Philosophical Studies (2024) 181:281–297
https://doi.org/10.1007/s11098-023-02086-3

Humean learning (how to learn)

Jeffrey A. Barrett[1]

Accepted: 18 November 2023 / Published online: 19 December 2023

Abstract
David Hume's skeptical solution to the problem of induction was grounded in his belief that we learn by means of custom . We consider here how a form of reinforcement learning like custom may allow an agent to learn how to learn in other ways as well. Specifically, an agent may learn by simple reinforcement to adopt new forms of learning that work better than simple reinforcement in the context of specific tasks . We will consider how such a bootstrapping process may lead to a system that includes trial-and-error forms of learning like win-stay/lose-shift, probe and adjust, and simple reinforcement itself together with higher-rationality inferential tools.

Keywords Humean learning · The problem of induction · Hume's skeptical solution · Learning how to learn · Pragmatism

1 Introduction

David Hume was skeptical regarding our ability to rationally justify beliefs concerning matters of fact, but he held that we nevertheless routinely learn matters of fact by means of custom. This sort of instinctive learning might be understood as a form of reinforcement where an agent's dispositions to act are strengthened on success and perhaps weakened on failure in action. While Hume was right to suppose that humans, like other animals, very often learn by means of reinforcement, we also learn in other ways.

We consider here how an agent may be led by simple reinforcement to adopt new forms of inductive learning. By such means she may evolve a learning system that includes trial-and-error forms of learning like win-stay/lose-shift, probe and adjust, and simple reinforcement itself together with higher-rationality inferential tools.[1]

[1] See Erev, I., Roth, A. E. (1996), Fudenberg and Levine (1998), Erev and Roth (1998), Bereby-Meyer and Erev (1998), Barrett and Zollman (2009), Skyrms (2010), Huttegger (2017), Cochran and Barrett (2021,

✉ Jeffrey A. Barrett
jabarret@uci.edu

[1] University of California Irvine, Irvine, USA

While simple reinforcement may lead an agent to adopt forms of learning that are more sophisticated or better-suited to the inferential tasks she faces, there is no magic. Even when she is led to adopt a learning rule that has been highly reliable for a particular purpose, it may fail to work well in the future.[2] That said, simple reinforcement provides a reliable way of tracking which learning rules have worked well, and insofar as reinforcement on success is in fact psychologically efficacious in tuning our dispositions, this will lead one to evolve more sophisticated forms of learning regardless of whether one is rationally justified in doing so. This will serve the agent well in future action should those rules continue to work.

2 Humean learning

Hume believed that we can never have rational justification for our expectations or beliefs regarding matters of fact.[3] While one may observe that an event of type *A* has always been followed by an event of type *B*, constant conjunction fails to entail any necessary connection. No sequence of past conjunctions, no matter how extensive, provides any reason whatsoever for concluding even that the occurrence of *A* makes the occurrence of *B* more likely. To get something like this, one would need to assume that what has happened in the past is a reliable guide to what will happen in the future, but such an assumption begs the question. Even if the past has in some ways been a reliable guide to the future in the past, it need not be in the future. As a result, our experience provides no ultimate justification for any beliefs at all regarding future events. And since the same line of argument applies to conclusions regarding causal relations generally, only by means of which Hume argued can one learn matters of fact, one can have no ultimate justification for believing any matter of fact (1975, 25–39).

This poses an immediate problem for rational action. Inasmuch as one cannot infer anything concerning the future from the past, Hume held that one can never have any rational justification for acting one way rather than another. That said, there is an important sense in which he was not at all skeptical regarding the expected efficacy of his actions or his judgments regarding matters of fact more generally. Understanding the position requires some care.

Footnote 1 (continued)

2022), Barrett and Gabriel (2022), and Barrett (2023) for descriptions and discussions of a great many alternative forms of learning. Each has potential virtues and vices depending on the learning problem at hand and the resources available to the learner. See Barrett (2023) for an extended discussion of learning how to learn and reflections on how various basic and task-specific forms of learning might self-assemble.

[2] In this regard, note that any particular learning algorithm *R*, no matter how subtle or sophisticated it may be, may routinely fail to provide successful predictions. Consider a world where whenever one learns by *R* to expect *E* on the basis of one's evidence so far ¬*E* occurs. See Putnam (1963) for a more elaborate version of this argument.

[3] He argued for this in both *A Treatise of Human Nature* (1739–40) and *An Enquiry Concerning Human Understanding* (1748). Here we will follow the argument of the latter. Regarding learning, we follow his natural propensity account grounded in custom.

Hume explicitly recognized that he, like everyone else, was in fact firmly committed to a rich collection of beliefs regarding future events and matters of fact. Further, he found that he remained committed to these beliefs even when he knew that he possessed no ultimate justification for believing them. As a result, he was perfectly comfortable using beliefs that he had formed in the context of experience to guide even his most important actions (1975, 42).

Hume held that beliefs regarding matters of fact, and expectations regarding the future in particular, 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. Hume 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). In other words, we learn just as animals do who "by the proper application of rewards and punishments, may be taught any course of action." The upshot is that, rather than being an activity grounded in reason, the ability to engage in empirical inquiry is one that "we possess in common with beasts" and "is nothing but a species of instinct or mechanical power, that acts in us unknown to ourselves" (1975, 108).[4]

It is by custom, then, that each repeated instance of a pattern of events strengthens an agent's dispositions to act as if that pattern will continue to hold in the future. In identifying human and animal learning, Hume suggests that to learn by custom is to learn by reinforcement on success and punishment failure in expectation and action.

That we learn matters of fact and form expectations regarding the future by reinforcement meshes well with Hume's insistence that inductive learning does not involve rational inference. Even someone with a broader understanding than Hume's concerning what should count as a rational faculty might find room for at agreement. An agent who learns by reinforcement may simply update her dispositions to act as a consequence of her past experience. To do so, she need not know any rules of logic or the ways of probabilistic inference or even that she is learning at all (1975, 41–2). The process may be entirely unreflective.[5]

Hume took the fact that we learn by reinforcement on experience to be a fortunate feature of our psychological nature. Reason is unable to justify our beliefs regarding matters of fact or our expectations, but even if it could, its psychological effects are too weak to guide us in practical action. In contrast:

[4] Again, on Hume's naturalistic account, custom is just an irresistible "mechanical tendency" to update our propensities (1975, 42 and 55). Its effect is as "unavoidable as to feel the passion of love, when we receive benefits; or hatred, when we meet with injuries" (1975, 46). See Allison (2008) and Morris and Brown (2019) for discussions of the nature and role of custom in Hume and Sect. 2 of Henderson (2022) for a discussion of the options available, given Hume's account of custom, for understanding the scope of his skeptical conclusion.

[5] Hume moves easily between custom as a principle that produces propensities and custom as a principle that produces beliefs. While this aspect of his epistemology suggests a dispositional account of belief, one can also get his main conclusions if custom produces dispositions of a sort that lead to appropriate corresponding beliefs.

> 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).

What matters for practical action is learned belief not justified belief. Shifting the focus from rational justification to how we in fact learn is the key move in Hume's naturalistic approach to practical knowledge. Empirical inquiry is a matter of learning by custom, not rationally justifying beliefs or predictions.

Hume's commitment to custom as a reliable and irresistible guide explains his own empirical practice. While he did not believe that the past success of custom in forming reliable beliefs provided rational justification for its use, he held that everyone, including himself, will nevertheless form beliefs and expectations regarding matters of fact by reinforcement on experience and that everyone, including himself, will find themselves believing that this practice will in fact lead to similar successes in the future to what it has in the past. His empirical account of human nature explains both of these points. In each case, custom leads one to form such beliefs in the context of regular past experience. Indeed, that one cannot resist the effects of custom explains his own acceptance of his empirical psychology on empirical grounds. That is, Hume's account of human nature explains why he could not help but accept that very account given his empirical evidence even when he knew full well that the evidence he marshaled provided no rational justification for accepting it.

The efficacy of custom, then, explains why Hume accepted the principles of his empirical psychology on empirical grounds while simultaneously recognizing that there is nothing that rationally justifies this acceptance. And it explains why he believed that his readers would be similarly convinced when they reflected on the evidence he provided regarding the efficacy of custom. It is not because they have rational justification for being convinced. Rather, it is because they form their beliefs in the same way he forms his, by means of reinforcement on experience. Empirical inquiry is a form of cognitive conditioning, and they will be led by custom and their experience to similar beliefs.

Hume's skeptical solution to the problem of induction trades justification for learning. In this, it is both naturalistic and pragmatic.[6] He was right to suppose that custom is in fact an essential guide to action for both animals and humans. We have a long history of experimental evidence that reinforcement learning, in its various forms, is ubiquitous in nature. These forms of learning evolved because they have

[6] See Henderson (2022) for a discussion of Hume's skeptical solution and a survey of approaches to the problem of induction. Our aim here is to investigate the potential scope of Hume's skeptical solution starting with how we in fact learn. This is in contrast with the long tradition of seeking to specify a meta-inductive practice that would ultimately justify one's inductive predictions. See Schurz (2008, 2019) for a recent example.

afforded adaptive fitness to the organisms that implemented them.[7] Insofar as one is concerned with successful action, one should want to learn in ways that work, and even the simplest form of reinforcement learning very often does. But even here one can see that there must be more to the story.

Custom is a great guide of human life, but it is not our only guide. In addition to there being a variety of forms of reinforcement learning, we often learn in ways that are not well characterized as reinforcement at all. This last point is pragmatically significant since reinforcement learning is often not the best way to learn. While Hume allowed for the involvement of other psychological faculties in learning, custom in even its simplest form provides a means of learning how to learn in more sophisticated ways. To see how, we will begin by considering the nature of custom in human and animal learning.

3 Learning by reinforcement

The psychologist Edward Thorndike (Thorndike, 1898) performed some of the first careful experiments to investigate the nature of reinforcement learning in animals. His experiments involved putting hungry cats, dogs, and chicks in puzzle boxes from which they might escape by performing a simple action like pulling a cord, pressing a lever, stepping on a platform, or turning a button.[8] Food was placed outside the box in the sight of the animal, and its actions were observed. If the animal escaped and got the food, the length of time it took was recorded. If it did not escape within a reasonable period of time, the animal was removed from the box without being fed. If it never figured out how to escape, the case was recorded as one of "complete failure" and the data for that animal was set aside.

Thorndike found that for those animals that were eventually able to figure out how to escape, as the experiment was repeated, it took less time for them to escape. The time it took eventually became very short and relatively constant (1898, 6–7). He used time curves to present the progress of learning. Figure 1 is an example of one of these from his experiments with cats. Cat no. 10 was a kitten 4–8 months old. To escape from a type C box, it had to turn a button from the vertical to horizontal position. In the time curve, the horizontal axis indicates the trials in temporal order and the vertical axis indicates the length of time it took for each. The marks on the horizontal axis indicate significant breaks in time between trials. As the number of trials increases, the time it takes to solve the problem decreases. Thorndike took the curve to represent the evolution of the animal's probabilistic dispositions in the context of the particular puzzle box.

It was essential to Thorndike's understanding that the process of learning involved the gradual evolution of dispositions. As he put it for his cat experiments, "gradually all the other non-successful impulses will be stamped out and the particular impulse

[7] Hume anticipated this virtue of custom in holding that it is a principle "necessary to the subsistence of our species, and the regulation of our conduct, in every circumstance of human life" (1975, 55).

[8] He reported that all of the animals used in his experiments "were apparently in excellent health, save an occasional chick."

leading to the successful act will be stamped in by the resulting pleasure until, after many trials, the cat will, when put in the box, immediately claw the button or loop in a definite way." (1898, 13). He considered reinforcement learning to be a physical process, one involving the nervous system of the animal:

> The gradual increase in success means a gradual strengthening of one set of nerve-connections, and a gradual weakening of others. This method of learning may be called the method of trial and error, or of trial and success. ... The cause of such strengthening and weakening is the resulting pleasure in one case and discomfort in the others. (Thorndike, 1901, 38–39)

And he considered the ability to learn in this fashion to be the result of natural selection, "The most important of all original abilities is the ability to learn. It, like other capacities, has evolved." (Thorndike, 1911, 278).

There is a great deal of subsequent evidence that, just as with Thorndike's cats, dogs, and chicks, humans and other animals also very often learn by reinforcement. Salient examples include R. J. Herrnstein's (1970) studies on birds and Alvin Roth's and Ido Erev's (1995) (1998) studies on humans.

In its most basic form, reinforcement learning works as follows. Let $q_i(t)$ be an agent's propensity for strategy i at time t. Her propensities evolve according to the update rule:

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

Here $\pi(t)$ is the payoff received by an agent taking the action i on round t. The payoff represents the degree of success or failure resulting from the action. It affects the agent's propensities, and her propensities fix her dispositions by determining the probability of each action on a future round of play. How this works is given by the response rule:

$$p_i(t) = \frac{q_i(t)}{\sum_j q_j(t)}$$

Here $p_i(t)$ is the probability that the agent takes action i on a play at time t.

In order to say how the process gets started, one must also specify a set of initial propensities. Lower initial weights allow for more agile early exploration. Higher initial weights make for more stable dispositions but also slow the process of learning. And uneven initial weights bias the early dispositions of the learner in a way that may lead to very different behaviors depending on the learning problem. While one can assign initial propensities any way one wants, we will suppose that each strategic option is given an equal and small initial weight, say $q_i(0) = 1$ for all i when $\pi(t) = 1$.

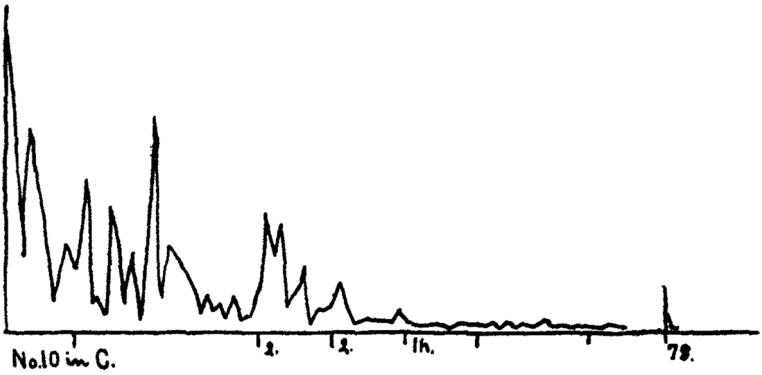

Fig. 1 Cat no. 10 in box type C

Both Hume and Thorndike allowed for reinforcement on success *and* punishment on failure. The formal scheme above also does by allowing for negative payoffs.[9] That said, we are primarily interested here in the simplest form of reinforcement learning. Specifically, we will begin by supposing that payoffs are always positive. Such *simple reinforcement learning* can be thought of as drawing a ball from an urn to determine one's action then adding balls of the type drawn to the urn if and only if the action was successful. We will consider precisely how this works in the next section.

Hume was right to suppose that we and other animals learn inductively by reinforcement, but we often use other similarly simple forms of trial-and-error learning, and we sometimes use significantly more sophisticated and cognitively costly forms of learning. In this we are also fortunate. While simple reinforcement learning is ubiquitous in nature, low cost, and very often effective, there are many tasks where other forms of learning work better.

4 Two problems

Consider two learning problems. In each a subject must determine the location of a food reward to be successful. The food is initially placed in one or the other of two opaque but distinguishable boxes A or B according to a placement rule. The subject chooses between the boxes. If she chooses the correct box, she is rewarded with the food. Whether or not she is successful, the boxes are concealed for a moment and the placement rule is applied again. Then the subject is given another chance to choose.[10]

[9] One just needs to guard against negative propensities. A natural way to do so is to stipulate that if a punishment would cause a propensity to fall below a small value $k > 0$, then the propensity is set to k.

[10] This sort of discrimination problem has a long history in experiments in comparative psychology. The present setup is much like the one studied by Harry Harlow (1949) in his learning-set experiments with monkeys. While our reflections here are clearly relevant to such experiments, there is good reason to suppose that Harlow's monkeys learned how to learn in a more subtle way than the sort of Humean

The two problems differ only in the placement rule. On each run of a *fixed-box problem*, the experimenter initially chooses one or the other of the two boxes at random, either A or B, as special for that run. He then places the food in the special box to begin the run. He does nothing when the boxes are concealed between plays unless the subject just chose the special box and ate the food. If so, the experimenter places food in the same special box again when the boxes are concealed between plays.

On each run of a *random-box problem*, the experimenter initially chooses one of the two boxes at random as special for that run, then places the food in that box with probability 2/3 and in the other box with probability 1/3 to begin. Then, regardless of whether the subject was successful on the last play, when the boxes are concealed, the experimenter randomly determines the location of the food for the next play using the same probabilities for the special box ($p = 2/3$) and the other box ($p = 1/3$). This process continues until the end of the run.

Now consider two learning rules an agent might use to decide which box to pick on a play. The first is *simple reinforcement*. On this dynamics the agent begins the run with an urn containing one ball labeled A and one ball labeled B. When given the chance to choose a box, she draws a ball at random from her urn then chooses the corresponding box. If she finds the food in the box on that play, she returns the ball she drew to the urn and adds a duplicate ball of the same type; otherwise, she simply returns the ball that she drew.

The second learning rule is *win-stay/lose-shift*, another simple trial-and-error form of learning. Here the agent randomly and unbiasedly chooses a box to open on the first play of a run. If she finds the food, and hence wins, she stays with the same box when given the chance to chose on the next play; otherwise she chooses the other box on the next play. On this dynamics the agent stays with a box as long as she finds food there and shifts to the other box if the present box is ever found empty. Then she stays with the new box only as long she finds food there.

The best learning rule for an agent to use depends on the type of learning problem she faces. Consider a sequence of ten-play runs of the fixed-box problem. Simple reinforcement does fairly well with a mean success rate on simulation of 0.74 of finding the food on a play.[11] But win-stay/lose-shift does better. Since the location of the food does not change during the run on a fixed-box problem, if a win-stay/lose-shift agent is right on her first guess, then she will be right on every play of the run. And if she is wrong on her first guess, she will be right on her second guess and from then on. Since the probability of each initial guess is 1/2 on a run, she will end up with a mean success rate of 0.95 over all the plays in the sequence of runs. As a result, a simple reinforcement learner should prefer to learn by win-stay/lose-shift rather than by simple reinforcement for this sort of problem.

Footnote 10 (continued)
bootstrapping that we are considering at present. See Barrett (2023) for an account of how bootstrapping might work in Harlow's experiments.

[11] This is for 1000 ten-play runs of the fixed-box problem.

While win-stay/lose-shift does better than simple reinforcement on short runs of the fixed-box problem, simple-reinforcement does better than win-stay/lose-shift on long runs of the random-box problem. Consider a sequence of one-thousand-play runs of the random-box problem. On simulation, simple reinforcement does nearly as well as possible given the rule for placing the food with a mean cumulative success rate of 0.63. Win-stay/lose-shift in contrast exhibits a mean cumulative success rate of 0.55.[12] This is better than chance, but only just. A simple reinforcement learner, then, should sometimes prefer to learn by simple reinforcement rather than by win-stay/lose-shift. So how might a simple reinforcement learner come to use win-stay/lose-shift for fixed-box problems and simple reinforcement for random-box problems?

5 Bandit games

While simple reinforcement learning is not always optimal for a given task, it very often provides a way for an agent to learn how to learn more effectively when presented with the task. Inasmuch as she tends to act in ways that have been successful in the past, a reinforcement learner will tend to reuse learning rules that have worked well. If a learning rule has in fact worked well in a salient way for a particular task, she may consequently learn to use that rule for that type of task when she desires the sort of success it affords. In this way, she may assemble a learning system where she continues to use simple reinforcement learning for some purposes but adopts other, possibly more sophisticated, forms of learning for others.[13]

To see how this works, we will begin by considering how a simple reinforcement learner might evolve optimal dispositions for playing an *n*-armed bandit game, then return to the problem of learning how to learn in the next section.[14] In an *n*-armed bandit game, an agent is presented with *n* slot machines with the goal of finding and playing the machine that in fact pays at the highest rate.

Consider three slot machines *A*, *B*, and *C* that each pays in dollar coins and has a maximum payoff of $10 on a play. Suppose that each machine pays randomly with a different, but unknown, expected return. There is a straightforward procedure by which a simple reinforcement learner will almost always evolve to play the machine with the highest expected return with probability 1.

Suppose that the agent starts with an urn containing one ball of each type *A*, *B*, and *C*. On each play, she draws a ball at random from the urn and plays the indicated

[12] Both of these results are for 1000 runs of the random-box problem with 1000 plays per run. Win-stay/lose-shift does better than chance since winning provides at least some evidence that one has found the higher-chance box.

[13] See Barrett and Skyrms (2017), Barrett (2020), and Barrett (2023) for more general discussions of such self-assembly.

[14] Bandit problems provide a natural framework for modeling simple inquiry. Mayo-Wilson et al. (2011) and Mayo-Wilson et al. (2013) use an approach similar to the one we consider below to model scientific inquiry within a community. See Berry and Fristedt (1985) for a survey of bandit problems and Huttegger (2017) for a discussion of rational learning in bandit problems.

machine. Then she returns the ball drawn to the urn and adds a duplicate ball for each dollar coin she received on the play. As she repeats this procedure, with probability 1 both the probability that she will play and the empirical frequency with which she will play the machine with the highest expected return will converge to 1 in the limit of play. As a result, she will almost always evolve dispositions to play optimally. This is true for any finite number of slot machines she might investigate if one of them in fact pays best.[15]

Not all learning dynamics have this property. Some are, as Brian Skyrms put it, *too hot* and some are *too cold*.[16] If the learning dynamics is too cold, then the agent may get stuck always playing the same suboptimal machine. If it is too hot, she may get stuck always exploring her options. Simple reinforcement solves the Goldilocks problem of finding an effective learning dynamics by being just right for the standard n-armed bandit game.

In contrast, consider a probe-and-adjust learner, another simple trial-and-error form of learning. Here the agent chooses an initial machine to play at random. On each round, she has a constant probability p of probing instead of playing her current machine. If she probes, she chooses another machine with unbiased probabilities and plays it. If the payoff is higher on that play than on her last play of the machine she was playing before the probe, she shifts to playing the new machine until her next probe. If the payoff on her play of the new machine is lower, she goes back to the machine she was playing before the probe. And if the payoffs are equal, she flips a coin to decide which machine to play. Skyrms (2015) shows that while the learner will spend more of her time playing higher paying machines, she will never settle on the highest paying machine and hence never learn to play optimally. This dynamics is too hot.[17]

In a standard n-armed bandit game, one supposes that each machine has a constant expected payoff and that the outcome of each play is independent. Simple reinforcement will find the optimal strategy in this case, but one can get something significantly stronger.

Alan Beggs (2005) showed that if there exists a constant $\gamma > 1$ such that the expected return of one action is in fact always greater than γ times the expected return of each other action at each step in the learning process, then with probability 1 both the probability of and the empirical frequency with which a suboptimal action will be played by a simple reinforcement learner goes to zero in the limit of

[15] See Skyrms (2015) for a discussion of this point in light of Hopkins and Posch (2005). Beggs (2005) provides a more general result that we will discuss shortly.

[16] See Skyrms (2010, 87–8) for a discussion and Barrett (2023) for a survey of the properties of various learning rules.

[17] A more sophisticated probe-and-adjust learner might track the statistical features of the machine she is playing, then when she probes, play the new machine *for a while* before deciding whether to shift. She only shifts if the new machine exhibits better statistics over the duration of the probe. Such a learner might do significantly better than a probe-and-adjust learner who only remembers the results of her last play on each machine. Even so, she will not converge to optimal play as there will always be a constant positive probability of shifting away from playing the machine with the highest expected return. As a quick example of a learning dynamics that is too cold, consider the simple strategy of always playing the same machine come what may.

play (6–7). The upshot is that if one machine dominates the others in this sense, an agent who learns by simple reinforcement will eventually evolve optimal dispositions, and this holds even if the expected payoffs of the machines change over time or if they depend on the history of play or even the behavior of the other machines.

This gives us something concrete to say concerning the reliability of simple reinforcement learning in contexts of practical choice. If one learns by simple reinforcement, then if one action always γ-dominates the other available actions in the sense just described, an agent who considers that action will almost always learn to act optimally. Depending on the situation, it may take a long time even to get close, but under these conditions, one is guaranteed with probability 1 to evolve optimal dispositions in the limit.[18] Hume was, in this sense, right to be optimistic regarding the reliability of custom as a guide to human life.

One consequence of all this is that, while there is no canonically best learning rule for all occasions, simple reinforcement is a form of low-rationality learning that is often particularly well-suited to learning how to learn.

6 Learning how to learn

Sometimes learning how to learn has the structure of an n-armed bandit game. Consider an agent who has the task of determining which of three learning rules A, B, or C works best for a particular type of learning problem. Suppose that she has a criterion such that a learning rule either succeeds or fails each time it is applied to a problem and that she starts with an urn containing one ball of each type A, B, and C. On each play, she draws a ball at random, then tries the learning rule indicated by that ball on the learning problem at hand. If the rule succeeds on her criterion, then she returns the ball drawn to the urn and adds a duplicate. Otherwise, she simply returns the ball drawn to the urn.

If the outcome of each trial of the learning rule is independent and if each rule has a constant reliability (probability of success given the learner's criterion) for the type of learning problem one is considering, then with probability 1, the probability that the learner will use that rule and the empirical frequency that she will use it will both converge to 1 as she continues learn by simple reinforcement on successful plays. This holds for any finite set of learning rules she might consider.[19] Depending on the task at hand, a simple reinforcement learner might even learn that simple reinforcement is best among the competitors for accomplishing it.

[18] Simple reinforcement learning is sometimes very slow as early chance reinforcements may lead away from the optimal strategy, and it can take a long time to recover. Further, depending on the task, expected reinforcements for suboptimal play may be nearly as high as those for optimal play, and it is hard to get traction on playing the best option when the others are nearly as good. See Beggs for a further discussion of these points (2005, 7). Allowing for both reinforcement and punishment, as Hume did in his conception of learning by custom (1975, 105–6), often yields much faster learning. But as Beggs' results are for simple reinforcement learning, we will stick with that for now.

[19] See again (Skyrms, 2015) and Hopkins and Posch (2005) for discussions.

But as we discussed in the last section, the efficacy of reinforcement learning does not depend on the independence of plays. This allows us to say somewhat more here as well.

Consider an oracle-selection game. Suppose that, in preparation for war with the Persians, Croesus wishes to determine which of seven oracles is most reliable. To this end, he places one ball representing each of the oracles in the royal urn. On each play, he draws a ball from the urn then sends an emissary to the corresponding oracle to ask a question of the oracle. If the oracle's answer proves correct, he reinforces by adding a duplicate of the ball drawn to the urn.

Some questions may be harder than others. Or the oracles may get better at making predictions with experience. Or the answer of one oracle may depend on the answer of another. But if one of the oracles is always in fact more reliable than a constant factor $\gamma > 1$ times the reliability of each other oracle in answering the questions asked, then, by Beggs' theorem, Croesus will almost always learn to consult that oracle. If so, he has learned which oracle is in fact best to learn from by reinforcing on the successes of each.

The dominance condition does not hold if the Pythia at Delphi is the most reliable in answering one type of question and the Sybil at Cumae is most reliable at answering another type of question and the king can decide which type of question to ask. Nor does it hold if the Pythia is always the most reliable oracle on weekdays and always the least reliable on weekends and the king can decide when to ask a question. It only holds if one oracle is in fact the most reliable in answering each question given that the king asks it. The most reliable oracle may answer a question incorrectly, but it must always have the lowest probability of doing so.

Just as the king will learn which oracle to consult if one dominates the others, a simple reinforcement learner will evolve to use a dominant learning rule for the task at hand if there is one and if it happens to be among those that she tries. An agent who begins as a simple reinforcement learner may, by such means, self-assemble a learning system where different rules gradually come to be used for different learning tasks. She may come to use a form of win-stay/lose-shift to decide where to get her morning coffee and a form of reinforcement with iterated punishment when learning complex signaling conventions.[20] She may even come use a form of Bayesian conditioning for learning problems where the stakes are high if she has the requisite cognitive capacity and the sort of background information she needs to implement such a dynamics. We will return to this in a moment.

In the Croesus story, an oracle either predicts correctly or not on each play, but a simple reinforcement agent may also learn which learning rule is best with respect to virtues that come in degrees. This works in the same way that an agent might learn which machine pays best in an n-armed bandit game.

Consider the food-location task that we started with and a simple reinforcement learner who wants to learn whether simple reinforcement or win-stay/lose-shift works better on a sequence of ten-play runs of the fixed-box problem. Suppose that

[20] See Barrett and Gabriel (2022) for a discussion of the latter type of learning and its efficacy in Lewis-Skyrms signaling games.

she cares, in particular, about the mean cumulative success rate on each run of the problem.

Suppose that the agent starts with equal propensities for using simple reinforcement and using win-stay/lose-shift on each ten-play run. Specifically, let $q_0(t_0) = q_1(t_0) = 1$, where $q_0(t)$ and $q_1(t)$ are the propensities of using simple reinforcement and win-stay/lose-shift respectively. At the beginning of each new ten-play fixed-box problem she chooses a learning rule to use for that problem. The probability that she will use rule i is given by

$$p_i(t) = \frac{q_i(t)}{\sum_j q_j(t)}.$$

After running the problem on that rule, the agent reinforces her propensity to use the rule by the cumulative success rate that it provided on the ten-play run. This represents the degree of success she achieved given what she cares about.

On the first round, the agent will use each of the two learning rules for the ten-play fixed-box problem at random and with equal probabilities. But as she plays, since she achieves a mean success rate of 0.95 when she uses win-stay/lose shift and a mean success rate of 0.74 when she uses simple reinforcement, she will reinforce somewhat more on average when she uses win-stay/lose-shift than when she uses simple reinforcement on a ten-play fixed-box problem. The cumulative effect of this difference will be to make it more likely that she will use win-stay/lose-shift on future problems.

The process is not fast, but it is sure. On simulation, when presented with a series of 1000 runs each consisting of 10^4 ten-play fixed-box problems, a simple reinforcement learner evolves to use win-stay/lose-shift better than 0.90 of the time cumulatively on approximately 0.36 of the runs with a mean probability of playing using the optimal rule at the end of a run of about 0.81. She continues to learn how to learn better over time. When presented with 10^6 ten-play fixed-box problems she uses the optimal rule better than 0.90 of the time cumulatively on approximately 0.67 of the runs with a mean probability of playing optimally at the end of a run of about 0.91.[21] And it follows from the results above that with probability 1 in the limit the simple reinforcement learner will learn to use the most effective learning dynamics for the problem at hand with probability 1. If she investigates her options for how she might learn best in the context of this type of learning problem in the way we have described, she is in this sense fated to learn to use the best learning dynamics for the problem at hand.[22]

An agent will similarly learn to use simple reinforcement instead of win-stay/lose-shift learning in the context of random-box problems. And she may do so

[21] These simulations were originally run in c++. See the supplementary material for a python version of the code.

[22] Concerning the speed of convergence, a learning rule need not be optimal to be useful. It may not matter much for the sake of practical action if the rule one is using is suboptimal if it is difficult for a simple reinforcement learner to distinguish it from an optimal rule by the associated track records of short- to medium-run success.

without reflective justification or even knowing the means by which she acquired her context-dependent dispositions regarding how to learn. She may even learn to act by a form of Bayesian conditioning if her situation calls for it. The Monty Hall game provides a simple example.

The Monty Hall game is played between a host and a contestant. The host randomly and without bias places a prize in one of three opaque boxes. The contestant chooses a box at random, then the host opens a box that he knows does not contain the prize. If the host can open either remaining box without revealing the prize, he opens one at random. The contestant is then given the option to switch her choice to the unopened box that she did not initially choose. It follows from the rules of the game, the axioms of probability theory, and the principle of strict conditionalization, that the contestant should expect switching boxes to be successful with probability 2/3 and staying with her original choice to be successful with probability 1/3. The upshot is that a rational agent should switch.

Given the significant difference in the expected payoffs, a simple reinforcement learner will quickly learn to switch rather than stay, just as a sophisticated Bayesian agent does who knows the rules of the game, by reinforcing on the payoffs associated with each type of action. Similarly, a more sophisticated reinforcement learner, one with the cognitive ability to represent the rules of the game and to choose an action by an application of Bayesian conditionalization given her evidence, will learn to do so in the context of this problem if this is among the options she considers and if she has access to sufficient evidence of this method's efficacy in this type of problem. Such are the virtues of simple reinforcement in learning how to learn.

The sort of reinforcement with punishment that Hume used to characterize custom in his analogy with animal learning can be significantly more effective than the special case of simple reinforcement learning we have been considering. While it is difficult to prove analogous results for reinforcement with punishment, a modest degree of negative reinforcement on failure often allows for much faster, and in many cases, more reliable learning. It also provides a way for agents to retool by unlearning dispositions that no longer lead to successful action given changes in the world they inhabit. In this sense, one might often expect to do yet better using reinforcement with punishment in learning how to learn.[23]

[23] See Barrett (2006) and Barrett (2007), Barrett and Zollman (2009), and Barrett and Gabriel (2022) for examples of the relative advantages of various forms of reinforcement with punishment learning in signaling games. Signaling games pose a particularly difficult type of learning problem since each agent is chasing an always moving target. Beggs' (2005) results for the efficacy of simple reinforcement learning do not apply to this type of problem, and it, indeed, typically fails to yield optimal dispositions. That said, reinforcement with punishment is not always preferable to simple reinforcement even here. If the level of punishment on failure is too high given the level of reinforcement on success and the difficulty of learning problem, reinforcement with punishment may fail to yield any useful dispositions at all.

7 Discussion

None of this means that a simple reinforcement learner will always learn how to learn optimally. It may be that none of the learning rules she considers work well for the type of problem at hand. Or it may be that there is no one learning rule that is always the most reliable given the problem. Or she may fail to conduct her trials or to reinforce with sufficient care. Or it may be that reinforcement learning is simply too slow, given her patience or lifespan, to yield reliable learning dispositions. Such reflections are consonant with the observation that we often fail to learn optimally.

That said, Hume recognized, even in himself, the psychological efficacy of custom. While he held that there can be no ultimate justification for the inductive practice afforded by custom, inasmuch as we are naturally constituted to pursue empirical inquiry in this way and to be satisfied by its results, he also held that the lack of such justification matters to neither inquiry nor our satisfaction in its results. Just as we tend to expect similar effects from similar causes, if a form of learning has proven successful for a particular task, we will tend to use it for that task. And just as constant conjunction between events often yields confidence in there being a connection between those events even when we have no ultimate justification for so believing, one should expect the regular success of a form of learning to be accompanied by a conviction that it will continue to serve as a reliable guide in future action. The learning rule may not in fact continue to facilitate successful action, but if it does, its continued success will tend to reinforce further both our use of the rule and our confidence in its use.

Hume's trade of rational justification for learning has another pragmatic virtue. As we have seen, even the simplest form of reinforcement learning provides a way of investigating the relative success of alternative ways of learning. A simple reinforcement learner will almost always learn the best way of learning in the context of a particular task if there is a best way (a way that γ-dominates the others), if she is fortunate enough to consider it, and if she is able to pursue the matter with sufficient care and diligence. And if there is variation in how we apply a learning rule over time, reinforcing on especially successful instances of application may even serve to tune the rule itself, better fitting our practice to successful inquiry.

Hume was prescient in his appeal to custom as the grounds for his skeptical solution to the problem of induction. There is good empirical reason to believe that reinforcement on experience often guides human action. That we have evolved to learn in this way is unsurprising given that reinforcement requires few resources to implement and is very often effective given the world we inhabit. And even the simplest form of reinforcement provides a path to the adoption of more reliable and potentially more sophisticated forms of learning.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11098-023-02086-3.

Acknowledgements I would like to thank Kyle Stanford, Brian Skyrms, Christian Torsell, Jack VanDrunen, Thomas Barrett, and two anonymous reviewers for comments on an earlier version of this paper. I would also like to thank Christian Torsell for many convesations on this topic and for the Python version of the simulation for inclusion as a supplement to this paper.

References

Allison, H. E. (2008). *Custom and Reason in Hume: A Kantian Reading of the First Book of the Treatise*. Oxford: Oxford University Press.

Barrett, J.A. (2023). *Self-Assembling Games*. Manuscript.

Barrett, J. A. (2020). Self-assembling games and the evolution of salience. *The British Journal for the Philosophy of Science, 74*(1), 75–89.

Barrett, J. A., & Gabriel, N. (2022). Reinforcement with iterative punishment. *Journal of Experimental & Theoretical Artificial Intelligence.* Published online: 13 Dec 2022. https://doi.org/10.1080/0952813X.2022.2153272

Barrett, J. A., & Skyrms, Brian. (2017). Self-assembling games. *The British Journal for the Philosophy of Science, 68*(2), 329–353.

Barrett, J. A., & Zollman, K. (2009). The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence, 21*(4), 293–309.

Barrett, J. A. (2007). Dynamic partitioning and the conventionality of kinds. *Philosophy of Science, 74*, 527–46.

Barrett, J. A. (2006). Numerical simulations of the Lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar. *Institute for Mathematical Behavioral Sciences*. Paper 54. https://www.imbs.uci.edu/files/docs/technical/2006/mbs06_09.pdf

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory, 122*, 1–36.

Bereby-Meyer, Yoella, & Erev, Ido. (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.

Berry, Donald A., & Fristedt, Bert. (1985). *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman & Hall.

Cochran, C. T., & Barrett, J. A. (2023). The efficacy of human learning in Lewis-Skyrms signaling games. *The British Journal for the Philosophy of Science.* https://doi.org/10.1086/724446

Cochran, C. T., & Barrett, J. A. (2021). How signaling conventions are established. *Synthese, 199*(1–2), 4367–4391.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88*, 848–81.

Fudenberg, D., & Levine, D. (1998). *Learning and the Theory of Games*. Cambridge, MA: MIT Press.

Henderson, L. (2022) The problem of induction. *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta and Uri Nodelman (eds.). https://plato.stanford.edu/archives/win2022/entries/induction-problem/

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review, 56*, 51–65.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior, 13*, 243–266.

Hopkins, E., & Posch, M. (2005). Attainability of boundary points under reinforcement learning. *Game and Economic Behavior, 53*, 110–125.

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

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

Mayo-Wilson, Conor, Zollman, Kevin, & Danks, David. (2013). Wisdom of crowds versus groupthink: Learning in groups and in isolation. *International Journal of Game Theory, 42*, 695–723.

Mayo-Wilson, Conor, Zollman, Kevin J. S., & Danks, David. (2011). The independence thesis: When individual and social epistemology diverge. *Philosophy of Science, 78*, 653–677.

Morris, William Edward, & Brown, Charlotte R. (2019). David Hume. *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2022/entries/hume/

Putnam, Hilary (1963). 'Degree of confirmation' and inductive logic. In Paul Arthur Schilpp (ed.), *The Philosophy of Rudolf Carnap* (pp. 761–783). Open Court: La Salle.

Roth, A. E., & Erev, I. (1995). Learning in extensive form games: Experimental data and simple dynamical models in the intermediate term. *Games and Economic Behavior, 8*, 164–212.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science, 75*(3), 278–305.

Schurz, G. (2019). *Hume's Problem Solved: the Optimality of Meta-induction*. Cambridge, MA: MIT Press.

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

Skyrms, B. (2015) Learning to signal with two kinds of trial and error. In *Foundations and Methods from Mathematics to Neuroscience: Essays Inspired by Patrick Suppes*. Colleen E. Crangle, Adolfo Garcia de la Sienra, and Helen E. Longino (eds). CSLI Publications, 2015.

Thorndike, E. L. (1898) Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, Vol. II., No. 4 (Whole No. 8), June, 1898. The Macmillan Company: New York and London.

Thorndike, E. L. (1901). *The Human Nature Club: An Introduction to the Study of Mental Life* (2nd ed.). New York: Macmillan.

Thorndike, E. L. (1911). *Animal Intelligence*. New York: Macmillan.