LEARNING HOW TO LEARN BY SELF-TUNING REINFORCEMENT

CHRISTIAN TORSELL

JEFFREY A. BARRETT

Abstract. Humans and many animals are capable of learning and learning how to learn better. We are concerned here with one way that reinforcement learners might learn how to learn better. In an experiment described by Harry Harlow (1949), a group of rhesus monkeys learn a new way of learning in the context of a specific type of problem. We will consider how such agents might coevolve a new learning dynamics and new attendant saliences. To this end, we propose a self-tuning dynamics that illustrates one way that a reinforcement learner might acquire forms of learning that are well-suited to context-specific problems.

1. INTRODUCTION

David Hume believed that we can never have rational justification for our expectations or beliefs regarding matters of fact (1975, 25–39). In his sceptical solution to the problem, he explained that beliefs regarding matters of fact, and expectations regarding the future in particular, were naturally learned not rationally justified. This shift in focus from rational justification to learning has both pragmatic and naturalistic virtues.

Hume held that beliefs regarding expectation and matters of fact were produced from experience by means of *custom* or *habit*. Custom, in the sense in which he used the term, is a principle of our psychological nature that acts to produce and adjust propensities when presented with experience. He explained that "wherever the repetition of any particular act or operation produces a propensity to renew the same act or operation, without being impelled by any reasoning or process of the understanding . . .this propensity is the effect of *Custom*" (1975, 43). We learn just as animals do who "by the proper application of rewards and punishments, may be taught any course of action" (1975, 105). To learn by custom, then, is to learn by reinforcement on success and punishment on failure.

Hume regarded such reinforcement learning to be a fortunate natural endowment of human psychology:

> Custom . . . is the great guide of human life. It is that principle alone which renders our experience useful to us, and makes us expect, for the future, a similar train of events with those that have appeared in the past. Without the influence of custom, we should be entirely ignorant of every matter of fact beyond what is immediately present to the memory and senses. We should never know how to adjust means to ends, or to employ

> our natural powers in the production of any effect. There would be an end at once of all action, as well as of the chief part of speculation (1975, 44–45).

And he took its effect to be as "unavoidable as to feel the passion of love, when we receive benefits; or hatred, when we meet with injuries" (1975, 46).[1]

There is a great deal of evidence that Hume was right to believe that both humans and other animals very often learn by means of some form of reinforcement with punishment.[2] That said, humans and many animals are also capable of learning in other context-specific ways.

A natural extension of Hume's account of how we learn would consider how a reinforcement learner might develop and learn to implement other methods of learning, methods better suited to particular practical contexts. Barrett (2023) takes up this theme, offering a general Humean strategy for how an agent might use simple reinforcement to learn how to learn better. Here, in contrast, we focus on a narrower problem regarding natural learning. Specifically, we consider how rhesus monkeys might

[1]See Barrett (2023) and Morris and Brown (2019) for further discussion of the nature and role of custom in Hume's account of learning and his "sceptical solution" to the problem of induction.

[2]See Herrnstein (1970) for a description of such experiments and a formal account of positive reinforcement. See Roth and Erev (1995), Erev and Roth (1998), and Bereby-Meyer and Erev (1998) for examples of reinforcement learning in humans subjects. See Fudenberg and Levine (1998) and Skyrms (2010) for discussions of reinforcement and closely-allied types of learning in the context of games and Huttegger (2017) for a discussion of a discussion of the basic ideas behind reinforcement learning and rational learning more generally.

learn how to learn more efficiently using a form of self-tuning reinforcement learning in the context of a classic experimental study by Harry Harlow (1949).[3] This dynamics allows a reinforcement learner to learn how to learn in a way that is better suited to a particular type of problem while also learning how to apply the new form of learning to the problem. We take this sort of reinforcement learning to accord well with Hume's commitment to custom.

The argument proceeds as follows. In section 2 we introduce two kinds of learning, reinforcement and win-stay/lose-shift. In section 3 we explain how the learning to learn achieved by Harlow's subjects can be thought of as a gradual transition from the former to the latter. In sections 4 and 5 we argue that this transition is well modeled by a kind of "heating up" where the two parameters governing a reinforcement learner's behavioral and attentional dispositions are gradually increased over time. In sections 6 and 7 we present a model where the heating-up process is realized by a higher-order reinforcement process that operates on a learner's dispositions to stay with or shift away from actions depending on whether they recently led to practical success or failure. This second model shows that learning to learn of the kind achieved by Harlow's monkeys can be accomplished by a self-tuning process that involves nothing more sophisticated than reinforcement of strategic dispositions when they produce successful actions and

[3]Harlow discusses learning to learn in terms of *learning set formation*, the development of different methods of learning suited to different practical contexts. The 1949 paper we discuss was the first of several papers, by Harlow and others, investigating the development learning sets in animals. The original paper remains a classic in comparative psychology. See Schrier (1984) for more details on the reception of Harlow's research and its influence on subsequent work on animal learning. See Barrett (2024) for a discussion of other self-tuning forms of reinforcement learning.

punishment of strategic dispositions when they produce unsuccessful actions. The two-tiered formulation of reinforcement with punishment that we describe is both simple and highly adaptable. In section 8 we briefly discuss the results.

2. Two forms of learning

Edward Thorndike (1898) was one of the first to investigate in detail how animals learn by reinforcement with punishment. He summarized the results of his experiments on cats, dogs, and chicks in two laws. The first was the law of effect:

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.

The second was the law of exercise:

> Any response to a situation will, other things being equal, be more strongly connected with the situation in proportion to the number of times it has been connected with that situation and to the average vigor and duration of the connections. (Thorndike 1911, 244)

Together these laws capture the key features of reinforcement with punishment. Namely, an animal is more likely to perform an action when it has been rewarded in connection

with that type of action, less likely to perform it when it has been punished, and both the magnitude and the number of rewards and punishments matter in a cumulative way to the animal's subsequent probabilistic dispositions.[4]

In its most basic form *reinforcement with punishment learning* can be modeled as follows.[5] Let $q_k(t)$ be an agent's propensity for action k at time t. Her propensities evolve according to the update rule:

$$q_k(t+1) = \begin{cases} q_k(t) + \pi(t) & \text{if action } k \text{ was taken} \\ q_k(t) & \text{otherwise.} \end{cases}$$

Here $\pi(t)$ is the payoff received by an agent taking the action k on round t. It may be positive for reinforcement or negative for punishment depending on the degree of success or failure resulting from the action. If one allows for punishment, then one needs to do something to prevent negative propensities. One strategy is to specify a limit $b > 0$, then to stipulate that if a punishment would result in $q_k(t+1) < b$, then $q_k(t+1) = b$.

An agent's propensities, in turn, determine her probabilistic dispositions. This works by means of the response rule:

$$p_k(t) = \frac{q_k(t)}{\sum_j q_j(t)},$$

where $p_k(t)$ is the probability that the agent takes action i at time t. In order to say how the process gets started, one must also specify a set of initial propensities $q_k(0)$.

While both humans and animals often learn by this or a similar variety of reinforcement, they also learn in other ways. Sometimes an agent considering multiple

[4]See Thorndike (1898) (1901) (1911) for descriptions of his experiments and his understanding of how reinforcement learning works.

[5]This way of characterizing the dynamics follows Roth and Erev (1995) and Erev and Roth (1998).

possible actions begins by picking one at random. If her guess leads to success, then she repeats the same action the next time she finds herself in a similar situation. But if her attempt results in failure, then she tries a different response at random the next time around.

Win-stay/lose-shift learning formalizes this type of trial-and-error learning. Consider a learner who confronts a series of trials each of which results in either success or failure depending on which of a finite number of acts she chooses on that trial. As above, we will use t to denote the current time-step. At $t = 0$, a win-stay/lose-shift learner chooses each available act with equal probability. At each subsequent step, if she chose act a at t and that choice led to successful action, then she chooses a again at $t + 1$; if she chose a at t and failed on that trial, she chooses an act at random and without bias from the set of all available acts except for a at $t + 1$.

Win-stay/lose-shift does better than reinforcement in some learning problems. Harlow (1949) presented rhesus monkeys with a series of such problems and recorded their behavior. The monkeys started as reinforcement learners then slowly learned how to learn by win-stay/lose-shift in a context-specific way that involved the coevolution of what they took to be salient as they learned. We show that the learning accomplished by Harlow's monkeys is well-modeled by a process in which they gradually shift from implementing a simple form of reinforcement learning to implementing a learning rule that closely approximates the behavior of a win-stay/lose-shift learner even as they learn by reinforcement how to apply the new form of learning to the particular type of problem they face.

One might think of this co-evolutionary process as a self-assembling discrimination game.[6] In a self-assembling game, structural features of a strategic interaction, such as the payoff structure or the players' strategy sets, evolve alongside the strategic dispositions of the players. In the game played by Harlow's monkeys, both the learning dynamics they use to update their dispositions and the features of the world on which they condition their actions coevolve as they play.

3. Harlow's monkeys

Harry Harlow (1949) performed a series of experiments to determine how rhesus monkeys might learn how to learn in the context of a particular type of problem. Here we are primarily concerned with his first experiment.

In Harlow's first experiment, the monkeys were presented with a series of discrimination problems. Each problem consisted of a different pair of objects O_1 and O_2 that were easily distinguishable, with one of these, say O_1, always covering a small piece of food. As a concrete example, O_1 might be a handkerchief and O_2 a small pillow for a given problem. The two objects were then placed randomly to the left and right before the monkey. The monkey was rewarded if it chose the object covering the food. Each problem was repeated a number of times with the objects O_1 and O_2 randomly placed before the monkey on each trial and with the same object O_1 always covering the food. Then the experimenter introduced a new learning problem with two new objects and with the food always under one of those. The full experiment consisted of a series of 344 such problems using 344 different pairs of stimuli (objects) run on a group of eight monkeys (1949, 52).

[6]See Barrett and Skyrms (2017) for a general account of the self-assembly of games by ritualization.

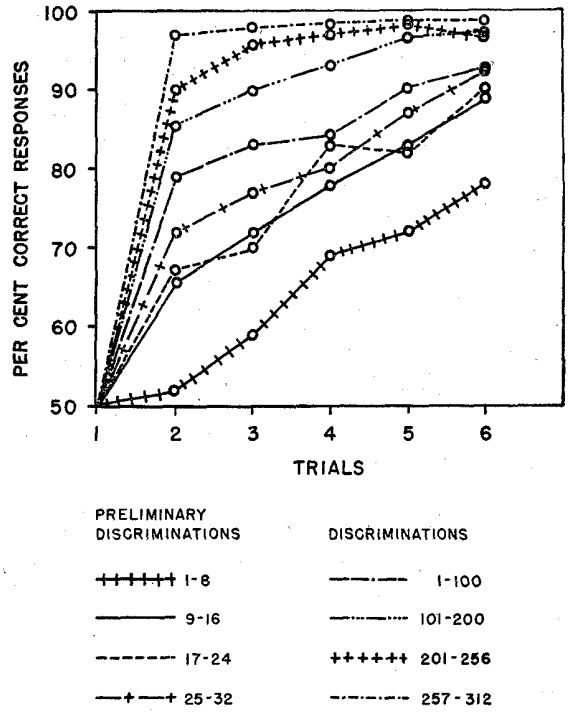

FIGURE 1. Harlow's 1949 learning set data: Discrimination learning curves on successive blocks of problems.

Harlow found that the monkeys initially learned to select the right object within a problem by means of a process that is closely modeled by simple reinforcement learning. But in later problems, the monkeys learned where the food was much faster and in a qualitatively different way. Figure 1, reproduced from Harlow's original paper, captures this phenomenon visually, plotting the monkeys' mean aggregate success rates in the first six trials of successive blocks of problems.

By learning across problems, the monkeys learned how to learn more efficiently within each problem. Instead of their usual reinforcement learning, they gradually began to learn by means of a form of a win-stay/lose-shift.[7] They would choose an object on their

[7]See Cochran and Barrett (2021) and (2022) for discussions of various forms of win-stay/lose-shift learning and their use by human subjects.

first trial blindly. If the food was there, they would stay with that object no matter where it might be located on a future trial. If the food wasn't there, they would choose the other object regardless of where it might be located on a future trial.[8]

Harlow referred to this type of acquired skill as a *learning set*, a way of learning in the context of a particular type of problem. In allowing for more efficient forms of learning, he said, the formation of a new learning set "delivers the animal from Thorndikian bondage." (1949, 59). The monkeys are no longer dependent on their usual reinforcement learning, a form of learning that does not work nearly as well as win-stay/lose-shift for the task at hand.

In learning how to learn better, the monkeys coevolve a new learning dynamics and new associated saliences that allow for the effective use of the new dynamics.[9] The monkeys' probabilistic dispositions gradually shift from those associated with reinforcement learning to those associated with win-stay/lose-shift learning over subsequent problems. And they learn that *objects* matter and *locations* don't, and they learn to use win-stay/lose-shift not simple reinforcement for this *type of problem*. In this way, the coevolved saliences provide conditions for both when and how the new dynamics is used.

[8]Harlow reported that some of the monkeys were eventually able to solve 20 to 30 consecutive problems with no errors whatsoever after their first blind trial (1949, 56).

[9]The notion of salience at work here is one on which a feature of an agent's environment is salient for that agent if she is disposed to notice and condition her response on that feature's state. When we speak of the saliences of an agent, we mean this as shorthand for the agent's dispositions to attend to and condition her actions on the various bits of her environment.

In brief, the monkeys begin as reinforcement learners who consider both position and object quality, then gradually learn to use win-stay/lose-shift on object quality.[10] In doing so, they self-assemble a new way of learning in the context of this particular type of task.

Harlow also described a series of experiments where children are presented with a similar task. The children learned in much the same way as the monkeys but were faster in moving from reinforcement learning to win-stay/lose-shift learning with the associated saliences (1949, 55 and 59).

While it is unclear precisely how the monkeys or children are learning how to learn, one can model how a reinforcement learner might learn to use win-stay/lose-shift with appropriate attendant saliences. We will consider how they might learn new saliences by reinforcing on what they attended to when their action was successful and how they might gradually evolve from learning by simple reinforcement to learning by win-stay/lose-shift by updating the magnitudes by which they reinforce on success and punish on failure as the play.

In the *first model*, we show that the monkeys' transition from gradual reinforcement learning to win-stay/lose-shift can be thought of as a kind of "heating up" of the monkeys' act- and salience-level learning, in which they are always learning by a form of reinforcement or punishment but the magnitudes by which they reinforce on success and punish on failure grow over time. The first model, however, does not consider how this heating-up process might be realized by a *learning* mechanism. This is addressed by the

[10]In his subsequent experiments, Harlow showed how the monkeys might learn to take position rather than object quality as salient and even switch between the two learned saliences (1949, 56–9). Of course, the monkeys are using pre-evolved and pre-learned saliences from the start. They must even to know that each trial involves making a choice.

second model, which we introduce in section 4. The second model shows how a shift in learning like that achieved by Harlow's monkeys might be accomplished by a learner who implements a self-tuning form of reinforcement with punishment learning over higher-order dispositions to repeat or shift away from actions depending on whether they were successful. [11]

4. THE FIRST MODEL

Consider an agent who learns by reinforcement with punishment over a sequence of learning problems both what to attend to and how to act. One might picture how she learns by considering a set of urns from which she might draw balls to determine her actions and add or remove balls to update her dispositions.[12] All draws from urns are random and without bias. We will start with a description of how learning occurs *within* a problem, then discuss how learning evolves across problems.

Each learning problem consists of a sequence of trials in which the agent chooses one of two objects O_1 or O_2. At the beginning of a problem, one of these objects is randomly selected as the reward object and remains the reward object for each trial of the problem. On problem n, the reward for success is i_n and the punishment for failure is j_n on each trial. The positions of the objects are randomly determined between trials.

[11]See Barrett (2020) and (2024) for how to model the evolution of salience and Herrmann and VanDrunen (2022) for an application to the evolution of saliences in the context of basic Lewis-Skyrms signaling games.

[12]We will also allow for fractional changes in the number of balls of each type in an urn. This will affect the probability of drawing a ball of a given type in precisely the way one would expect.

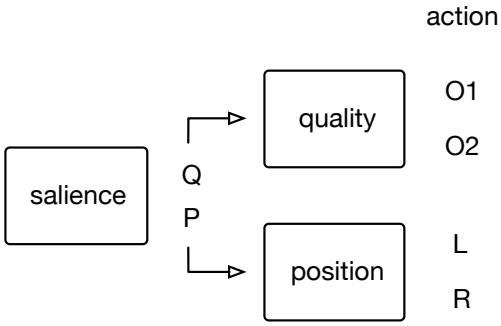

FIGURE 2. The urn model

At the beginning of a trial, the agent draws from a *salience urn* containing Q balls and P balls as in figure 2. Before the first problem, this urn contains one ball of each type. The result of the draw determines which type of stimulus the monkey attends to in determining her action.

If the agent draws a Q ball from the salience urn, then she chooses an object to select on the trial by a draw from her *quality urn*. Before the first trial of each problem, this urn contains one O_1 ball and one O_2 ball. If the agent chooses the reward object, then she is successful, and she returns the balls she drew from the salience urn and the quality urn and adds i_n new balls of the same type to each. If she selects the non-reward object, then she is unsuccessful, and as long as doing so will not drive the weight associated with the relevant type below a small $l > 0$, she returns the ball she drew to the urn from which she drew it then removes j many balls of its type from each of the two urns. If removing j_n balls would drive the associated weight below l, then she sets the weight associated with that type to l. The weights associated with each disposition are thus bounded from below by l. This prevents initially possible strategies from being completely eliminated.[13]

[13]While thinking of whole balls provides an intuitive picture of the process, the weights for each type in an urn are typically fractional. Specifically, each of the simulations in the

The process is analogous if a P ball is drawn from the salience urn. In this case, the agent determines which object to select on the trial by a draw from her *position urn*. Before the first trial of each problem, this urn contains one R ball and one L ball. If an R is drawn, the agent selects the object on the right; and if an L is drawn, she selects the object on the left. Reinforcement and punishment on the trial works the same way as it does on a quality draw.

An agent also adjusts how she learns between trials by updating the magnitude by which she reinforces on success i_n and punishes on failure j_n. Specifically, we will suppose that an agent's $(+i_n, -j_n)$ reinforcement with punishment learning evolves by the following recursive rule:

$$i_{n+1} = \alpha i_n + \beta$$

$$j_{n+1} = \alpha j_n + \beta$$

where $\alpha > 0$ and $\beta \geq 0$ are constant over the full multi-problem experiment. While a more complex model would allow for different scale and shift parameters for

---

next section starts with one ball of each type in each urn, a punishment level of 0.25, and a lower bound on each weight of $l = 10^{-14}$. The lower bound allows the reinforcement learner to retain an unsuccessful strategy that has a long track-record of failure as at least an in-principle possibility and perhaps even try it again later, particularly if nothing else has worked well either. While our particular choice of l is more or less arbitrary, choosing a value that is small relative to the initial propensities ensures that punishments occurring early in the learning process can significantly affect subsequent choice probabilities, as is necessary for the possibility of generating win-stay/lose-shift-like behavior in an agent implementing reinforcement learning with punishment.

reinforcement and punishment, we will suppose that the two parameters are the same in both contexts. We will also suppose that the magnitudes of reinforcements and punishments for act-level learning change from trial to trial according to the recursive rule, while the magnitudes of the reinforcement and punishments are constant for the learning of saliences.

An agent's salience urn is not reset between problems. This is so she might learn whether the *series of problems* she faces involve object quality or position. In contrast, her object and position urns are reset at the beginning of each new problem. This corresponds to the appearance of a new set of objects for which the agent needs to evolve effective dispositions.

While we are interested in the quantitative fit with Harlow's data, our primary concern is the basic structure of the model. We will start with a particularly simple set of parameters then discuss other settings that also work well.

Harlow's experiments do not allow us to determine precisely how the monkeys learn how to learn, but they do say something about how they learn how to learn in aggregate when repeatedly presented with the same special sort of problem.[14] The recursive rule is designed to capture this aspect of their higher-order learning. Later, we will consider a model in which this higher-order learning is modeled explicitly.

[14]Harlow also presents evidence from experiments where saliences may change from problem to problem. This is a step in the right direction, but to get a better understanding of the second-order dynamics of individual agents, one would need data for each agent rather than aggregate data. Further, it would also be useful to know how the monkeys behave when faced with problems where, for example, the reward alternates between objects within a problem.

Monotonically increasing act-level reinforcements and punishments represent a monkey's sharpening sense of the *type of learning problem* it faces in Harlow's first experiment. If the monkey tries an object and succeeds, then it will reinforce more on that object than it would have in earlier problems in the degree to which it has learned that when an object works in a problem, then it will work again if it tries it again. Similarly, if the monkey tries an object and fails, then it will punish more on that object than it would have in earlier problems in the degree to which it has learned that when an object doesn't work in a problem, then it will still not work if it tries it again.

5. FIRST MODEL: RESULTS

Following Harlow's experimental design, a single run of the model consists in a series of 344 problems. The first 32 problems involve 50 trials each, followed by 200 six-trial problems and 112 nine-trial problems.

The following parameters for a single simulated agent provide a close qualitative fit with Harlow's experimental data for the mean aggregate behavior of his eight monkeys:

$$i_1 = 1$$

$$j_1 = 0.25$$

$$\alpha = 1$$

$$\beta = 0.0004$$

Since $\alpha = 1$, this transformation just additively shifts the reinforcements and punishments with no rescaling between trials. And since $\beta = 0.0004$, the difference in learning dispositions between contiguous trials is small.

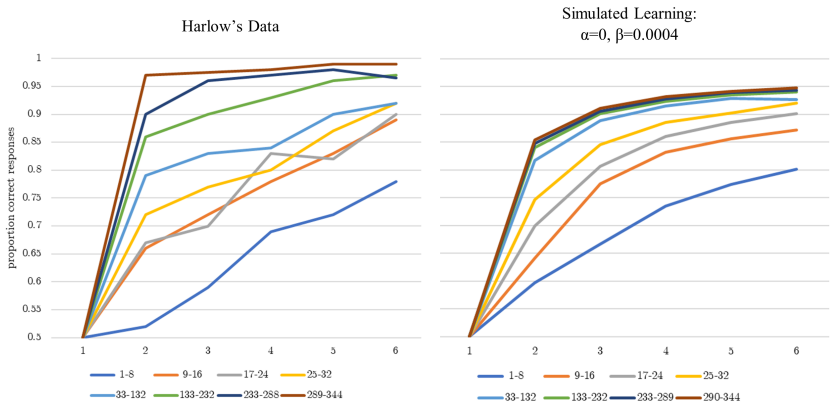

FIGURE 3. Harlow's experimental data compared to simulation data from our model on parameters $i_1 = 1$, $j_1 = 0.25$, $\alpha = 0$, $\beta = 0.0004$.

These parameters generate a sequence of learning curves that capture the steepening pattern across problems that Harlow reports in his experiment. This is illustrated in the comparison between Harlow's experimental data and the simulation data from the treatment where act-level and salience learning coevolves as illustrated in figure 3.[15]

Although we are primarily interested in the basic structure of the model and the qualitative steepening pattern reflecting the gradual transition from simple reinforcement to win-stay/lose-shift, it is worth noting that the quantitative fit with the experimental data is close. The model thus offers not only an account of how a reinforcement learner might in principle come to implement win-stay-lose-shift; it is able to closely approximate the aggregate learning data from a particular case of such learning to learn. That said, the closeness of the fit varies somewhat across problem blocks.

Figure 4 reports the mean absolute difference between the percent correct responses in Harlow's experiment and in the two simulated treatments using the parameter settings

[15]The experimental data is estimated from Harlow's figures. The resolution of each data point is approximately 2%.

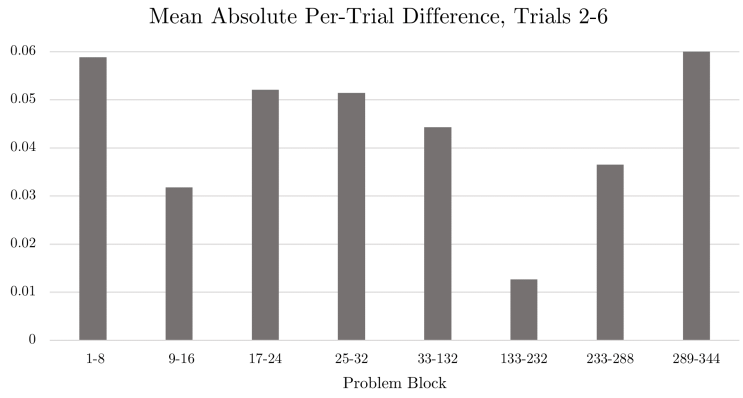

FIGURE 4. Mean absolute difference between the percent correct responses in Harlow's experiment (figure 1) and in the simulations. Only trials 2–6 are counted in each problem, and the average is taken over all of the trials in each block of problems.

above. Only trials 2–6 are counted in each problem. Starting with trial 2, the monkeys have the chance to shift on object choice if their first guess was incorrect, and we only use data up to trial 6 as that is all Harlow reports. Averages are taken over all of the trials in each block of problems.

As indicated in figure 4, the worst match between simulation results and the experimental data is in the first and last problem blocks, 1–8 and 289–344, but even here the difference between the predicted success rate on each treatment and the experimental data is never more than 6%. Given the relatively low resolution of the experimental data itself, this is a very close quantitative fit.

The upshot is that additively shifting reinforcements and punishments by a constant between trials provides an account of how the monkeys learn how to learn that fits well with the experimental data. As they update their own learning dynamics, they gradually come to act as win-stay/lose-shift learners. And they learn to pay attention to objects,

not locations, in the context of this type of problem. In this the simulated agents behave just as the rhesus monkeys do in aggregate in Harlow's experiment.

The model does surprisingly well under quite different parameter settings. Instead of increasing reinforcements and punishments by means of iterated additive shifts ($\alpha = 1$, $\beta > 0$), one might rescale reinforcements and punishments after each trial ($\alpha > 1$, $\beta = 0$). Starting with the same initial values for i_0 and j_0 as above, a pure rescaling of $\alpha = 1.0005$ and $\beta = 0$ delivers a qualitative overall fit approximately as good as the pure additive shift of $\alpha = 1.0005$ and $\beta = 0.0004$, and it does somewhat better on the final problem block than the additive shift. As with the original parameters, the mean absolute difference between the model's learning on the rescaling parameters and that of Harlow's monkeys for trials 2-6 never exceeds 0.06 for any problem block. Unsurprisingly, there are also parameter settings that involve both an additive shift and rescaling that provide a good qualitative match with Harlow's data.

The robustness of the model under different parameter settings means that the model gets the basic structure of the monkey's higher-order learning right. In particular, monotonically increasing levels of punishment and reward capture the aggregate shift in the dispositions of the monkeys as they evolve from reinforcement learners to win-stay/lose shift learners.

That different parameters settings work similarly well, however, also represents a significant limit regarding what can infer from Harlow's experiments. A second-order learning dynamics that works by simple reinforcement (additive shift) is different in kind from one that works by multiplicative reinforcement (rescaling).

There are two further things to note regarding the present model. The first concerns the evolution of salience. The second concerns the form of reinforcement learning required to capture the behavior of the monkeys.

To this point, we have only considered how an agent might learn that object quality is salient to learning within a problem, but the story is much the same for location. In subsequent experiments, Harlow tried always placing the reward in the same position rather than under the same object in a problem. The monkeys were able to learn how to learn in the context of such problems by win-stay/lose-shift on position just they had on object quality. The present model captures this behavior. Since quality and position are symmetric, if the reward is always put in the same position, a simulated agent gradually learns how to learn by win-stay/lose-switch on position rather than object quality, agreeing well with the aggregate behavior of Harlow's monkeys.

The second thing to note is that punishment is an essential feature of the present model. While there are parameter settings that allow for the emergence of something roughly akin to win-stay/lose-shift learning without punishment, one cannot get a good match with Harlow's aggregate data without punishment. The reason is relatively straightforward. Since the object urns are reset between problems, the expected success rate in trial 2 within a problem for a simple reinforcement learner without punishment but with optimal saliences is bounded from above by 0.75.[16] Hence no level of positive

[16]Suppose that with probability 1 a simple reinforcement learner without punishment attends to the relevant dimension in every trial of problem n. At the beginning of the problem, the object act urn is reset to one ball of each type. The agent will, hence, choose the rewarded object in trial 1 with probability 0.5. Let i_n be large so that the probability that the agent will choose the rewarded object in trial 2 conditional on having chosen the correct object in trial 1 is 1-ϵ, where ϵ is small. If the punishment level j_n is zero, then the agent will choose the rewarded object in trial 2 with probability 0.5, conditional on having chosen the unrewarded object in trial 1. Thus the probability of success in trial 2

reinforcement alone, can generate a success rate of 0.97 in trial 2, as observed in the final problems of Harlow's first experiment.

6. THE SECOND MODEL

The first model closely approximates the behavior of Harlow's monkeys as they learn how to learn in this type of discrimination problem. There is good reason, however, to hesitate in taking the model as illustrating how they *learn* how to learn. The modeled agent's salience- and act-level learning is transformed by iterated additive transformations of reinforcement and punishment values, but those transformations occur automatically between problems independently of the agent's experience. In a genuine learning process, one should expect an agent's dispositions to change over time *in response* to the specific content of her experience.

In this section, we consider higher-order learning process that would lead an agent to gradually shift from implementing slower reinforcement learning to fast win-stay/lose-shift-like learning in Harlow-style discrimination problems. The meta-learning dynamics that describes the evolution of an agent's first-order learning parameters is a variety of reinforcement learning in which the agent reinforces and punishes four higher-order strategies: (i) stick with strategies when they succeed, (ii) abandon strategies when they succeed, (iii) stick with strategies when they fail, and (iv) abandon strategies when they fail.

Reinforcing and punishing dispositions (i)-(iv) is operationalized in terms of adjustments of the reinforcement and punishment levels governing the agent's act- and

---

$\overline{\text{is } (0.5)(1-\epsilon) + (0.5)(0.5)}$. Letting ϵ go to zero for very large reinforcements, the value of expression asymptotically approaches 0.75 from below.

salience-level learning. Before describing the model in detail, it will be helpful to consider the motivation.

Note that for any fixed initial assignment of propensities over first-order acts, the higher a reinforcement learner's level of reinforcement, the *more* likely she is to repeat successful actions; and the lower her level of reinforcement, the *less* likely she is to repeat successful actions. Thus, increasing the level of reinforcement can be thought of as reinforcing the disposition to repeat actions which just led to success, and decreasing the level of reinforcement can be thought of as punishing the disposition to repeat actions which just led to success. Similarly, for any fixed initial assignment of propensities over first-order acts, the lower the learner's level of punishment, the *more* likely she is to repeat unsuccessful actions; and the higher the level of punishment, the *less* likely she is to repeat unsuccessful actions. So, decreasing the level of punishment can be thought of as reinforcing the disposition to repeat actions which just led to failure, and increasing the level of punishment can be thought of as punishing the disposition to repeat actions which just led to failure.

These observations provide the basis for the second model's higher-order reinforcement dynamics. Suppose that an agent performs two identical actions A_1 and A_2 under identical conditions at contiguous times t_1 and t_2. If A_1 was successful and A_2 was successful, then the agent would want to reinforce in the future more strongly than she did since an identical action succeeded when it was repeated. In contrast, if A_1 was successful and A_2 was unsuccessful, she would want to reinforce in the future less strongly than she did, since an identical action failed when it was repeated. Similarly, if A_1 was unsuccessful and A_2 was successful, then the agent would want to punish in the future less strongly than she did since an identical action succeeded when it was repeated. And

if A_1 was unsuccessful and A_2 was unsuccessful, she would want punish in the future more strongly than she did, since an identical action failed when it was repeated.

Consider an agent facing a series of n many k-trial object quality discrimination problems, formalized precisely as in the first model. And like the first model, suppose the learner's saliences and act-level dispositions evolve by reinforcement with punishment, as described by the urn model in figure 2. As above, let i_t represent the level of reinforcement for salience and act learning at time t, and let j_t represent the corresponding level of punishment at t, where timesteps are cumulative across problems (so that, e.g., the first trial of the second problem occurs at $t = k + 1$, not $t = 1$). Let $s(t)$ denote the salient dimension on that trial, i.e. the feature of the stimulus objects the learner attended to at t, and let $I_s(t, t + 1)$ be a function whose value is 1 if $s(t) = s(t + 1)$ and 0 otherwise. Let $o(t)$ denote the outcome of the trial at timestep t, where $o(t) = 0$ if the trial was unsuccessful and $o(t) = 1$ if the trial was successful. $a(t)$ will denote the act chosen at t, and $I_a(t, t + 1)$ is a function whose value is 1 if $a(t) = a(t + 1)$ and 0 otherwise. $\gamma > 1$ and $\lambda < 1$ are constants. Figure 5 describes precisely how reinforcement and punishment dispositions evolve, along with qualitative descriptions relating the formal characterization to the interpretation in terms of reinforcing and punishing on stay/shift dispositions.

In the present model γ and λ are constants by which the learner's reinforcement and punishment levels may be rescaled when the agent learns to learn from trial to trial. This involves several modifications of the original model. Most importantly, higher-order reinforcement on the present model is not automatic; rather, first-order levels of reinforcement and punishment are only modified as a result of higher-order learning on the basis of the agent's experience. In the considering results from the first model, we focused on the case in which reinforcement and punishment levels are updated by

$I_s(t, t+1)$	$I_a(t, t+1)$	$\langle o(t), o(t+1)\rangle$	i_{t+1}	j_{t+1}	qualitative description
1	1	⟨0,0⟩	i_t	γj_t	act chosen at t repeated at $t+1$, both trials unsuccessful; learner reinforces disposition to switch from unsuccessful actions
1	0	⟨0,0⟩	i_t	λj_t	act chosen at t not repeated at $t+1$, both trials unsuccessful; learner punishes disposition to switch from unsuccessful actions
1	1	⟨0,1⟩	i_t	λj_t	act chosen at t repeated at $t+1$, first trial unsuccessful but second successful; learner punishes disposition to switch from unsuccessful actions
1	0	⟨0,1⟩	i_t	γj_t	act chosen at t not repeated at $t+1$, first trial unsuccessful but second successful; learner reinforces disposition to switch from unsuccessful actions
1	1	⟨1,0⟩	λi_t	j_t	act chosen at t repeated at $t+1$, first trial successful but second unsuccessful; learner punishes disposition to stick with successful actions
1	0	⟨1,0⟩	γi_t	j_t	act chosen at t not repeated at $t+1$, first trial successful but second unsuccessful; learner reinforces disposition to stick with successful actions
1	1	⟨1,1⟩	γi_t	j_t	act chosen at t repeated at $t+1$, led to success both times; learner reinforces disposition to stick with successful actions
1	0	⟨1,1⟩	λi_t	j_t	act chosen at t not repeated at $t+1$, both trials successful; learner punishes disposition to stick with successful actions
0	---	---	i_t	j_t	shift in agent's framing of choice situation between t and $t+1$; no update

FIGURE 5. The higher-order learning dynamics (The last row indicates that on any trial in which the agent's salience shifted from the previous trial, reinforcement and punishment levels remain unchanged.)

additive translations; in the second model, higher-order reinforcement is accomplished by rescaling as this provides a natural way of avoiding the possibility of negative first-order reinforcement or punishment levels. And while the first model modifies reinforcement and punishment in lockstep, first-order reinforcement and punishment levels are never modified simultaneously in the present model. Rather, they respond independently to the learner's experience.

Another feature of the present model is that reinforcement and punishment levels remain unchanged whenever the agent's saliences shift between contiguous trials (i.e., whenever $I_s(t, t+1) = 0$). The thought is that which dimension of the stimuli is salient to the agent in a given trial determines a *framing* of the choice problem at hand and that choices made in trials with different frames are not comparable. In particular, it is not

meaningful to treat the sequence of choices made in two contiguous trials in which the learner switched saliences as an instance of *staying with* or *shifting from* a given strategy.

An example may be helpful. Consider a Harlow discrimination problem in which the two stimulus objects are a red bowl and a green cup. Suppose that object quality is salient to the learner in trial t. She therefore frames the problem at hand in terms of choosing the right kind of object. On trial $t+1$ her salience changes. Now, she attends to position, and thus sees the problem as requiring her to choose the correct *location*.

The example illustrates how a shift in the salience a learner uses marks a change in how she frames her options: in trial t, she faces a choice between different types of objects; in trial $t+1$ it faces a choice between different locations. Suppose she first (at t) chooses the green cup, and then (at $t+1$) chooses the right-hand position, which happens to be occupied by the green cup. Of course, from an outside perspective, the same object was selected between the trials. But from the learner's perspective, the acts *choose the right-hand position* and *choose the green cup* are not comparable in the way they would have to be in order for talk of the learner *keeping the same strategy* or *switching to a new strategy* between t and $t+1$ to be meaningful. As a result, we suppose that the agent does not update her dispositions to stay with or switch away from successful or unsuccessful actions after two-trial sequences when her saliences change between trials.

7. SECOND MODEL: RESULTS

To investigate whether the second model can replicate the desired shift from gradual reinforcement learning to win-stay/lose-shift, we ran a series of 1000 computer simulations, each consisting in 1000 blocks of 10-trial Harlow discrimination problems. Propensities for acts and saliences were bounded from below at 0.01. All initial

propensities were set to 1; γ was set to 1.02 and λ was set to 0.98. Initial reinforcement and punishment levels were $i_1 = 1$ and $j_1 = 0.25$, as in the earlier model.

On simulation, the modeled agent reliably learned that success in the type of problem she was given required her to use large reinforcement and punishment levels, which led to dispositions approximating those of win-stay/lose-shift learner. The mean cumulative success rate over all 1000 problems, averaged across the 100 runs, was 0.91. Restricted to the last 100 problems, the mean success rate across all runs rises to 0.929. This is close to the optimal expected success rate of 0.95 for a true win-stay/lose-shift learner (with fixed task-appropriate saliences) in this problem, indicating that the agent typically successfully learned to attend to the task-relevant dimension and to very closely approximate win-stay/lose-shift.[17]

The mean cumulative success rate on the last hundred problems was less than 0.9 on just 47 of the 1000 runs. On all but two of these runs, the agent had mistakenly learned to attend to the task-*irrelevant* dimension of the stimulus objects with high probability. On these runs, the agent performed approximately as well as chance, with mean success rates lying between 0.47 and 0.52.[18] It is unsurprising that the agent's performance was

[17]Recall that, conditional on its attending to the task-appropriate dimension, a win-stay/lose-shift learner will succeed on the first trial of a given problem half the time on average, and will succeed on every subsequent trial in that problem.

[18]For the two outliers, success rates for the last hundred problems were 0.57 and 0.62. In one of these runs, the agent had learned to attend to the task-relevant dimension with probability very close to 1; in the other, the agent's final probability of attending to object quality was 0.377.

close to chance in this case as the task-irrelevant dimension is uncorrelated with the location of the reward.[19]

8. CONCLUSION

It is natural to understand learning by Humean custom as learning by means of a form of reinforcement with punishment. But for custom to provide a compelling account of natural learning, one also needs to explain how an agent might start as a reinforcement learner, then learn how to learn in a manner well suited to a particular type of problem. Here we consider one way this might work in the context of a famous type of discrimination problem.

Harlow showed that his monkeys were able to learn how to learn by win-stay/lose-shift, a learning dynamics that is better suited to the type of problem they face than their default reinforcement learning, and that they were able to learn how to apply this new form of learning in a context-specific way by co-learning the saliences relevant to that type of problem. The first model illustrates how a simple reinforcement learner might learn saliences appropriate to the type of discrimination problem Harlow describes while gradually shifting to learning by means of win-stay/lose-shift. Building on this, the second model shows how a more subtle sort of reinforcement learner, one equipped with a higher-order dynamics that allows her to reinforce and punish the magnitudes of first-order reinforcements and punishments, might learn how to learn more effectively in a Harlow-style problem as she learns.

[19]When the 47 outlier runs are removed, the mean success rate over the final hundred problems of each run is 0.954, with noise just slightly better than the optimal expected success rate.

The learning dynamics in the second model is self-tuning. As an agent uses it, the higher-order dynamics provides a way for her to learn how to adjust her first-order learning to make it more effective given how well it is performing in the task at hand. We have shown that this form of reinforcement is highly effective in the context of Harlow-type problems. It is a topic for future research how well it will work in the context of other learning problems where tuning matters.[20]

Hume was right to believe that humans and other animals very often learn by a form of reinforcement. Here we have described a type of reinforcement learner who can learn how to learn in a way that is well suited to a type of problem for which simple reinforcement is not at all well suited. While Hume did not consider this type of self-tuning reinforcement, it is compatible with his insistence that we learn by means of custom. Here custom itself provides a mechanism for an agent to better learn by custom.

[20]Another place tuning matters is in Lewis-Skyrms signaling games. The sender and receiver in a basic $n \times n \times n$ signaling game can evolve successful signaling conventions quickly using reinforcement with punishment if the levels at which they punish and reinforce are well-tuned to the complexity of the game. If the level of punishment it too low, they get stuck in suboptimal pooling equilibria. If it too high, they cannot get the traction required to learn. See Barrett and Gabriel (2023) for simulation results and a discussion. Tuning also matters is in learning from neighbors on a network. As Zollman and others have shown, if one learns too quickly, this can generate consensus without sufficient exploration. How well suited the present dynamics is for problems like these is an open empirical question.

References

[1] Barrett, Jeffrey A. (2024) *Self-Assembling Games*. Forthcoming with Oxford University Press.

[2] Barrett, Jeffrey A. (2023) "Humean Learning (How to Learn)" *Philosohical Studies* Volume 181, pages 281–297.

[3] Barrett, Jeffrey A. (2020) "Self-Assembling Games and the Evolution of Salience," *British Journal for the Philosophy of Science*. https://www.journals.uchicago.edu/doi/10.1086/714789

[4] Barrett, Jeffrey A. and Brian Skyrms (2017). "Self-Assembling Games," *The British Journal for the Philosophy of Science*, 68(2), 329–353

[5] Beggs, Alan W. (2005) "On the Convergence of Reinforcement Learning," *Journal of Economic Theory* 122: 1–36.

[6] Bereby-Meyer, Yoella and Ido Erev (1998) "On Learning to Become a Successful Loser: A Comparison of Alternative Abstractions of Learning Processes in the Loss Domain." *Journal of Mathematical Psychology* 42(2–3): 266–286.

[7] Cochran, Calvin T. and Jeffrey A. Barrett (2022) "The Efficacy of Human Learning in Lewis-Skyrms Signaling Games."

[8] Cochran, Calvin T. and Jeffrey A. Barrett (2021) "How Signaling Conventions are Established," *Synthese* 199(1-2): 4367–4391.

[9] Erev, Ido and Alvin E. Roth (1998) Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review* 88: 848–81.

[10] Fudenberg, David, Drew Levine (1998) *Learning and the Theory of Games* MIT Press: Cambridge, MA.

[11] Harlow, Harry F. (1949) "The Formation of Learning Sets," *Psychological Review* 56:51–65.

[12] Herrmann, Daniel A. and Jacob VanDrunen (2022) "Sifting the Signal from the Noise," forthcoming in *The British Journal for the Philosophy of Science*.

[13] Herrnstein, Richard (1970) On the law of effect. *Journal of the Experimental Analysis of Behavior* 13:243–266.

[14] Hume, David (1975) *Enquiries Concerning Human Understanding and concerning the Principles of Morals*. Oxford: Oxford University Press.

[15] Huttegger, Simon (2017) *The Probabilistic Foundations of Rational Learning* Cambridge: Cambridge University Press.

[16] Morris, William Edward and Charlotte R. Brown (2019) "David Hume," *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2022/entries/hume/¿.

[17] Roth, Alvin E. and Ido Erev (1995) Learning in extensive form games: experimental data and simple dynamical models in the intermediate term. *Games and Economic Behavior* 8: 164–212.

[18] Schrier, A. M. (1984). "Learning How to Learn: The Significance and Current Status of Learning Set Formation." *Primates*, 25, 95-102.

[19] Skyrms, Brian (2010) *Signals: Evolution, Learning, & Information*, New York: Oxford University Press.

[20] Thorndike, Edward (1898) "Animal Intelligence: an Experimental Study of the Associative Processes in Animals" *T*he Psychological Review: Monograph Supplements, Vol. II., No. 4 (Whole No. 8), June, 1898. The Macmillan Company: New York and London.

[21] Thorndike, Edward (1901). The human nature club: An introduction to the study of mental life (2nd ed.). New York: Macmillan.

[22] Thorndike, Edward (1911). Animal intelligence. New York: Macmillan.