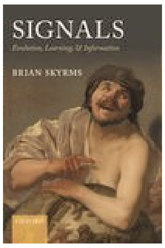

Signals: Evolution, Learning, and Information

Brian Skyrms

https://doi.org/10.1093/acprof:oso/9780199580828.001.0001

Published: 08 April 2010 **Online ISBN:** 9780191722769 **Print ISBN:** 9780199580828

CHAPTER

7 7 Learning

Brian Skyrms

https://doi.org/10.1093/acprof:oso/9780199580828.003.0008 Pages 83–92

Published: April 2010

Abstract

This chapter argues that investigation of reinforcement learning is a complement to the study of belief learning, rather than being a ‘dangerous antagonist’. It begins at the low end of the scale, to see how far simple reinforcement learning can get us, and then move up. Exactly how does degree of reinforcement affect the strengthening of the bond between stimulus and response? Different answers are possible, and these yield alternative theories of the law of effect.

Keywords: law of effect, belief learning, reinforcement learning

Subject: Philosophy of Science, Epistemology, Philosophy of Language

Collection: Oxford Scholarship Online

> “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.”
>
> Edward Thorndike, *Animal Intelligence*, 1911

The Law of Effect

When Edward Thorndike, as an undergraduate English major at Wesleyan University, read William James' *Principles of Psychology*, he switched his interests. After graduating in 1895 he moved to Harvard, and eventually conducted learning experiments on chickens in the basement of William James' house. The epigraph to this chapter is a statement of Thorndike's famous “Law of Effect,” taken from his 1911 *Animal Intelligence*.

Despite Thorndike's high regard for James, his studies of what is now known as reinforcement learning were a move away from introspective cognitive psychology towards a theory more focused on behavior. In the spirit of

p. 84 Darwin, he focused on commonalities between human and animals. That this focus sometimes generated a ↳ certain amount of hostility is evident from his introductory paragraph of an article on the law of effect:[1]

> It (*the Law of Effect*) has been even more odious to philosophers and educational theorists, who find it a dangerous antagonist to, or an inferior substitute for, their explanations of behavior by purposes…

We think that investigation of reinforcement learning is a complement to the study of belief learning, rather than being a "dangerous antagonist." Our strategy will be to begin at the low end of the scale, to see how far simple reinforcement learning can get us, and then move up. Exactly how does degree of reinforcement affect the strengthening of the bond between stimulus and response? Different answers are possible, and these give us alternative theories of the law of effect.

Roth–Erev reinforcement

In 1995 Al Roth and Ido Erev used a version of reinforcement learning to account for how subjects actually behave in experiments.[2] The experiments have the subjects repeatedly playing a game, and sophisticated rational choice fails to explain the experimental data. Roth and Erev, following pioneering early investigations by Patrick Suppes and Richard Atkinson show that reinforcement gives a much better explanation.[3]

Roth and Erev's basic model works like this. The probability of choosing an action is proportional to the total accumulated rewards from choosing it in the past. They trace the idea back to the psychologist Richard Herrnstein.[4] Some initial equal inclinations, p. 85 ↳ or propensities, are assumed to get the process started by choosing at random.[5]

We can visualize the operation of the law of effect in terms of drawing balls from an urn. For instance, suppose you have two actions and start out with an urn containing one red and one black ball. On the first trial, you draw a ball and choose act 1 if it is red and act 2 if it is black. Suppose you choose act 1 and get a reward of two. Then you put two more red balls in the urn and draw again. Now the chance of drawing a black ball is 1/4. But suppose you draw the one black ball and get a reward of six. Then you put in six black balls, and draw again. In this way the urn keeps track of accumulated rewards. We don't really need an urn. Organisms may keep track of accumulated rewards by strength of neural connections,[6] or concentrations of pheromones,[7] or any number of ways.

We can summarize the basic Roth–Erev reinforcement process as follows: (i) there are some initial inclination weights; (ii) weights evolve by addition of rewards gotten; (iii) probability of choosing an alternative is proportional to the inclination weights.

When the magnitudes of the rewards are fixed, there is only one parameter of the process. That is the magnitude of the initial equal weights. If they are very large, learning starts off very slowly. If they are small, initially probabilities can move a lot. But either way, as reinforcements pile up, individual trials can move probabilities less and less. Learning slows down. In psychology, the qualitative phenomenon of learning slowing down in this way is called the *Law of Practice*.

p. 86 Bush–Mosteller Reinforcement

In 1950 Bush and Mosteller suggested a different realization of the law of effect. Today's rewards act directly on probabilities of acts—there is no memory of accumulated reinforcements. The most basic model looks like this: If an act is chosen and a reward is gotten the probability is incremented by adding some fraction of the distance between the original probability and probability one.[8] Alternative action probabilities are decremented so that everything adds to one. The fraction used is the product of the reward and some learning parameter. Rewards are scaled to lie in the interval from zero to one, and the learning parameter is some fraction.

For example, suppose that there are just two actions and the current probability of act one is .6. Suppose that you happen to choose act 1 and get a reward of 1. Take the learning parameter to be .1, then your new probability of act one is .6 + .1 (1-.6) =.64. Your new probability of act 2 is .36. At this point the learning parameter is the only parameter. If it is small you learn slowly; if it is larger, you learn fast. But learning does not slow down as it does in Roth–Erev reinforcement. Basic Bush–Mosteller does not obey the Law of Practice.

Bush–Mosteller learning has also been used to explain empirical data.[9] Both Roth–Erev reinforcement and Bush–Mosteller reinforcement have led to versions with various modifications and lots of parameters, but for now we stick with the simplest versions of each. We would like to compare them. In the long run they can behave quite differently.

p. 87 Slot machines and medical trials

Consider two slot machines that pay off at different unknown rates—or alternatively, different drugs with different unknown probabilities of successful treatment. A trial-and-error learner has to balance two different considerations. She doesn't want to lock onto the wrong machine just because it got lucky in a few initial trials. But nor does she want to explore forever, and never learn to play the optimal machine. Likewise, in medical research we don't want to *jump to conclusions*. We want to explore long enough to have a valid study. But if one treatment is clearly better, we don't want to deny it to those who need it by dithering about. There is a tension between gaining knowledge and using it. As John Holland put it in 1975,[10] there is a tradeoff between exploration and exploitation.

If our gambler sometimes freezes into always playing the wrong machine, we will say her version of reinforcement is *too cold*. She learns too fast. If she gets stuck in exploring, and never learns to play one machine, even in the limit, we will say that her version of reinforcement is *too hot*. She never fully learns. Goldilocks reinforcement learning would be neither too hot nor too cold.[11] It would always converge to playing the optimal machine with probability one. In the drug trials model, it would always learn to use the best treatment.

Is there Goldilocks reinforcement learning? In 2005, Alan Beggs proved that Roth–Erev reinforcement has the
p. 88 Goldilocks ↳ property.[12] The theorem was already proved in this context in 1978, by Wei and Durham[13]—using
different m:mathematical techniques.

Bush–Mosteller learning, as presented above, is too cold. It can freeze into playing the worse slot machine. This is because Bush–Mosteller reinforcement does not slow down with more practice. It learns too fast.[14]

Reinforcement and evolution

Reinforcement learning is probabilistic; at any juncture alternative acts may be selected and alternative paths taken. But inside the probabilistic process lies a deterministic dynamics describing the expected motion at every point in the process.

This is called the average, or mean-field dynamics.[15] If learning is very slow—if it proceeds in tiny steps—then with high probability the real learning path will, for some time, approximate the mean-field dynamics. So it is of some interest to ask what this mean-field dynamics is for our two basic models of reinforcement learning.

In 1997 Tilman Börgers and Rajiv Sarin showed that the mean-field dynamics for Bush–Mosteller learning is a version of the replicator dynamics. In 2005, Alan Beggs and also Ed Hopkins and Martin Posch showed that the mean-field dynamics of Roth–Erev learning is a version of the replicator dynamics. In Chapter 1, we started p. 89 with one question and ended up with two: *How can ↳ interacting individuals spontaneously learn to signal? How can species spontaneously evolve signaling systems?* Now we see that these two questions are closely intertwined.[16]

Why then, are the dynamics so different in the long run? The key is that the Roth–Erev dynamics slows down in such a way that replicator dynamics is a good indication of limiting behavior.[17] Bush–Mosteller does not slow down, so while it may be likely to stay close to replicator dynamics for a finite stretch of time, it may not be close at all in the long run. There is a theory of slowing down in such a way that the mean-field dynamics is a good guide to limiting behavior, the theory of stochastic approximation.[18] This is the theory that Beggs used to prove that Roth–Erev learning has the Goldilocks property in medical trials and the slot machine problem. As we shall see in the next chapter, it is also the tool for analyzing this kind of reinforcement in signaling games.[19]

Variations on reinforcement

Both of the realizations of the law of effect that we have discussed have given rise to various modified versions. Negative payoffs have been considered, with the zero point either fixed or itself evolving. Errors have been introduced. A little bit of forgetting the past has been introduced into the Roth–Erev model (Bush–Mosteller already forgets the past.) Different ways of translating inclination weights into choice probabilities have been tried with Roth–Erev.

One popular approach is to use an *exponential response rule*. The basic idea is to make probabilities proportional p. 90 to the exponential of ↳ past reinforcements.[20] Or more generally, past reinforcements are multiplied by some constant, lambda, and probabilities are proportional to the exponential. In analogy with thermodynamics, the reciprocal of lambda is sometimes called the temperature. If lambda is zero the temperature is infinite, and everything is tried with equal probability. If lambda is large, the act with the largest accumulated rewards is chosen with high probability. Starting with a high temperature and slowly cooling off is called *simulated annealing*, which has been shown to have nice properties for exploring a fixed environment. The effect of rewards piling up in Roth–Erev reinforcement, modified with the exponential response rule, is to slowly cool off the system.

Belief and decision

Reinforcement learners do not have to know their opponent's payoffs; they do not have to know the structure of the game. If acts are reinforced, they do not have to know that they *are* in a game. But now we will move up a level. Individuals know that they are in a game. They know the structure of the game. They know how the combination of others' actions and their own affect their payoffs. They can observe actions of all the players in repeated plays of the game. They can think about all this.

New possibilities for learning now open up. Individuals form beliefs from past experience about how others are likely to act. They then use these beliefs and their knowledge of the game to decide what to do. Different varieties of belief learning dynamics arise from different accounts of how beliefs are formed and different ways of reaching decisions.

p. 91 The very simplest way to form beliefs is to just assume that others will do the same thing that they did last time. The most straightforward way to choose is to pick the best for you, given your beliefs. The combination is called the *best response dynamics*.

It was first studied in the nineteenth century by the m:mathematician, philosopher and economist Antoine Augustin Cournot.[21] For this reason it is sometimes called *Cournot dynamics*.

Inductive logic

The foregoing uses an almost laughably simple method of belief formation. With a history of past play, it would be possible to form beliefs by Bayesian inductive logic. The simplest Bayesian model treats the others as flips of a coin, or rolls of a die, with unknown bias. This gives us *Laplace's rule of succession*. If you choose the best response to these beliefs, the resulting dynamics is, for the following odd historical reason, known as *fictitious play*. In the very early days of game theory, it was thought that a good way to find equilibria in games would be to program one of the (then new) computers to simulate the play of actual players—thus *fictitious play*. The model of the learning dynamics of the players that was suggested at the time by G. B. Brown in 1951 is essentially that just described.[22]

There are a variety of learning models that interpolate between pure reinforcement learning and fictitious play,[23] and experimental studies that fit them to experimental data.[24]

p. 92 Learning to signal

In Chapter 1, we articulated the strategy of starting with reinforcement learning and moving up to belief learning only if reinforcement learning fails. There was a dual rationale for this approach. First, a positive result for reinforcement learning would apply not just to humans, but also to many sorts of animals. Second, reinforcement learning was supposed to be a worst-case scenario. If it allowed us to learn to signal, surely more sophisticated forms of learning would do so too. Is that right? It is time to take a look.

Notes

[1] Thorndike 1927.

[2] Roth and Erev 1995; Erev and Roth 1998.

[3] Suppes and Atkinson 1960.

4 Herrnstein 1961, 1970.

5 Later on, we will also consider variations of the process where the initial propensities are unequal.

6 For a summary of what is known of the neurology, see Schultz 2004.

7 The pheromones in food trails of ants act as a transient record of food obtained that is essentially a reinforcement memory stored outside the individual ants. Evaporation of the pheromone strongly discounts the past so that if it is not continually reinforced the trail vanishes. See Hölldobler and Wilson 1990.

8 The dynamics may be more familiar in the form of updating with a weighted average of the old probability and some maximum attainable probability, which I here take to be 1. Thus, if A is tried and the product of the reward gotten and the learning parameter is α, then $pr_{new}(A) = (1\text{-}\ \alpha)\ pr_{old}(A) + \alpha\ (1)$. This is equivalent to $pr_{new}(A) = pr_{old}(A) + \alpha\ (1—pr_{old}(A)$, which is the way I said it in the text.

9 Macy 1991; Flache and Macy 2002; Macy and Flache 2002; Borgers and Sarin 2000.

10 Holland 1975.

11 The reference is to the tale of Goldilocks and the Three Bears:

> At the table in the kitchen, there were three bowls of porridge. Goldilocks was hungry. She tasted the porridge from the first bowl. “This porridge is too hot!” she exclaimed. So, she tasted the porridge from the second bowl. “This porridge is too cold,” she said. So, she tasted the last bowl of porridge. “Ahhh, this porridge is just right,” she said happily and she ate it all up.

12 Beggs used stochastic approximation theory, which will enter again later in this chapter. In stochastic approximation theory, the Goldilocks property has a precise characterization (connected with decreasing step size of order 1/n). Beggs 2005; Pemantle 2007.

13 Wei and Durham 1978.

14 Some fancier versions of Bush–Mosteller with dynamically adjusting aspiration levels can be too hot. They may exhibit some degree of “probability matching” and never converge to one machine. See Borgers and Sarin 2000.

15 For any state of the learner, various things can happen with various probabilities leading by stochastic dynamics to a new state. Some state is the probability weighted average of the possible new states, or the expected new state. The deterministic dynamics that maps any state onto the expected new state of the stochastic dynamics is called the associated mean-field dynamics.

16 Indeed, some models of evolution in a finite population are remarkably similar to the Roth–Erev model of reinforcement learning. Payoffs are in offspring, and offspring are just individuals of the same type. The difference is that individuals die at random, so with some bad luck types (or even the whole population) may go extinct. But if this doesn't happen and the population grows, then the probabilistic process approximates the replicator dynamics just as reinforcement learning does. Shreiber 2001; Benaim, Shreiber, and Tarres 2004.

17 Of necessity, there is a little over-simplification here.

18 Benaim 1999; Pemantle 1990, 2007.

19 Argiento et al. 2009.

20 See, for instance, Blume et al. 2002 for an experimental study of signaling games that evaluates reinforcement learning with an exponential response rule. They, somewhat misleadingly, call this Roth–Erev reinforcement, but it differs from the Roth and Erev model in the response rule.

21 Cournot had his players, who were two duopolists controlling a market, alternate best responses. One might vary the dynamics by having the players best-respond at random times, and just keep doing the same thing otherwise. This is called *best response with inertia*.

22 Brown 1951. A lot has been learned about the properties of fictitious play since it was introduced. For a review see Fudenberg and Levine 1998.

23 See Fudenberg and Levine 1998 and Camerer and Ho 1999.

24 In the short run it may be hard to discriminate between these models. See Salmon 2001.