcer contingency is negative and latter because the reinforcer is aversive. So once again the product, and hence the propensity to respond, is positive. The reinforcer revaluation procedure employed by Fernando and colleagues should have decreased the negativity of the incentive value of the shock, thereby reducing the positive product of *g* and *I*.

Influenced by Baum's correlation-based law of effect (Baum, 1973), we suggested that *g* is determined by a mnemonic system that deploys a short-term memory (STM) to compute the current local correlation between the rate of responding and the rate of reinforcement. As illustrated in Figure 1a, the contents of the STM consist of a number of time samples each of which records the number of actions and the number of outcomes or reinforcers that occur in that sample. At the end of each time sample, the current correlation between the response and outcome rates, *r,* is calculated across the time samples currently in

the STM before one of the memory samples is randomly deleted from memory and the registration of responses and outcomes in a new sample is started[2]. Figure 1A illustrates the operation of the system with a STM of five time samples across two recycles of the memory, whereas Figure 1B displays the data used to calculate the current *r* at each memory recycle, which is then used to update the running average *r*.

The strength of goal-direct control, *g,* is a weighted mixture of the current *r* and the current mean *r*, which when multiplied by the current incentive value of the outcome, *I,* determines the probability of goal-directed responding in each second of the next time sample. Of course, the current *r* can only be calculated if the current STM has registered at least one action and one outcome. In other words, the agent will compute the rate correlation only if in any particular memory recycle there is at least one action and reinforcer registered in memory. In the absence of a registered action and/or reinforcer, *g* is determined by the mean *r* prior to the recycle, which is not then updated until current memory registers at least one of each event.

With the reinforcer incentive value set to one, this goal-directed system yields qualitative matches to the rates of lever pressing by rats under ratio and interval schedules of positive reinforcement across variations in the probability of the reward per press, the reward rate, and the delay between a reinforced

[2] This procedure for deleting memory samples differs from that used by Perez and Dickinson (2020) which deleted the oldest sample at a recycle. However, that procedure requires the system to represent the age of a sample, whereas random deletion ensures that the older a sample the more likely it is to have been deleted at any given recycle without requiring a representation of its age.

press and reward delivery using a consistent set of parameters (Perez & Dickinson, 2020). Of course, under a negative reinforcement, or an avoidance contingency, the rate correlation will be negative, and at issue is whether the same system can also account for the impact of variations in the important determinants of free-operant avoidance when the incentive value of the negative reinforcer is set to minus one to reflect its aversive properties. As a consequence the product of g and I yields a positive p reflecting the probability of responding in each second.

Simulations

To implement the avoidance simulations, our agent used an STM consisting of thirty 20-s samples, which are mnemonic parameters similar to those used for the simulations of positively reinforced behavior reported by Perez and Dickinson (2020). Simulations were performed using the R programming language under the RStudio IDE (RStudio Team, 2020).

The response-reinforcer relationship was computed and assessed by the agent at each memory recycle by a Pearson correlation coefficient between the number of responses and shocks per sample across the current samples in the STM (see Figure 1B). The strength of goal-directed control, g, throughout the next memory sample $k + 1$ was set at a weighted mean of the correlation yielded by that last recycle and the mean of the correlations at all previous recycles, that is:

$$g_{k+1} = \theta r_k + (1 - \theta)\bar{r}_k \quad (0 < \theta < 1) \qquad (1)$$

, where $\bar{r}_k$ is the mean correlation experienced during the experiment, up to

memory recycle k, which the agent can compute online as $\underline{r}_k + \beta(r_k - \underline{r}_k) \quad (0 < \beta < 1,\ k > 0)$. The weighting controls the extent to which goal-directed control depends upon the most recent or distant experiences of the instrumental contingency between the action and the reinforcer, in this case the absence of a shock.

We assume that the probability of a response being performed at each second is constant in each memory recycle k, so that $p_k = g_k \sim geom(p_k)$, where $geom()$ is a geometric distribution with parameter p_k. In de Villiers' (1974) procedure, each schedule contained a fixed number of intervals that approximated a constant probability of shock per second that yielded the appropriate scheduled interval. Consequently, our simulated schedules also maintained a constant shock probability per second generated in a similar way to responding by using a geometric distribution with parameter $s_t = geom(1/T)$, where T denotes the average interval between shocks, or the average shock-shock (SS) interval. For example, for a VC 15-s, shocks were generated by $s_t = geom(1/15)$ so that on average the SS interval was 15-s. The parameter s was kept fixed across training for all our simulated agents. Following de Villiers' paradigm, each response performed by the agent canceled the next programmed shock; additional responses during the SS interval did not have any effect on subsequent programmed shocks.

For the present simulations we used the same parameters across all our artificial agents with exception of the parameter T (the SS interval), which was

varied to study the sensitivity of our model to the schedules. The weighting θ given to the current correlation relative to the mean correlation in determining g_k was set to 0. 8. Each simulation was run for 200 mnemonic cycles and the mean response rate over the last 50 cycles was used to assess sensitivity to T . We performed 100 simulations of our virtual rats under this free-operant avoidance paradigm using VCs of 15-s, 30-s, 45-s and 60-s schedules - the same SS parameters used by de Villiers in his study. To this end, we programmed shocks to come at random times and canceled the following shock every time one response was performed before the shock; further responses before the canceled shock did not have any impact on the delivery of future shocks (see Figure 1A). The average sensitivity to variations in these schedule parameters was assessed by the mean response and shock rates and shock rate reductions generated for each schedule.

Figure 1C illustrates the mean response rates generated by de Villiers' rats and the strength of goal-directed responding $|g|$ for the simulations of the rate correlation system as a function of the mean received shock rate (left top and bottom panels, respectively) and as a function of the mean reduction in shock rate (right top and bottom panels, respectively)[3]. In both cases, the simulations show a qualitative match to the avoidance performance of the rats trained by de Villiers (1974), with a systematic reduction in responding as the shock rate and shock reduction rates decrease.

[3] We note that the response strength is a direct function of the product of the experienced rate correlation r and the incentive value of the shock, I. This interaction yields a positive response strength. We show the absolute value of g in the plot, which is equivalent.

Discussion

In response to the discovery by Fernando and colleagues that free-operant avoidance can be conjointly controlled by goal-directed (Fernando et al., 2014a) and habitual learning (Fernando et al., 2014b), we investigated whether our dual-system model of positively reinforced free-operant behavior (Perez & Dickinson, 2020) can be extended to the corresponding negative reinforcement. Within this model goal-directed control is determined by the experienced correlation between the response and reinforcement rates as assessed by a STM system, which in the case avoidance yields a negative correlation. Consequently, the causal relationship between the instrumental response and the reinforcer is represented as a negative value, which interacts with the negative incentive value of the aversive reinforcer to generate a positive response strength. By simulation we demonstrated that this rate-correlation system yields response rates that are positively related to two empirical variables of free-operant avoidance, the experienced reinforcer rate and the reduction in this rate produced by responding.

Recently, Baum (2020) presented an account of the relationship between free-operant avoidance and the experienced reinforcement or shock rate. Central to his account is the process of induction whereby experiences of a covariation, either positive or negative, between an activity, such as lever pressing, and a reinforcer enables the presentations of the reinforcer to induce the activity with the rate of induction being determined by the reinforcement rate. It is this process of induction that explains why the rate of lever pressing increased with

the shock rate in the extensive data set considered by Baum (2020), including the de Villiers' (1974) experiment that is the focus of our analyses. Importantly, however, the induction process is constrained by the avoidance schedule with the equilibrium response rate occurring at the point where the induction and schedule feedback functions intersect. If the response rate falls below this point the resulting increase in shock rate induces more responding, whereas if the response rate increases, the induction of responding is reduced by the decrease in shock rate. What remains unclear is the processes by which experience of a response-reinforcer covariation produces induction. If it is exposure to the rate correlation between responding and reinforcement originally identified by Baum fifty years ago as an important variable in free-operant responding, our rate correlation system provides no more than a mechanism for induction.

Whatever the relationship between Baum's induction and our goal-directed system, our dual-system model also includes a role for a second, habit system that instantiates a version of the classic two-process account of learning in the sense that this system involves the interaction of Pavlovian and instrumental habit learning. The relevant version of the two-process account of avoidance learning is the safety signal theory of Konorski and Miller who proposed that the avoidance response is positively reinforced by its feedback stimuli through their acquisition conditioned aversive inhibition. As in case of rewarded responding, we assume that the two systems conjointly determine the resultant response rate.

Currently, there is insufficient data to suggest how such a habit system should

operate computationally to produce the habit strength h that we proposed for rewarded behavior in our dual-system theory (Perez & Dickinson, 2020). A habit system for free-operant avoidance would require a fully-fledged Pavlovian theory of the role of the safety signals in the acquisition of the habit, including the strength of the Pavlovian aversive inhibition conditioned to these stimuli. The prediction-error underlying rewarded behavior should in this case reflect the difference between the strength of Pavlovian inhibition elicited by the feedback stimuli at the time when a response is performed, I_t^k , and the current habit strength, h_t^k , at time-step t in the current memory cycle k. Therefore the prediction error at that time (PE_t^k) should be given by $PE_t^k = I_t^k - h_t^k$. This prediction error would potentially serve as a teaching signal for the habit strength, and updated at each second t and cycle k by $h_{t+1}^k = h_t^k + \alpha\, PE_t^k = h_t^k + \alpha\,(I_t^k - h_t^k)$. How I_t^k could be derived from a Pavlovian theory of inhibitory learning still requires theoretical and empirical research.

A further issue concerns the motivation of free-operant avoidance that is evident in demonstrations of Pavlovian-instrumental transfer (PIT). Having trained dogs to shuttle back and forth over a barrier to avoid a shock on a free-operant avoidance schedule, Rescorla and LoLordo confined each dog on one side of the barrier where they received unavoidable shocks, each signaled by a tone. When subsequently presented while the dogs were engaged in shuttling,

the tone increased the rate of responding (Rescorla & Lolordo, 1965). It is unlikely that increase was due an expectation of a shock during the tone because LoLordo subsequently demonstrated that a signal for a loud aversive klaxon produced as great an elevation of shock-reinforced avoidance responding by the dogs (LoLordo, 1967), a finding recently replicated with rats (Campese et al., 2020). Rather it would appear that the avoidance PIT reflects a general motivating effect of aversive signals.

In our discussion of positively reinforced free-operant behavior (Perez & Dickinson, 2020), we attributed general appetitive PIT to a motivational influence on habitual responding, and it is possible that general aversive PIT also operates through the habit system. To reiterate, we appealed to a Konorskian two-process mechanism whereby the habitual avoidance response was self-reinforced by the aversive Pavlovian inhibition conditioned to its feedback stimuli through their negative temporal correlation with the shock. There is now good evidence that this form of conditioned inhibition is mediated by the aversive excitation conditioned to the contextual stimuli. For example, Miller and colleagues reported that extinguishing the aversive excitation elicited by the contextual stimuli following inhibitory conditioning reduced the subsequent inhibition exerted by the conditioned stimuli that had been previously trained under a negative correlation with the shock in the context (Miller, Hallam, Hong, & Dufore, 1991). In this sense, inhibition is a 'slave' process to excitation and consequently the self-reinforcement of responding during a strong conditioned excitor will be enhanced in the PIT procedure.

It is also possible that the aversive PIT is mediated by the goal-directed system. We argued that incentive values of reinforcers in the goal-directed system (Perez & Dickinson, 2020) are acquired through a process of instrumental incentive learning that enables a motivational state to control the current incentive value of a reinforcer (Dickinson & Balleine, 1994). Presumably, avoidance learning, especially during the early stages, takes place while the animal is fearful with the result that this state comes to control the negative incentive value assigned to the animal's representation of the shock, for example. As a consequence, the shock representation has a more negative incentive value when the animal is fearful than when it is not. Aversive PIT could therefore be due to an increment in negative incentive value of the shock representation during the fear-inducing signal.

Whatever the processes by which Pavlovian conditioning modulates avoidance, it is likely that it also contributes to the extinction of avoidance through the extinction of aversive excitatory conditioning to the context. According to our dual-system account (Perez & Dickinson, 2020), extinction of rewarded behavior is a complex of interacting systems. As the rate correlation is undefined when the STM is cleared of any reinforcer representations following the onset of extinction, g remains fixed at the weighted average value after the last recycle containing an outcome representation. However, as g predicts the same rewarding outcome that reinforced the habit, g contributes to the prediction error generating the habit strength, h. As a consequence, during extinction h acquires a sufficient negative value to counteract the contribution of the terminal positive g to responding. By contrast, this interaction does not occur

in avoidance because the outcome represented by the goal-directed system and the reinforcer of habits are different events: the shock and the feedback stimuli, respectively. Therefore, the residual *g* could maintain persistent responding.

This predicted persistence might be thought to be a virtue of the model as it often claimed that avoidance responding is abnormally persistent in extinction. However, there is little reason to believe this this persistence is a feature of free-operant avoidance. Perhaps the most pertinent study is one by Uhl and Eichbauer in which rats were trained to avoid a shock by lever pressing on a VC schedule similar to that employed by de Villiers (1974) before responding was extinguished by omitting the shock (Uhl & Eichbauer, 1975). Persistence of avoidance responding was contrasted with that following VC training in which a lever press during a cycle delivered sugar water to hungry rats at the end of the cycle. Following both positive (reward) and negative (avoidance) reinforcement training, extinction was similarly rapid with responding during the first 3-h extinction session being about a fifth of that at the end of training which Uhl and Eichbauer attributed to generalization decrement produced by the absence of the reinforcer (Uhl & Eichbauer, 1975). In agreement with this account, transferring from extinction to a non-contingent shock presentations under a variable time (VT) schedule immediately reinstated avoidance responding to the level seen in the last reinforced session.

This marked loss of performance in extinction contrasted with that observed when the rats were transferred from the VC to a VT schedule under which the reinforcer is delivered with the same recycling time as during training but

independently of responding. Although relative to the terminal VC rates, initial avoidance was marginally more persistent than rewarded responding, in both cases the level was greatly elevated above that in extinction before progressively declining to a low level. Contrasting performance under the VC schedule with that under the VT schedule rather than standard extinction provides an unconfounded measure of the impact of positive (reward) and negative (avoidance) instrumental contingencies by controlling the role of the discriminative and Pavlovian functions of reinforcement. For both rewarded responding and avoidance, the goal-directed system predicts a progressive decline of responding as *g* gradually decreases as cycles of the STM with a zero rate correlation accumulate. Similarly, the feedback stimuli lose their inhibitory properties as the avoidance response, if anything, is followed by a shorter time to the next shock than the inter-shock interval of the VC schedule. In summary, in according with Uhl and Eichbauer's (1975) results, our dual-system theory predicts not only that a VT schedule should produce greater persistence than extinction but also that positive and negative reinforcement training should yield similar profiles of responding as a common rate correlation mechanism mediates rewarded responding and avoidance.

When taken in conjunction with Perez and Dickinson (2020), the dual system model of free-operant behavior that incorporates a rate correlation system accounts for two of the four interactions between the response-reinforcer contingency and the valence of the reinforcer: positive reinforcement (reward) when both factors are positive and negative reinforcement (avoidance) when both factors are negative. In the case of an omission schedule under which the

contingency is negative and valence positive, Dickinson and colleagues failed to find any evidence for goal-directed control (Dickinson et al., 1998): devaluing the omitted reinforcer had no impact on the level of response reduction produced by prior omission training. Whether or not the response reduction produced by free-operant punishment, under which the contingency is positive and the valence negative, has a goal-directed component remains unexamined to the best of our knowledge. Therefore, the role of goal-directed control in punishment and omission training remain as important empirical issues.

Declarations

Funding

This work was supported by a SIA grant from ANID (SIA 85220023), a FONDECYT ANID grant, and the Instituto Sistemas Complejos de Ingenieria ANID PIA/PUENTE AFB220003, all of them awarded to Omar D. Perez.

Conflicts of interest/Competing interests

Not applicable.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Code availability

The code for the simulations presented in this paper can be found at github.com/omadav/2s_avoidance.

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B : Comparative and Physiological Psychology*, *33*(2), 109–121. https://doi.org/10.1080/14640748108400816

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

Baum, W. (2020). Avoidance, induction, and the illusion of reinforcement. *Journal of the Experimental Analysis of Behavior*, *114*(1), 116–141. https://doi.org/10.1002/jeab.615

Daw, N. D., & O'Doherty, J. P. (2013). Multiple Systems for Value Learning. In *Neuroeconomics: Decision Making and the Brain: Second Edition* (pp. 393–410). https://doi.org/10.1016/B978-0-12-416008-8.00021-8

de Villiers, P. A. (1974). The Law of Effect and avoidance: A Quantitative Relationship Between Response Rate and Shock-Frequency Reduction. *Journal of the Experimental Analysis of Behavior*, *21*(2), 223–235. https://doi.org/10.1901/jeab.1974.21-223

DeVito, P. L., & Fowler, H. (1994). Positive and negative transfer of conditioned aversive stimuli to a conditioned appetitive excitor as a function of aversive US intensity. *Animal Learning & Behavior*, *22*(2), 195–202.

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

Dickinson, A., & Balleine, B. W. (1994). Motivational control of goal-directed

action. In N. Mackintosh (Ed.), *Animal Learning & Behavior* (Vol. 22, pp. 1–18). London: Academic Press. https://doi.org/10.3758/BF03199951

Dickinson, A., & Dearing, M. F. (1979). Appetitive-aversive interactions and inhibitory processes. *Mechanisms of Learning and Motivation: A Memorial Volume to Jerzy Konorski*, *203*, 231.

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

Dickinson, A., Squire, S., Varga, Z., & Smith, J. W. (1998). Omission learning after instrumental pretraining. *The Quarterly Journal of Experimental Psychology: Section B*, *51*(3), 271–286.

Dinsmoor, J. A. (2001). Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing. *Journal of the Experimental Analysis of Behavior*, *75*(3), 311–333. https://doi.org/10.1901/jeab.2001.75-311

Dinsmoor, J. A., & Sears, G. W. (1973). Control of avoidance by a response-produced stimulus. *Learning and Motivation*, *4*(3), 284–293.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Fernando, A., Urcelay, G., Mar, A., Dickinson, A., & Robbins, T. (2014a). Free-operant avoidance behavior by rats after reinforcer revaluation using opioid agonists and D-amphetamine. *Journal of Neuroscience*, *34*(18), 6286–6293.

Fernando, A., Urcelay, G. P., Mar, A. C., Dickinson, A., & Robbins, T. W. (2014b). Safety signals as instrumental reinforcers during free-operant avoidance. *Learning & Memory*, *21*(9), 488–497.

Heyes, C., & Dickinson, A. (1990). The intentionality of animal action. *Mind {&} Language*, *5*(1), 87–103. https://doi.org/10.1111/j.1468-0017.1990.tb00154.x

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. CUP Archive.

Konorski, J. (1967). *Integrative activity of the brain; an interdisciplinary approach*. University of Chicago Press.

Laurent, V., Balleine, B. W., & Westbrook, R. F. (2018). Motivational state controls the prediction error in Pavlovian appetitive-aversive interactions. *Neurobiology of Learning and Memory*, *147*(November 2017), 18–25. https://doi.org/10.1016/j.nlm.2017.11.006

Lovibond, P. (2006). *Fear and avoidance: An integrated expectancy model.*

Perez, O. D., & Dickinson, A. (2020). A Theory of Actions and Habits: The Interaction of Rate Correlation and Contiguity Systems in Free-Operant Behavior. *Psychological Review*. https://doi.org/10.1037/rev0000201

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1–5. https://doi.org/10.1037/h0025984

Rescorla, R. A. (1969). Conditioned inhibition of fear resulting from negative CS-US contingencies. *Journal of Comparative and Physiological Psychology*, *67*(4), 504.

Rescorla, R. A., & Lolordo, V. M. (1965). Inhibition of avoidance behavior. *Journal*

of Comparative and Physiological Psychology, *59*(3), 406.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*(3), 151.

Seligman, M., & Johnson, J. (1973). *A cognitive theory of avoidance learning.*

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

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

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

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Uhl, C. N., & Eichbauer, E. A. (1975). Relative persistence of avoidance and positively reinforced behavior. *Learning and Motivation*, *6*(4), 468–483.

Wang, O., Lee, S. W., O'Doherty, J., Seymour, B., & Yoshida, W. (2018). Model-based and model-free pain avoidance learning. *Brain and Neuroscience Advances*, *2*, 239821281877296. https://doi.org/10.1177/2398212818772964

Weisman, R. G., & Litner, J. S. (1969). Positive conditioned reinforcement of Sidman avoidance behavior in rats. *Journal of Comparative and Physiological Psychology*, *68*(4), 597–603.