26. *Learning Models*

The most important attitude that can be formed is that of desire to go on learning.

—John Dewey

In this chapter, we study models of individual and social learning. We apply each in two contexts. The first setting involves learning the best choice in a set of alternatives. In that setting, both types of learning, individual and social, converge on the optimal choice. The choice of learning rule only affects the rate of convergence. We then apply the learning rules to actions in games. In a game, an action's payoff depends on the action of the other player or players. In that setting, both learning rules favor risk-averse equilibrium outcomes over efficient ones. We also find that individual and social learning need not produce the same result and that neither performs better in all environments.

These findings bolster our many-model approach to representing behavior. Learning models lie in between the rational-choice models, which assume that people think through the logic of situations and games and take optimal actions, and rule-based models that assign behaviors. Learning models do assume that people follow rules, but those rules enable behavior to change. In some cases, the behavior converges to optimal behavior. In those cases, learning models can be used to justify the assumption that people optimize. However, learning models need not also converge to equilibria; they might produce cycles or complex dynamics. If the models do converge, they may select some equilibria more than others.

The chapter begins by describing a reinforcement learning model and applies it to the problem of choosing the best alternative. The model reinforces actions with higher rewards. Over time, the learner takes only the best action. This is a baseline model that proves ideal for learning about learning. It also fits quite well with experimental data, and not just for humans. Sea slugs, pigeons, and mice all reinforce successful actions. It may be a better model of sea slugs, which possess fewer than 20,000 neurons, than of humans, who have more than 85 billion. That extra capacity allows humans to consider counterfactuals when learning, a phenomenon left out of the reinforcement learning model.

We then introduce social learning models, where individuals learn from their own choices and the choices of others. Individuals copy the actions or strategies that are most prevalent or that are performing above average. Social learning requires observation or communication. Some species create social learning through *stigmergy*: a process in which successful actions leave a trace or residue that others can follow, such as when goats who roam a mountain range leave trampled grass, reinforcing routes to water or food.

In the third section, we apply both types of learning models to games. As already noted, games present a more complicated learning environment. The same action might produce a high payoff in one period and a low payoff the next. As might be expected, we find that both social and individual learning models can fail to converge to efficient equilibria. They can also produce different outcomes. We conclude with a discussion of more sophisticated learning rules.[1]

Individual Learning: Reinforcement

In *reinforcement learning*, an individual chooses actions based on the weights of those actions. Actions with a lot of weight are chosen more often than actions with little weight. The weight assigned to an action depends on the reward (payoff) that a person has received from taking that action in the past. This reinforcement of high-reward payoffs leads to better actions being taken. The question we explore is whether reinforcement learning converges to only choosing the alternative with the highest reward.

At first, it may seem that to choose the most rewarding alternative is a trivial task. If the rewards are expressed in numerical form, such as money or time, we would expect people to choose the best. In Chapter 4, we invoked that line of thinking to argue that a person choosing a route to work in Los Angeles would settle on the shortest one.

If rewards do not take numerical form, which is generally the case, people must rely on memory. We grab lunch at a Korean restaurant. We find the kimchi delicious, so we are more likely to eat there again. On Monday, we eat an oatmeal cookie an hour before running and find we can sustain a strong pace for ten kilometers. If prior to Wednesday's run we again grab an oatmeal cookie and perform well, we add weight to that action. We learn that cookies improve our performance.

Other species do the same. Edward Thorndike, an early psychologist who studied learning, conducted an experiment in which cats who pulled a lever to escape a box were rewarded with fish. When returned to the box, the cats pulled the lever within seconds. Thorndike's data revealed a process of continued experimentation. He found that cats (and people) learned faster when he increased the reward. He called this the *law of effect*.[2] This finding has a neurological explanation. Repetition of an activity builds neurological pathways that induce that same behavior in the future. Thorndike also found that more surprising rewards, rewards that far exceeded past or expected outcomes, produced faster learning in people, a phenomenon known as the *surprise principle*.[3]

In our reinforcement learning model, the weight assigned to a chosen alternative is adjusted based on how much the reward from that alternative exceeds our expectations (our *aspiration level*). This construction embeds both the law of effect (we take actions that produce higher rewards more often) and the surprise principle (the amount of weight we add to a choice depends on how much its reward exceeds the aspiration level).[4]

A Reinforcement Learning Model

A collection of **alternatives** $\{A, B, C, D,\ldots, N\}$ have associated **rewards** $\{\pi(A), \pi(B), \pi(C), \pi(D),\ldots,\pi(N)\}$ and a set of strictly positive **weights** $\{w(A), w(B), w(C), w(D),\ldots,w(N)\}$. The probability of choosing K is as follows:

image

After choosing allternative K, $w(K)$ increases by $\gamma \cdot (\pi(K) - \Theta)$, where $\gamma > 0$ equals the **rate of adjustment,** and $\Theta < \max_K \pi(K)$ equals the **aspiration level**.[5]

Notice that the aspiration level must be set below the reward for at least one alternative. Otherwise, any alternative chosen becomes less likely to be chosen in the future and all of the weights converge to zero. It can be shown that if the aspiration level is below the reward for at least one alternative, eventually almost all of the weight will be placed on the best alternative. This occurs because each time the best alternative is selected, its weight increases by the most, creating stronger reinforcement of that alternative. This occurs even if we set the aspiration level below the reward from each alternative. In that case, every alternative increases in weight when selected. Thus, the model can capture *habituation,* where we do more of something just because we have done it in the past. Even with a low aspiration level, the alternatives with the highest rewards increase in weight the fastest, so the best alternative wins out in the long run. However, the time required for convergence on the best alternative may be long. It will also be true that as we add more alternatives, time to convergence also increases.

To avoid these complications, we can build in *endogenous aspirations*. We emend the model so that the aspiration level adjusts over time by setting it equal to the average reward. Imagine a parent learning whether a child prefers apple pancakes or banana pancakes. Assign a reward of 20 to apple pancakes and 10 to banana pancakes. Set the initial weights on both alternatives to 50, the rate of adjustment to 1, and the aspiration level to 5. If the parent makes banana pancakes the first day, the weight on banana pancakes increases to 55. As this was the first day, set the aspiration level equal to 10. If the parent makes banana pancakes the next day as well, the weight on banana pancakes does not change because the reward equals the aspiration level.

If, on the third day, the parent makes apple pancakes, this choice produces a reward of 20, 10 above the aspiration level. The weight on apple pancakes increases to 60, making them the more likely choice. The average payoff now equals 13.3. It follows that if the parent makes banana pancakes again, the weight on banana pancakes will decrease because the reward lies below the new aspiration level. Reinforcement learning therefore converges to only apple pancakes being selected.

It can be proven that reinforcement learning will converge toward selecting the best alternative with probability 1. That means that the weight on the best alternative will become arbitrarily large compared to the weights on all other alternatives.

Reinforcement Learning Works

In the **learning-the-best-alternative framework,** reinforcement learning with the aspiration level set equal to the average earned reward (eventually) almost always selects the best alternative.

Social Learning: Replicator Dynamics

Reinforcement learning assumes an individual acting in isolation. People also learn from watching others. Social learning models assume that individuals see the actions and rewards of others. This can speed the rate of learning. The most widely studied model of social learning, *replicator dynamics*, assumes that the probability of taking an action depends on the product of its reward and its popularity. We can think of the former as a *reward effect* and the latter as a *conformity effect*.[6] Most often replicator dynamics models assume an infinite population. We can then characterize the actions taken as a probability distribution across the various alternatives. In the standard construction, time advances in discrete steps so that we can capture learning by changes in the probability distribution.

Replicator Dynamics

A collection of **alternatives** $\{A, B, C, D, \ldots, N\}$ have associated **rewards** $\{\pi(A), \pi(B), \pi(C), \pi(D), \ldots, \pi(N)\}$. The actions of a population at time t can be written as a probability distribution across the N alternatives: $(P_t(A),$

$P_t(B),\ldots, P_t(N)$). The probability distribution changes according to the **replicator equation**:

image

where image equals the average reward in period *t*.

Consider a community in which parents choose between apple, banana, and chocolate chip pancakes. Assume that that all of their children have identical preferences and that the three types of pancakes produce rewards of 20, 10, and 5. If initially 10% of parents make apple, 70% make banana, and 20% make chocolate chip, the average reward equals 10. Applying replicator dynamics, the probabilities of choosing each of the three alternatives in period two are as shown in the table below:

The Replicator Equation

image

Applying the replicator equation, in the next period twice as many parents make apple pancakes. This occurs because the reward for apple pancakes equals double the average reward. Half as many parents make chocolate chip pancakes because that reward equals half of the average reward. Finally, the proportion of parents making banana pancakes, which produce exactly the average reward, does not change. Combining all of these changes, we can show that the average reward increases to 11.5.

As noted above, replicator dynamics includes a conformity effect (more popular alternatives are more likely to be copied) as well as a reward effect. In the long run, the reward effect dominates, because high-reward alternatives always grow in proportion to lower-reward alternatives. In replicator dynamics, the average reward performs a function similar to that of the aspiration level in reinforcement learning when the aspiration level adjusts to equal the average reward. The only difference is that in replicator dynamics, we calculate the average reward for a population. In reinforcement learning, the aspiration level equals an individual's average reward. That distinction matters insofar as a population provides a larger sample. Thus, replicator dynamics produce less path dependence than reinforcement learning.

In our construction of replicator dynamics, we assume that every alternative exists in the initial population. Given that the highest-reward alternative always has a higher-than-average reward and its proportion increases in every period, (eventually) replicator dynamics converge to the entire population choosing the best alternative.[7] Thus, in a setting of learning the best alternative, both individual and social learning converge to the alternative with the highest reward. That will not be true in games.

Replicator Dynamics Learns the Best

In learning the best from a finite set of alternatives, replicator dynamics with an infinite population converges to the entire population choosing the best alternative.

Learning in Games

We now apply our two learning models to games.[8] Recall that in a game, a player’s payoff depends both on her own action and on the actions of the other players. The payoff from a given action, such as cooperating in the Prisoners’ Dilemma, could be high in one period and low in the next depending on the action of the other player. We begin with the *Guzzler Game,* a two-person game in which each player must choose whether to drive an economy car or a gas guzzler. Choosing the gas guzzler always produces a payoff of 2. Choosing an economy car when the other player also chooses an economy car produces a payoff of 3—both drivers have good lines of sight, require less fuel, and have no fear of being crushed by an enormous gas guzzler. If the other player chooses a gas guzzler, a player driving the economy car must be cognizant of the other driver. To capture that effect, we assume that her payoff falls to zero. We represent these payoffs in figure 26.1.

image

Figure 26.1: The Guzzler Game

The Guzzler Game has two pure strategy equilibria: both players can choose economy cars or both players can choose gas guzzlers.[9] The equilibrium in which both choose the economy car produces the higher payoff. It is the efficient equilibrium.

image

Figure 26.2: Reinforcement Learning (γ = image) Probability of Choosing Guzzler

We first assume that both players use reinforcement learning. Figure 26.2 shows results from four numerical experiments with the initial weights on each action set equal to 5, an aspiration level of zero, and a learning rate (γ) of image. In all four experiments, both players learn to select the gas guzzler, the inefficient pure strategy equilibrium. To see why this occurs, we need only look at the payoffs. The gas guzzler always returns a payoff of 2. The economy car sometimes returns a payoff of 3 and sometimes returns a payoff of zero. By assumption, both actions will be equally represented in the initial population. Therefore, the economy car produces an average payoff of only 1.5 to the gas guzzler's payoff of 2. More players choose the gas guzzler, and the payoff from selecting the economy car decreases further.

image

Figure 26.3: Replicator Dynamics (100 Players): Probability of Choosing Guzzler

Next, we apply replicator dynamics to the same game. Again we assume an initial population consisting of equal proportions of people choosing gas guzzlers and economy cars. We further assume that each player plays the game against every other player in the population. People who choose the gas guzzler receive higher payoffs, and because initially equal numbers choose each action, in the second period more people will choose gas guzzlers.[10] Applying the replicator equation a second time, shows that the number of players choosing gas guzzlers would again increase. Continued application of the replicator equation results in the entire population choosing guzzlers. Figure 26.3 shows results from four runs of discrete replicator dynamics with 100 players. By assuming a finite population, we introduce a small amount of randomness. The proportions adopting each action may not be exactly equal to those stated in the replicator equation. In each of the four runs, all of the players choose the gas guzzler after only seven periods. Convergence occurs quickly because both the conformity effect and the reward effect push people to choose the gas guzzler after the first period. For example, when 90% of the population chooses gas guzzlers, the payoff from choosing an economy car will be less than one-sixth of that from choosing the gas guzzler. The conformity effect amplifies the reward effect, making social learning much faster than individual learning, which took, on average, more than 100 periods to reach 99% guzzlers.

In this game, both learning rules converge to choosing the gas guzzler because it has the higher payoff when both actions are equally likely. Such actions are called *risk dominant*. Both learning rules favored the risk-dominant equilibrium over the efficient equilibrium. We next construct a game in which our two learning rules converge to different equilibria.

The Generous/Spiteful Game

Our next game, *The Generous/Spiteful Game*, builds on a much-analyzed question about human behavior: Do we care more about our absolute or relative payoffs? A person who would prefer a $10,000 bonus when all of his colleagues receive $15,000 bonuses over an $8,000 bonus when all of his colleagues receive only $5,000 cares more about his absolute payoff. A person who would accept less money to have the largest bonus cares more about his relative payoff. An extreme preference for relative payoffs is captured in the story of the spiteful man and the magic lamp.

The Spiteful Man and the Magic Lamp

A spiteful man finds a bronze lamp while on an archeological expedition. He rubs the lamp and a genie appears. The genie proclaims, “I will grant *you one wish for anything that you desire, and because I am a benevolent genie, I will give everyone you know double what I give you." The man ponders the proposition, grabs a stick, and says, "Poke out one of my eyes."*

The spiteful man takes an action that gives him a low absolute payoff and a high relative payoff.[11] A similar tension exists in foreign affairs. Neoliberals believe that countries want to maximize absolute payoffs measured by military power, economic prosperity, and domestic stability. Another camp, known as neorealists, believes that countries value relative payoffs. A country would rather have a lower absolute payoff but be stronger than its enemies. Kenneth Waltz, a neorealist, wrote at the height of the Cold War, "The first concern of states is not to maximize power but to maintain their positions in the system."[12] Neorealists would claim that during the height of the Cold War, had either Russia or the United States rubbed the magic lamp, each would have handed the genie a stick.

We can embed the conflict between absolute and relative gains in an *N*-person game with a generous action that increases absolute payoffs for everyone along with a spiteful action that only increases one's own payoff. This game differs from a collective action game, where generosity comes at a cost.[13] The formal game with payoffs is shown in the box. The generous action is a dominant strategy. Regardless of the actions of the other players, a player choosing generous receives a higher payoff. However, on average the players choosing spiteful earn higher payoffs.

These may at first appear to be contradictory statements. They are not. By being generous a player raises his absolute payoff by 3 but also raises the payoffs of all other players by 2. A player who chooses to be spiteful raises his payoff by only 2 but does not raise the payoffs of the other players. Each player improves his payoff by choosing to be generous. When a player chooses to be spiteful, he reduces his payoff, but (and here's the key assumption) he reduces the payoff to everyone else by an even larger amount.

The Generous/Spiteful Game

Each of *N* players chooses to be *generous G* or *spiteful S*.

$\text{Payoff}(G, N_G) = 1 + 2 \cdot N_G$

$\text{Payoff}(S, N_G) = 2 + 2 \cdot N_G$

If we apply reinforcement learning in the Generous/Spiteful Game, the players learn to be generous. To see why, suppose that the players have almost converged to an equilibrium, with N_G of the players choosing to be generous. A spiteful player earns a payoff of $2 + 2 \cdot N_G$. This will be his aspiration level. If he chooses *G*, which occurs with small probability, he earns a payoff of $1 + 2 \cdot (N_G + 1) = 3 + 2 \cdot N_G$, which is above his aspiration level. He will become more likely to be generous. By continuing to apply this logic, we see that all players will learn to be generous.

If we apply replicator dynamics, the population learns to be spiteful. This can be seen by referring to the replicator equation. In every period, players who choose to be spiteful earn higher payoffs than players who choose to be generous. Therefore, the proportion of players choosing to be spiteful increases in each period.

These findings highlight a key difference between individual and social learning. Individual learning leads people to choose the better action, so people learn a dominant action if one exists. Social learning leads people to choose actions that perform well relative to other actions. In most cases, those actions would also produce higher payoffs. That is not the case in the Generous/Spiteful Game, where the spiteful action has a higher average payoff, while the generous action is dominant. Notice that our analysis arrives at the rather paradoxical finding that if people learn individually, they learn to act more generously than if they learn socially. That occurs because in social learning the players copy the actions of players who perform relatively well.

We might now take a moment to consider an earlier comment: that we can think of replicator dynamics as an adaptive rule or as the selection of fixed rules. If we assume the latter, then our model says that selection could favor spiteful types. Selection need not produce cooperation. This result runs counter to what we found when studying the repeated Prisoners’ Dilemma, where repetition led to cooperation. In that case, we considered repeated games and allowed for more sophisticated strategies.

Combining Learning Models

We have seen how individual and social learning both find the best solution among a fixed set of alternatives, but that when applied to games, they can produce different outcomes. This lack of agreement is a strength. Imagine a giant set consisting of all possible games. Imagine a second set consisting of all learning models. We could apply every learning model in the second set to every game in the first set and evaluate their performance. We can then partition the set of all games into two sets: those in which the learning rule produces the efficient outcome and those in which it does not. We could also look to experimental data and evaluate each learning rule as a predictor of actual behavior. That exercise would undoubtedly reveal contingencies. Each learning rule would result in efficient outcomes for some games but not for others. Each learning rule would also vary in the contexts in which it accurately describes behavior. Hence, we advocate many models.

In this chapter, we covered two canonical models. Each includes only a few moving parts. Our goal was to provide a gentle introduction to a large and exciting literature. By adding more details to either learning model, we would better fit experimental and empirical data. Recall that in the reinforcement learning model, individuals add or subtract weight to an alternative or action depending on whether its reward (payoff) exceeds the aspiration level. Individuals do not add weight to actions not taken: we do not increase the probability of taking some action that would have given a high payoff had we taken it.

That assumption may not make sense in all cases. Consider the case of an employee who decides not to take his cell phone on vacation. While he is away his boss calls with an important question. The employee misses the call and is passed up for a promotion. In the reinforcement learning model, the employee would not attach more weight to bringing his phone on vacation in the future. The *Roth-Erev learning model* amends the standard model so that alternatives that are not chosen also receive weight based on their hypothetical payoffs. In the example, the employee would attach more weight to bringing his phone.

This modification creates a *belief-based learning* rule. The amount of the increase in weight for the alternatives not chosen is determined by an experimentation parameter. The higher the experimentation parameter, the more individuals take into account the effect of others' actions and the more they increase the weights on those actions. Roth and Erev also discount the past to take into account that other players are learning as well and their strategies likely change.[14]

These additional assumptions make intuitive sense and have empirical support, but they do not fit all cases. If we go back to our example of the parent making pancakes, the first assumption implies that after the parent makes banana pancakes, additional weight is added to the alternative of making apple pancakes and that weight is proportional to the payoff from apple pancakes. Such an assumption makes sense only if the parent knows the payoff from apple pancakes. That would be true only if people can see or intuit the payoffs of unchosen actions.

A model by Camerer and Ho creates a functional form that admits both reinforcement learning and belief-based learning as special cases. A parameter that can be fitted to data allows a determination of the relative strength of each type of learning rule.[15] The ability to combine models was one motivation for mastering many models. That said, combining models necessarily leads to a better fit because of the increase in parameters. Even taking into account the parameter increase, Camerer and Ho's model produces better predictions and deeper explanations.

Modeling learning creates several challenges. Learning rules that work well in one setting may not capture other situations as well. Furthermore, what people learn to do can depend on their initial beliefs, so two people may learn differently in the same setting and the same person may learn differently in different settings. Even if we could construct an accurate learning model, we again confront the *exploitability principle*: if a model explained how people learned, then others could apply that model to anticipate (and in some cases exploit) that knowledge. It is then likely that people would learn not to be exploited, and our original learning model would no longer be accurate. We encountered this phenomenon earlier when discussing the Lucas critique and in our analysis of the efficient market hypothesis. We cannot necessarily conclude that because people learn that they optimize. We can assume learning will winnow out poor actions in favor of better ones.

Does Culture Trump Strategy?

We now apply contagion models and learning models to address the longstanding claim from organizational theory that culture trumps strategy.[16] In brief, the claim states that strategic incentives to change behaviors fail. The pull of culture, the existing set of repertoires and beliefs, proves too powerful. Economists argue the opposite: that incentives drive behavior.

To turn these opposite proverbs into conditional logic, we first apply a version of the network contagion model. In this model, the manager, or possibly the CEO, announces a new strategy and produces evidence for the benefits of the change. The CEO may even redefine the organization’s core principles to reflect this new behavior. Individuals in the organization then choose whether or not to adopt the behavior based on how compellingly the manager makes her case. Some initial proportion of people buy into the initiative. When they make contact with others in their work network, they spread their enthusiasm. There also exists a pull against the new strategy, causing people to no longer adopt the new strategy. The three features that determine if the new strategy spreads—the contact rate, the spreading rate, and the rate of abandonment—map naturally into the parameters in the basic reproduction number, R_0:

image

If we add in the possibility of superspreaders, then we might conclude that culture trumps strategy provided any of three conditions hold: if people do not believe in the new strategy, if they are quick to abandon it, or if the strategy’s advocates are not well connected. Otherwise, strategy may well trump culture.

Our second model applies replicator dynamics to a *Culture/Strategy Game* that models interactions between pairs of employees. We can represent these choices in game form as a *cultural action* (doing what they currently do) and as an *innovative strategic action*. We assume that the manager constructs payoffs so that both players earn higher payoffs if both choose to be innovative. However, a single innovative player earns less.

image

The Culture/Strategy Game

The game has two strict pure-strategy Nash equilibria: one in which both innovate (strategy trumps culture) and one in which neither innovates (culture trumps strategy). The manager appears to have constructed incentives so that the employees will take the innovative new action, as it has the higher payoff. If we write down a learning model, we see that the manager needs sufficient initial buy-in for the innovation to take hold. In the game above, it can be shown that if the initial buy-in, that the proportion that adopts the innovation in the first period, does not exceed 20%, then culture trumps strategy.[17] If we were to increase the payoff from the innovative strategic action, then initial buy-in could be even lower yet still result in the efficient outcome.

These two models show that the opposing proverbs “Culture trumps strategy” and “People respond to incentives” can both be correct conditionally. According to the first model, charismatic CEOs who can convince well-connected employees can introduce new strategies that trump culture. According to the second model, culture trumps weak incentives but not strong ones.