Chapter 7

From probabilities to actions

Nick Chater, Thomas L. Griffiths &
Mark K. Ho

This book has so far focused primarily on questions of inductive inference: inferring the structure of the environment, a sentence, or a category, from samples of data together with background knowledge. But the acquisition of new knowledge is ultimately only of practical value to an agent if it can help guide decisions concerning *action*. So, for example, an animal may classify possible foodstuffs to decide whether they should or should not be eaten; or interpret a looming shadow in order to initiate fight or flight. Perceptual inferences about the state of the environment or one's own body may be used to guide reaching, maintain balance, or avoid collision. Causal inferences about the operation of a piece of physical apparatus, or a computer interface, will determine how a user can achieve their goals. And in the social and economic realm, inferences about the motives of another person may determine whether they are friend or foe, what they do or don't know, etc, and hence how we should best to interact with them.

Understanding decision-making is also important from a methodological point of view, because the vast majority of experimental data record behaviors which result from decisions. So, while studying perception using a psychophysical method, we often rely on participants' responses indicating what they saw, or whether a stimulus was visible at all; and such responses are, of course, themselves the result of a decision-making process[1]. Moreover, observing people's decisions can be used to make inferences about the subjective probabilities that underlie those decisions. Indeed, in experimental economics, inference from observed decisions is the primary methodology for inferring a person's subjective probabilities—an approach which is rooted in theoretical results about ideal rational agents, as we shall see. Thus, we can see decision-making both as the ultimate object of most cognitive processes; and as a medium through those processes may be glimpsed.

In this chapter, we summarize the key ideas needed to translate from a probabilistic representation of the world to rational action. This is in itself a topic that could occupy an entire book, and indeed there are several that we recommend (e.g., Robert, 2007; Sutton & Barto, 2018; Russell & Norvig, 2021). Our goal here is to provide an introduction to these ideas at a level of detail that makes it possible to understand the topics presented in the second half of the book. We begin by introducing **statistical decision theory**, which tells us how rational agents should balance probabilities with rewards. We then consider how those rewards should be represented, introducing the idea of **utility functions**, and connect this approach to Bayesian inference via evidence accumulation. Many real-world settings do not involve making a single decision in isolation but rather a sequence of decisions, which leads us to the topic of **reinforcement learning**. A further extension to Bayesian theories of decision-making is required to account for decisions concerning the learning process itself. Rather than passively processing whatever data comes our way, the brain is engaged in **active learning**, directing its limited information processing capacities to sample and process information that is likely to prove most valuable or interesting. We conclude the chapter by considering the apparent contrast between Bayesian theories of decision-making (especially of basic cognitive processes, from detecting sensory signals, recalling memories or planning and executing motor movements) and the vast empirical literature in psychology and behavioral economics which appears to show that people's decisions often radically depart from rational action.

7.1 Minimizing losses: statistical decision theory

To begin, we consider the question of what an agent should do if it has well-defined subjective probabilities (in line with Bayes' theorem and the rest of the laws of probability, of course), and has a well-defined objective that can be quantified in numerical terms. Suppose, for example, that a person is attempting to detect faint targets – such as brief flashes – in an environment where there are also distractors (and also noise in the perceptual system itself). Suppose we study this detection problem in the lab, running an

[1] By contrast, brain imaging, or neural recordings, and typically fine-grained behavioral analysis, such as studying patterns of eye-movements or autocorrelations in reaction time distributions, pick up *incidental* products of cognitive processes which are not the result of deliberate decision. We decide to respond, say, that a stimulus was present on the screen, we do not decide to respond in 750ms or with a particular pattern of neural activity.

experiment with a series of discrete trials, where on each trial the experimental participant either presses a button (to signal that the target is present) or does nothing (signalling that the target is absent on that trial). Perhaps the simplest way to evaluate performance is just to count up the number and types of right and wrong answers that the person gives. There are two types of right answer: a "hit," when the target flash was present, and the button was pressed, and a "correct rejection" when the target was absent, and the button was not pressed. And there are two types of wrong answers: a "miss," when the button was not pressed even though the target was present, and a "false positive," where the button was pressed while there was no target.

This type of set-up can be modeled via statistical decision theory (Berger, 1993). For largely historical reasons, in statistical decision theory, we normally talk about minimizing a loss, rather than maximizing an objective. So, applying the simplest possible loss function, a 0-1 loss function, we can assign a score of -1 to every mistake; and 0 to every correct answer. Then the objective of our hypothetical person can be modeled as minimizing the sum of the losses. When choosing an action, of course, the agent doesn't yet know what these losses will be. Hence, the natural strategy is to choose the action that minimizes *expected* losses, where the expectation is based on one's current subjective probabilities. So, for example, with the 0-1 loss function, on each trial the agent simply needs to press the button if the subjective probability of the target, given the sensory evidence and prior information, is greater than 1/2; if the probability is less than 1/2, then the agent should not press the button. And if it is exactly 1/2, then either pressing or not pressing the button has the same expected loss, so either option is equally good, and the choice can be made arbitrarily.[2]

7.1.1 Asymmetric loss functions

This type of signal-detection task has long been studied by psychologists (Green & Swets, 1966), in applications as diverse as detecting brief flashes of light, radar images indicating approach enemy aircraft, or medical scans which may indicate cancer. And in general a 0-1 loss function will be too simple: some errors are much more important than others. So, for example, a "false positive" in which a person is wrongly suspected of having cancer, and then subjective to further testing, is an annoyance; but a "miss" in which a person with cancer is overlooked, and hence does not receive potentially life-saving treatment is a disaster. To account for this, we need what is called an asymmetric loss function—which can apply different penalties for the two losses (and still a loss of zero for "hits" and "correct rejections"). Suppose, for example, that in our cancer detection case, we judge that a "miss" should incur a loss which is 1,000 times greater than a false alarm. For concreteness, let's set these losses to -1 and -1,000 respectively. Suppose that our priors and sensory information from the scan lead us to a posterior probability p that the person has a cancer, where p is a smallish number, say 1/100 (or 0.01).

We have two actions: declaring a "positive" or a "negative" test result. If we were to stick to the 0-1 loss function, the expected loss of declaring a positive test result is $p.(0) + (1-p)(-1) = p - 1 = -.99$, with $p = 0.01$. The expected loss of declaring a negative test result is $p.(-1) + (1-p).(0) = -p = -.01$. We want to minimize our expected loss, so we should declare the test negative (and presumably send the patient away with no plans for further investigation). But suppose we switch to our asymmetric loss function, taking account of the fact that missing a cancer is much more serious than a false positive. Now the expected loss of declaring a positive test result is $p.(0) + (1-p)(-1) = p - 1 = -.99$; and the expected loss of declaring a negative test result is $p.(-1000) + (1-p).(0) = -p = -10$. Now the expected loss is minimized by declaring a positive test result, even though the actual probability of having cancer, given the declarative of a positive test, is rather remote. The asymmetric loss function causes the agent

[2]The prior will typically involve incorporate base-rates, concerning how often the target tends to appear, but might capture more complex patterns, for example of the present or location of the target seems to follow some regularity. Equally, the presence or absence of the target might depend on previous experience concerning what the target looks like, how variable it is, and so on—the calculations concerning the subjective probabilities can be arbitrarily complex.

to "err on the side of caution," and most likely make some further investigations rather than sending the patient home with a clean bill of health.

The type of approach can be generalized in many ways. For example, rather merely detecting a target, the aim might be to categorize a target and respond appropriately (e.g., eating fruit that is ripe, storing fruit which is unripe, and disposing of fruit that is over-ripe or moldy). Now there will be payoff matrix between categories and actions, and the loss function will of course not be symmetrical. Eating moldy fruit is a much more serious error than occasionally disposing of potentially edible fruit. But the same approach, choosing actions which minimize expected loss given the agent's subjective probabilities, can be applied.

7.1.2 Continuous actions

In many contexts, the natural measure of performance depends on not just choosing the right category of action, but on the real-valued precision of an continuous-valued action[3]. So, for example, when reaching for an object, it might matter how close we are to that target. In other cases our output may not be a physical movement, but could be a numerical judgment: for example, the market value of an antique, the length of a river, or the population of city. In statistics, two particularly popular loss functions for real-valued outputs are the squared (or quadratic) loss function, where loss is the sum of the squared distance between estimates y and the targets t, $(y-t)^2$, and the absolute value loss function, where loss is the sum of the absolute distance between estimates and the targets, $|y-t|$. The squared loss function is, of course, widely used as a default in regression problems in statistics and machine learning (see, e.g., Hastie, Tibshirani, & Friedman, 2009). Both loss functions are minimized if the error is zero: if an action or prediction precisely hits the target value, while the quadratic loss function is more sensitive to large errors (because those errors are squared).

From the point of view of choosing actions, these loss functions are very simple special cases. The loss function involved in real-world behavior will usually need to be tailored to the specific behavior being considered. So, for example, in guessing the age of a small child, the "loss" in terms of the irritation of the child might be large for underestimation, but small for overestimation. So minimizing expected loss will encouraged people to give upwardly biased estimates. Or suppose our action is making an offer on a second-hand car. If an agent has a probabilistic model of the likely minimum price that the seller will accept, given the characteristics of the car, the seller, and so on, how should the agent choose what sum to offer? Here, too, the loss function must be tailored to the situation. If we offer too little, we fail to buy the car, and have to continue searching (incurring costs of time and inconvenience); if we offer too much, we lose financially. A Bayesian approach to decision-making requires that we put these different types of loss on a single scale, and choose the action (here, our bid), which minimizes the overall expected loss, taking into account these different factors. And in reality, the story is complicated further by the possibility of the initial bid being followed by subsequent bargaining, and so on. So while choosing actions using Bayesian decision theory is conceptually straightforward—we just minimize expected loss—in practice it is typically very complex. Thus, a realistic cognitive model will typically assume that such computations must be approximated, perhaps drastically.

In simple experimental contexts, this approach to decision-making can provide a good model of behavior. For example, Trommershäuser, Maloney, and Landy (2003) ask participants rapidly to touch a green target on a touch screen (earning points) while avoiding a nearby, or even overlapping, red target (losing points). The experimenters can measure perceptual and (more importantly) motor noise in this task, roughly correponding to a Gaussian distribution around the true target. Here, then, the participant must choose where to aim, in the light of the gains and losses available (as noted above, the task could

[3]There is also another important generalization, to minimize a function of the entire action trajectory, rather than merely the end-point of the action. We do not consider this case here. These might include choosing a trajectory to minimize energy expenditure, imitating another person's actions, or dancing in synchrony to music, or in a particular style

be reframed entirely in terms of losses, and remain formally identical[4]). The elegant aspect of this task is that by aggregating over many trials, the experimenters can observe the degree of noise in participants' responses directly, and also infer where they are aiming. It turns out that people's behavior is well-predicted by the assumption that they are attempting to minimizing expected loss in this task—that is, they "aim off" from the center of the green target by the appropriate amount, as a function of the degree of loss associated with the red target.

7.1.3 Deviations from optimality

Results like these seem promising for a Bayesian model of action, and indeed, there is a large literature on Bayesian models of motor control which operate within this general framework (e.g., Kording & Wolpert, 2006; McNamee & Wolpert, 2019). On the other hand, though, there are extremely simple tasks in which people's behavior seems to depart dramatically from the optimal Bayesian response. One particular striking example is the phenomenon of probability matching (for a review, see Vulkan, 2000). In a typical task, on each trial a light is either green or red, and the participant has to guess the color on the next trial—the reward is often just the sum of the number of correct answers. Suppose, in reality, the green and red colors are selected by independent flips of a biased coin, with probability p of green and probability $1 - p$ of red. If the participant can figure out this distribution (or, in some variants, is also told about the underlying mechanism), then the Bayesian choice is straightforward. Suppose that the participant's estimate of the probability of the next coin being green is q (which is in general not quite equal to the true p, of course). Sticking with a framing in terms of losses, let us assign a loss of -1 to incorrect guesses and the usual loss of 0 for correct guesses. Then the expected loss for choosing green is just the subjective probability of a red, $1 - q$; and then expected loss for choosing red is the subjective probability of a green, q. Given that the aim is to minimize losses, we should consistently choose green when $1 - q < q$, i.e., when the subjective probability of green q is greater than $1/2$, and choose red if q is less than $1/2$, and choose arbitrarily if the probability of green and red is equal. This is simply a roundabout way of stating what might appear to be completely obvious: that we should always choose green if we think green is the most likely next item; and always choose red if we choose red is most likely.

While this may be the "obvious" strategy, it is surprisingly rarely observed in experiments. So, for example, Shanks, Tunney, and McCarthy (2002) find that even after hundreds of trials, and where after each block of fifty trials people are explicitly told how well they are doing, and how much better they would be doing if they used an optimal strategy, only a rather small number of people end up consistently choosing the more probable option. In many experiments, people's choices are better captured by the simple model by which they select green with probability q, and select red with probability $1 - q$: that is, their responses *match* the probabilities of each outcome. The precise conditions under which probability matching occurs, and how it is to be explained, have been widely debated. Linking back to the sampling approximations of Bayesian inference that we described in the last chapter, it is interesting to note that one simple explanation is that people are choosing red or green by drawing samples from the underlying distribution, rather than minimizing expected losses (see, e.g., Vul, Goodman, Griffiths, & Tenenbaum, 2014). But for the moment, the key point is that people seem systematically to fail to solve a Bayesian decision problem despite its extreme simplicity.

[4]It is well-known, of course, that while framing a task in terms purely of losses, or in terms of gains, or in terms of a mixture of losses and gain, makes no mathematical difference, it may affect the actions and choices that people actually make, in ways that are typically viewed as inconsistent with statistical decision theory and indeed with the more general rational choice framework which we shall describe below (Kahneman & Tversky, 1979).

7.2 Utilities and beliefs

So far, we have taken a loss function as given. But the question of what (if anything) human behavior is attempting to optimize in a particular setting (or indeed more broadly) is typically challenging. Only in very restricted circumstances, such as maximizing points in a video game, is there a clear and externally given "objective." But our daily lives require us to chooses complex courses of action without any externally given well-defined objective, or where many objectives need somehow to be traded off against each other.

Let us step back a little. In general terms, deciding what to do should, as we have indicated, depends partly on one's beliefs (about the external world, and sometimes also the thoughts and likely actions of other agents it contains); and the formation of beliefs, and the concepts of which they are constructed, has been the focus of this book so far. But decisions depend not only on what an agent believes, but also on its **desires**, **goals**, or **objectives**, for which we shall use the blanket term **utility**. Most normative theories of decision-making propose a rather strict conception of utility: that each relevant state of the world, S_i, (which might arise as a consequence of one's action) is associated with a number, representing the utility $U(S_i)$ of that outcome, for the agent.[5]

A rather stripped-down notion of utility is in play here. So, for example, an agents utility may not turn purely on its own "well-being," or its ability to meet its own objectives, but might depend on the well-being of others, or the attainment of some purely external goal. There is no assumption that utility need be reducible to sensory pleasures and freedom from physical pain (although this viewpoint was popular among early utilitarian economists and political philosophers such as Bentham, Edgeworth and Sidgwick, see, e.g., Cooter & Rappoport, 1984), but might be determined by abstract goals; and there is no assumption that the agent need be aware of their desires, or indeed have awareness of any kind.

In the laboratory, the objective of experimental participants can sometimes be specified externally: to maximize the number of points in a game, such as the game of hitting green (and avoiding red) targets that we discussed above. Similarly, our performance might be scored based on our ability to produce the correct answer in an arithmetical calculation, to recall accurately which items on a list have been seen before in an earlier phase of the experiment, or to determine correctly when a signal is present against a noisy background. In this type of case, computational models can straightforwardly be associated with utilities that directly capture the structure of the task—we are in the familiar territory described in the previous section. But we can also make some headway in modeling thought and behavior even when there is no externally-given objective.

To start, let us note that, quite generally, when choosing an action, we do not know with certainty what the consequences of that action will be. Indeed, if each action had only one possible outcome,[6] then deciding which action to choose would be fairly straightforward: simply choose the action leading to the outcome with the greatest utility. The standard (although by no means the only) way to choose an action is to aim to maximize *expected* utility—though, crucially, this utility is not an externally given standard, but is assumed to reflect the aims of the agent. So, consider an agent contemplating an action, a. If the agent takes action a, it believes that each possible state of the world which will be the outcome of this action s, has a probability $P(s|a)$. Then the expected utility of action a, $EU(a)$, is the sum of the utilities of each possible outcome of the action, weighted by the probability of each outcome:

$$EU(a) = \sum_s P(s|a)\, U(s). \tag{7.1}$$

The principle of maximizing expected utility provides a general-purpose criterion which applies, in principle, to a wide variety of decisions, and has been applied to foraging, investment, partner choice, shot

[5] We will henceforth use on the standard terminology in economics and psychology in which people are seen as maximizing utilities rather than minimizing losses. Of course, any maximization problem can be recast as a minimization problem, and vice versa, simply by flipping the sign of the function to be optimized.

[6] This special case, where actions have a certain outcome, is known as a deterministic MDP.

selections in tennis, and many more. For each action that you might take, just consider the probabilities and utilities of the possible consequences of each action, and hence derive that action's expected utility. Then just choose the action with the greatest expected utility.

This bracingly general and direct formula for deciding what to do is, however, not necessarily easy to follow in practice, and we shall explore some of the complexities that arise in building cognitive models of decision-making in the following sections. A first complication is that the very existence of a meaningful utility measure, which the agent can be viewed as maximizing, cannot be taken for granted. It is this problem to which we now turn.

7.3 When can a utility scale be defined?

The Bayesian approach to decision-making proposes that behavior can be understood as maximizing expected utility, perhaps approximately. This approach can only get started, of course, where a notion of utility is well-defined. As mentioned above, in psychological experiments the participant's objective is often specified directly—e.g., to maximize the number of points, or to make as few errors as possible. Similarly, in evolutionary arguments in biology, e.g., concerning sex ratios, mating or child-rearing strategies, some variant of Darwinian "fitness," perhaps defined at the level of the individual gene rather than the whole organism, is a useful externally-defined objective (Dawkins, 1978). But, in general, the goals being pursued in human behavior are not pre-specified. Indeed, people typically have a myriad of diverse aims that appear to compete for their attention. Thus, a driver may want to arrive quickly, drive safely, avoid traffic violations, plan a meeting, and send an urgent message to a colleague. It may be difficult to satisfy these objectives simultaneously: objectives, such as speed and safety, may conflict and need somehow to be traded-off against each other.

To apply the expected-utility perspective, we need to be able to combine diverse constraints and aims into a single overall measure (a scale of utility) reflecting the relative importance of each. If such a scale of overall utility can be constructed, then in principle at least, the driver's problem is now clear: the best sequence of actions is that which leads to the maximum expected utility. But when can such a utility scale be defined? That is, what rationality constraints need be imposed on a person's choices, such that it is possible to explain their behavior in expected-utility terms at all (see Chapter 2 for a broader discussion of constraints on rational coherence)?

A naïve approach to this problem is simply to construct a utility function directly: for example, we might try to measure each objective on a continuous scale and take, for example, a weighted sum of these as our overall utility function. But of course this approach is unlikely to capture an agent's preferences successfully. For example, it is not clear how objectives concerning properties as diverse as speed, safety, probability of traffic violations and so on, can be measured on comparable scales, how they should be combined, and what weight should be attached to each.[7] Fortunately, though, there are general results that establish the conditions under which a utility scale can be defined, merely by looking at the structure of an agent's preferences.

7.3.1 From preferences to utilities

Suppose, for a moment, we ignore any issues of risk and uncertainty and consider an agent's choices between outcomes which are certain: e.g., between known foods, activities or consumer goods. An ideally rational decision-maker might be presumed to have preferences which follow a number of natural rules. Suppose, for example, that the decision-maker can compare any two outcomes A and B, either

[7] Indeed, diverse scales may have fundamentally different properties, e.g., being nominal, ordinal, interval or ratio scales, in the terminology of measurement theory (Narens & Luce, 1986), making combining such scales inherently problematic.

preferring B to A (which we write $A \preceq B$), or A preferring to B ($B \preceq A$), or being indifferent between them.

Suppose further than the decision-maker's preferences are transitive: if $A \preceq B$, and $B \preceq C$, it seems reasonable that $A \preceq C$. These *completeness* and *transitivity* assumptions are sufficient to ensure the existence of a utility function U, which assigns numbers to outcomes, A, B, so that $A \preceq B$ if and only if $U(A) < U(B)$, and the agent is indifferent between X and Y just when $U(X) = U(Y)$.[8]

This utility function encodes an *ordering* over outcomes, from most to least preferred; but it does not capture the *strength* of preference between those outcomes. Any stretching or squashing of those numerical values, so long as order is preserved, will do equally well as a utility function, if we are choosing, say, whether we would prefer to be given an apple or an orange. All that matters is which items have higher (or lower) utility values. This dependence only on order is captured in the term *ordinal* utility—and it turns out that the minimal notion of ordinal utility provides a sufficient foundation to construct many parts of microeconomics, such as theory of supply and demand in the formation of market prices (e.g., Kreps, 1990).

From the point of view of cognitive science, however, a richer notion of utility is required. The theme of this book is that cognition involves dealing with an uncertain world, and that probability theory provides a framework for understanding how this is possible. Accordingly, we require an account of decision-making that can reflect the fact that our actions may often lead to a variety of possible outcomes. Consider, for example, a simple action such as picking up a cup of coffee. On the one hand, we do not want to expend inordinate amounts of time and care in such a simple action; on the other hand, as our movements become more hurried, the probability of spillage increases; in this, as in many actions, some balance between effort and probability of success must be found. And to make this trade-off sensibly we need to have to know more about *how much* we prefer different outcomes.

For simplicity, consider the case where actions correspond simply to choosing between monetary gambles (e.g., let us imagine our decision-maker is at the casino). Suppose, for example, a person has the choice of \$50 for certain, or a .5 probability of either \$0 or \$100. Given only an ordinal scale of utility, we can say only that $U(\$100) > U(\$50) > U(\$0)$, given the minimum assumption that more money is better than less. But to determine whether our decision-maker should gamble or play safe, we need to know *how much* the utility of \$100 exceeds the utility \$50, in relation to how much the utility \$50 exceeds the utility of \$0 dollars. Hoping to buy a last-minute ticket to an expensive concert, the decision-maker might very much prefer \$100 to either \$50 or \$0, because only this amount is enough to buy a ticket; such a decision-maker might be expected to take the gamble. Other decision-makers, requiring just \$5 for a pizza, on the other hand, might have the opposite preference, particularly disliking the \$0 dollar outcome, which might leave them hungry, and only marginally preferring \$100 to \$50. In short, what is required is a *cardinal* utility scale: A scale that assigns a meaningful numerical value to each state, so that, in particular, the difference in utility between states can be quantified.

7.3.2 Deriving real-valued utilities

It turns out that our previous mentioned assumptions of completeness and transitivity, now applied to *gambles* rather than certain outcomes, imply, along with fairly mild technical assumptions, that outcomes of those gambles can be associated with real-valued utilities, so that preferences between lotteries over those outcomes are captured by the *expected* utilities of those lotteries.[9]

[8] This is true if the number of outcomes is finite or countably infinite. By contrast, e.g., if we must choose between outcomes parameterized by one or more real-values, an additional technical continuity assumption is also required.

[9] In addition to a version of the continuity assumption, a more substantive assumption is *independence*: roughly, that utility assigned to each outcome is not affected by the other outcomes that might alternatively occur. For any three lotteries L, M, N and any probability $p > 0$, $L \preceq M$ if and only if $pL + (1-p)N \preceq pM + (1-p)N$. Intuitively, if the lottery formed by the probabilistic combination of L and N is no more preferable than the lottery formed by the same probabilistic

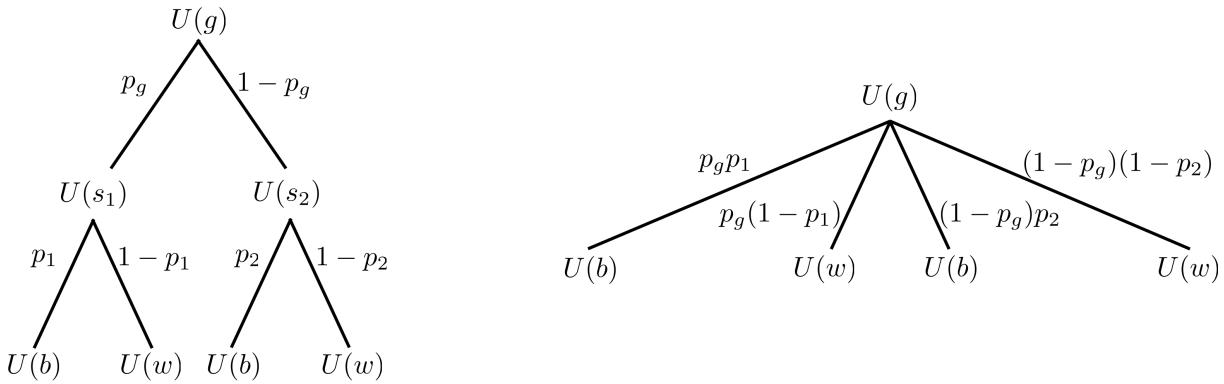


Figure 7.1: Assigning utilities to an arbitrary gamble g using best/worst mixtures of subgambles s_1 and s_2 is equivalent to a single-stage gamble with mixtures of the best and worst gambles b and w.

Such a scale can, at least in principle, be derived constructed from preferences, if we allow preferences to range over gambles rather than just fixed outcomes. Consider the following procedure. First pick the worst possible outcome under consideration, w, and the best, b, and arbitrarily assign these two to have numerical utilities $U(w)$ and $U(b)$, where, of course, $U(w) < U(b)$. For concreteness, and without loss of generality, let us set $U(w) = 0$ and $U(b) = 1$, so that the utilities of all states under consideration will lie on the $[0, 1]$ interval. Then, pick any other outcome, s_i, which is preferred to w, but less preferred than b. Using the assumption that any relevant options can be meaningfully compared, just as in the case of ordinal utility above, then we can ask whether s_i would be preferred to a gamble having the best outcome, b, with probability p_i and the worst outcome, w, with probability $1 - p_i$. If p_i is high enough, the gamble will, of course, be preferred; if p_i is low enough, then it will be rejected. For each s_i, there must be some value p_i at which the balance is struck – the decision-maker is indifferent between the certainty of the outcome s_i, and a gamble with a probability p_i of yielding b, and a probability of $1 - p_i$, of yielding w (we shall leave aside discussion of the assumptions required to make this line of reasoning rigorous, and whether those assumptions are justified (Neumann & Morgenstern, 1944; Edwards, 1954; Kreps, 1988)).

If we follow this procedure for each outcome s_i, the corresponding probabilities, p_i, provide a real-valued measure of the "goodness" of those outcomes. The best state b has a value of one (b is, of course, trivially equivalent to a gamble which yields b with probability one); a worse state w has a value of zero (because this state is trivially equivalent to the gamble which yields b with probability zero and w with probability one). Then the higher the value p_i that is associated with the outcome s_i, the higher its utility. Indeed, this value can play the role of the cardinal utility of s; it will us to determine preferences of a decision-maker both over outcomes, and gambles over outcomes.

From this standpoint, then, how do we assign a utility to an arbitrary gamble g, which has outcome s_1 with probability p_g, and outcome s_2 with probability $1 - p_g$? First, we associate each outcome s_1 and s_2 with the equivalent subgambles involving the best and worst states, b and w, associated with probabilities p_1 and p_2, respectively. Let us call these gambles *best/worst mixtures*. Then the decision-maker should be indifferent between our original gamble g, and a gamble with a probability p_g of facing

combination of M and N, then and only then $L \preceq M$. So if I prefer yogurt to ice-cream, then I should prefer, say, a fifty-fifty chance of yogurt or fruit to a fifty-fifty chance of ice cream or fruit. Note, of course, that independence is a claim about the irrelevance of *alternative* possible outcomes; it is entirely possible that I might prefer yogurt to ice-cream when eaten alone, but prefer ice-cream to yogurt when eaten with fruit (so that preferences about combinations of items will in general not be independent). Some decision theorists reject the independence axiom (and the closely related Sure Thing principle; Savage, 1972), which has some counterintuitive implications, most famously, Allais's Paradox (Allais, 1953). Indeed, there are decision theories which involve weakening various principles of standard expected utility theory (for discussion see, e.g., Bradley, 2017).

a best-worst mixture parameterized by probability p_1; and a probability $1 - p_g$ of facing a best/worst mixture parameterized by probability p_2 (see Figure 7.1). But, assuming that the goodness of a gamble purely depends on its outcomes and their probabilities, then we can collapse this two-stage gamble into a one-stage gamble. Specifically, in the two-stage gamble, there are two independent ways in which the best possible outcome b can arise: with probability p_g, we face the subgamble parameterized by p_1 and "win" – a sequence with probability $p_g p_1$; and with probability $1 - p_g$, we face the subgamble parameterized by p_2 and "win" – a sequence with probability $(1 - p_g)p_2$. Thus the total probability of attaining the best state b is the sum $p_g p_1 + (1 - p_g)p_2$; otherwise, the decision-maker faces the worst outcome w. Thus, we have a new best/worst mixture, parameterized by the probability $p_g p_1 + (1 - p_g)p_2$ of the best option (otherwise, the outcome is, of course, the worst option). This probability can be viewed as a measure of the goodness of a compound gamble—the greater the probability of getting the best, rather than the worst, outcome, the better.

The generalization to gambles with many outcomes follows the same pattern. A gamble with n possible outcomes $s_1, \ldots, s_n$, where the ith outcome has probability $P(s_i)$ should be equivalent to a best/worst mixture with a probability $\sum_i P(s_i)p_i$, where p_i is a parameterization of the best/worst mixture which the decision-maker treats as equivalent to the outcome s_i.

To see that the probability of "winning" in the best/worst mixture can be used as a measure of utility, let us blithely rewrite such probabilities as utilities. That is, let us replace p_i with u_i, so that our formula for the value of a gamble becomes not $\sum_i P(s_i)p_i$, but $\sum_i P(s_i)u_i$. And this is, of course, just the familiar formula for expected utility: the utility of each possible outcome, weighted by its probability.

So far, we have identified cardinal utility with a particular probability – the probability of winning a best/worst mixture – so these utilities are necessarily defined only on the $[0, 1]$ interval. But this restriction is not necessary. All preferences will be unchanged if all utilities are multiplied by an arbitrary factor; or if an arbitrary constant is added or subtracted to all utilities. That is, cardinal utility is our defined only up to a positive linear transformation – they can be defined on any portion of the real line. The absolute size of the numbers used to represent utilities, and whether these numbers are positive or negative, is not important; it is the relative differences between the utilities of different states that matters.

Indeed, more general and sophisticated arguments of this type can be provided. Given surprisingly minimal consistency criteria concerning the preferences of our hypothetical decision-maker (though these criteria may, nonetheless, be violated in reality by human and animal decision-makers), it can be shown that there exist a set of utilities *and* subjective probabilities, such that the decision-maker's preferences between options, whether simple states or gambles, perfectly follow the principle of maximum expected utility (e.g., Savage, 1972).[10] The specific utilities and subjective probabilities will, though, vary from person to person—even fully rational agents can still have distinctive beliefs and preferences, and hence make very different choices. Rationality simply ensures that these choices are coherent within the individual.

7.3.3 Revealed preference and cognitive science

The result presented in the previous section is particularly interesting from a methodological point of view. It suggests that, given sufficient information about a (rational) agent's preferences, we should be able to infer both the utilities and the probabilities the agent assigns to different outcomes. In economics, this observation has been the foundation of the **revealed preference** approach to utility (and, by extension, probability) (Samuelson, 1938; Savage, 1972) – the idea that probabilities and utilities are

[10] The Bayesian approach, here and through this book, models uncertainty with probabilities; some theorists, particularly in economics, argue that uncertainty can sometimes be so open-ended that probability cannot always meaningfully be applied, see (e.g., Knight, 1921; Binmore, 2008).

revealed by an agent's choices, rather than being directly measurable psychological or neural properties. From this standpoint, choice behavior is seen as primary, and probability and utility are merely convenient theoretical variables for predicting such behavior. The revealed preference standpoint has been viewed as providing a crucial separation between economics (which requires that minimum consistency assumptions are obeyed, so that convenient utility and probability scales can be inferred) and cognitive science. Crudely, from this point of view, economics need only be concerned with what people *choose*, not what they *think*.

This revealed preferences style of argument has been taken to imply that, given fairly minimal consistency and other conditions (which we have skated over here), there must exist a notion of utility such that a rational decision-maker should always prefer actions with the highest expected utility. As we noted above, the conditions required to establish this result may not always apply to real human or animal decision-makers. Nonetheless, the principle that choices should be determined by maximizing expected utility, where a suitable notion of utility is well-defined, is a gold standard in rational models of decision-making across a range of disciplines, ranging from economics and the social sciences, to behavioural ecology, artificial intelligence, and cognitive science.

How should we view such explanation, and, in particular, how should we view rational explanation in cognitive science? Taking up the standpoint of traditional economics, one possibility is that we view the type of Bayesian analysis outlined in this book as claiming only that the mind (or brain) behaves "as if" it makes probabilistic calculations: the probabilities are presumed to be constructions of the theorist, rather than descriptions of internal mental or neural states.

While this may indeed be the appropriate interpretation for some Bayesian models, it is also possible that probabilities (and perhaps utilities) *are* mentally represented, and that behaviour is not merely patterned "as if" the brain carries out Bayesian calculations and calculate maximum expected utility, but, rather choice behaviour is the outcome of such mental calculations: i.e., as a result of algorithms for probabilistic calculations. From this perspective, the brain is able to behave "as if" it were a probabilistic inference and expected utility maximizing engine, precisely because, in some domains at least, it *is* a probabilistic inference and expected utility maximizing engine. And, as we saw in Chapter 6, probabilistic inference need not carried out precise symbolic manipulation of the mathematical formulae of probability theory, but through approximate methods, such as through sampling from probability distributions. In the next section we consider how the problem of accumulating the evidence required to make a simple decision might be achieved by simple psychological and neural mechanisms.

7.4 The accumulation of evidence

Let us consider a particular illustration of how we might go beyond the "as if" viewpoint. As discussed above, perhaps the simplest type of decision, and one extensively studied by psychologists, is signal detection. A person is instructed, say, to respond "yes" on every trial in which a brief flash is present and to respond "no" otherwise (Green & Swets, 1966). The optimal strategy is to say "yes" whenever the posterior probability of the light being present is above some threshold, with the threshold being determined by the loss incurred by different outcomes. We can also derive this optimal strategy within the expected-utility framework introduced in the previous section.

For example, suppose that the participant obtains a reward of $5c$ each time the signal is correctly detected; loses $50c$ in the case of a false alarm; and receives nothing otherwise. Under this regime, the participant is likely to be extremely tentative. Suppose, on a particular trial, the participant estimates that the probability that the signal is present is q. Then their expected utility for saying "yes" is $q\,U(5c) + (1-q)\,U(-50c)$. The expected utility for saying "no," by contrast is $U(0c)$, for convenience, we can set $U(0c) = 0$ (this is possible with no loss of generality, because the utility scale is only defined up to a positive linear transformation). Thus, choosing the option "yes" yields strictly greater expected

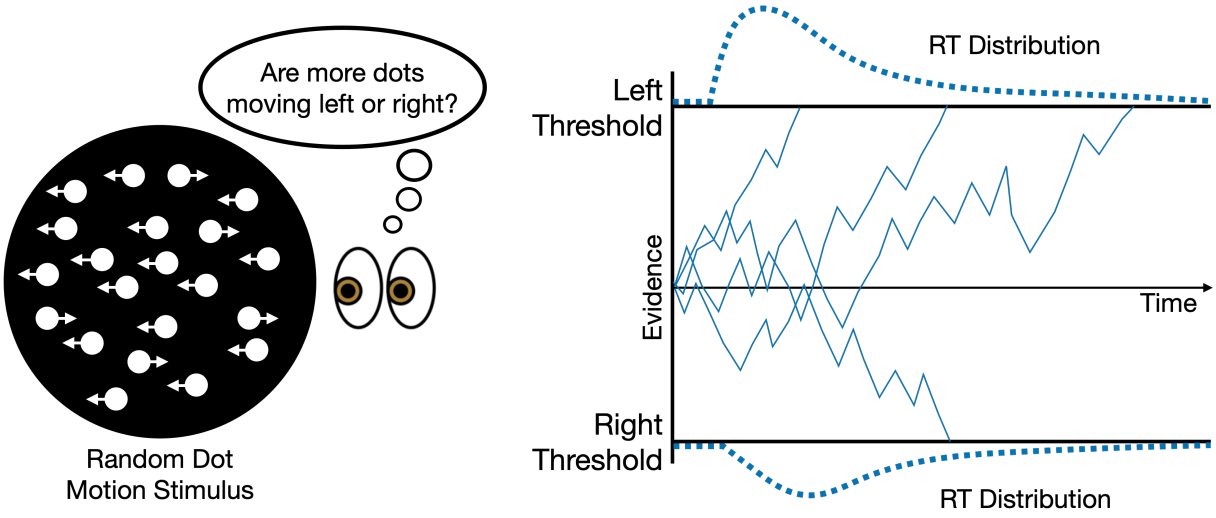


Figure 7.2: The random dot motion paradigm and a visual depiction of the basic drift diffusion model (Ratcliff, 1978). As the perceptual system accumulates evidence for the dots moving to the left or the right, the weight of evidence (depicted with solid blue lines corresponding to different stimuli) changes in value. When the evidence hits a threshold – either for motion to the left or the right – a decision is made. The response time (RT) distribution results from this stochastic process.

utility than choosing the option "no" when $q\,U(5c) + (1-q)U(-50c) > 0$, assuming that the absolute value of $U(-50c)$ is very much greater than the absolute value of $U(5c)$ (roughly, losing $50c$ is a great deal worse than gaining $5c$), then this will only be true when q is high. In the special case where utility is a linear function of money losing $50c$ will be exactly ten times worse than gaining $5c$, and simple algebra shows that the "yes" response will only have strictly greater expected utility than the "no" response when $q > 10/11$. Notice that we have already seen this type of explanation, when minimizing a loss function—but here, of course, we are seeing the problem as maximizing a utility. But, as we've seen already, really there is no difference: maximizing utility is just the same as minimizing a loss function which is set equal to negative of that utility.

Signal detection theory has proved to be a highly effective descriptive model across a wide range of psychophysical tasks. And it has traditionally been viewed from the revealed preference standpoint prevalent in economics: i.e., as assuming only that the experimental participants' behavior conforms descriptively with the theory. But it turns out that signal-detection models can also map naturally onto a simple computational mechanism – **diffusion models** – which can capture how the propensity to push for one decision or another can build up over time (Ratcliff, 1978; Usher & McClelland, 2001; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Brown & Heathcote, 2008; Ratcliff, Smith, Brown, & McKoon, 2016; Forstmann, Ratcliff, & Wagenmakers, 2016).

From a Bayesian point of view, these models can be viewed as accumulating relative strength of evidence which favors taking one decision or another (e.g., evidence that a signal is present or that it is not). Or, to pick what has become an important experimental task (Newsome & Pare, 1988; Britten, Shadlen, Newsome, & Movshon, 1992; Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012), suppose that we must decide whether a noisy random dot pattern, briefly presented on a computer screen is flowing, overall, to the left or to the right.

The flow of dots of noisy: that is, while dots are predominantly flowing either left or right, a fraction of the dots are flowing in the opposite direction. Making an overall judgment therefore requires accumulating fragments of information from individual dots—a process that will unfold over time. The sum

of these fragments traces a random walk in the strength of evidence favoring one decision or the other, sometimes moving towards one decision, sometimes the other. Specifically, suppose that h_L and h_R are the hypotheses concerning whether the dots are drifting left or right; and let us call the different pieces of sensory data $d_1, d_2, \ldots, d_n$. A simple application of Bayes' rule allows us to compare the posterior odds of the two hypotheses, given the data:

$$\frac{P(h_L|d_1, d_2, \ldots, d_n)}{P(h_R|d_1, d_2, \ldots, d_n)} = \frac{P(h_L)}{P(h_R)} \frac{P(d_1, d_2, \ldots, d_n|h_L)}{P(d_1, d_2, \ldots, d_n|h_R)}. \tag{7.2}$$

Assuming each piece of data d_i is independent, given the hypotheses, h_L and h_R, then the likelihood term for all the data can be decomposed into a product of the likelihoods of each piece of data,

$$\frac{P(h_L|d_1, d_2, \ldots, d_n)}{P(h_R|d_1, d_2, \ldots, d_n)} = \frac{P(h_L)}{P(h_R)} \prod_{i=1}^{n} \frac{P(d_i|h_L)}{P(d_i|h_R)}. \tag{7.3}$$

Taking logarithms of both sides, the likelihood term now becomes a sum of likelihood terms, each reflecting the strength of evidence in favour of one or other hypothesis, in the light of each new piece of data d_i. Given the noisy nature of the stimulus, some pieces of data will favor one hypothesis, and some will favor the other,

$$\log \frac{P(h_L|d_1, d_2, \ldots, d_n)}{P(h_R|d_1, d_2, \ldots, d_n)} = \log \frac{P(h_L)}{P(h_R)} + \sum_{i=1}^{n} \log \frac{P(d_i|h_L)}{P(d_i|h_R)}. \tag{7.4}$$

As more data is processed, the overall sum will gradually drift in the direction of the hypothesis which is best supported by the evidence.[11]

When the random walk hits a pre-defined boundary—which signifies the strength of evidence required to trigger a decision—a choice is made. The location of these boundaries will reflect the utilities involved in the decision. In a standard signal-detection experiment, these utilities will be shaped by the different numbers of points or monetary payoffs for hits, misses, and false-alarms. The same considerations arise, of course, for real-world decisions. For example, if a person or animal foraging for food must decide whether a fungus should be treated as a mushroom or a toadstool, then considerable evidence that it is a mushroom (i.e., is edible) will be required; even a slight suspicion that it may be a toadstool (i.e., poisonous) may be enough for the fungus to be cast aside. This asymmetry in the position of the decision boundaries captures the fact that the utility gain from eating the mushroom is small in comparison with the utility loss from being poisoned (though, of course, in extreme circumstances, the utilities calculations may be rather different—when near starvation, even hazardous food may have greater expected utility that no food at all).

Diffusion models, whether or not interpreted in Bayesian terms, are widely used and quantitatively successful in many areas of psychology, modeling aspects of perception, categorization, the initiation of movements and recognition memory, among many other topics (e.g., Hanes & Schall, 1996; Lamberts, 2000; Ratcliff, 1978; Smith & Ratcliff, 2004). One attraction of these models is that they provide fine-grained predictions about distributions of reaction times, speed-accuracy trade-offs, sensitivity of changes of payoffs, and confidence judgments (Pleskac & Busemeyer, 2010; Berg et al., 2016) (see the right hand panel of Fig 7.2 to see how reaction time distributions are generated). Note that, as described, such models are limited to dealing deciding between pairs of options; various generalizations have, though, been proposed (e.g., Usher & McClelland, 2001).

[11] The pioneering Bayesian statistician I. J. Good advocated the second term, the log-odds ratio in favor of one hypothesis as against the other, as an general measure of weight of evidence (e.g., Good, 1950). He attributed the idea to Alan Turing during his development of methods to break the Enigma code (Good, 1979). For a broader discussion of other measures of confirmation, see Crupi, Chater, and Tentori (2013). Non-Bayesian approaches to the analysis of sequential data ignore the priors (Wald, 1947; Green & Swets, 1966). For simple decisions such as those considered here, this is not a significant restriction, however, as the role of the prior can still implicitly be captured by shifting the decision bounds—that is, the weight of evidence required before a decision is triggered.

Might the brain be accumulating sensory evidence and making simple decisions by implementing a diffusion model of such kind, thus providing evidence for a neural implementation of Bayesian calculations? An important line of research, involving neural recording in the monkey, suggests that this might be the case (Gold & Shadlen, 2007). In a typical experiment, the monkey is presented with the random dot motion detection task, as described above; and it is rewarded on each trial if it moves its eyes in the same direction as the flow. It turns out that the firing rates of some populations of neurons in the monkey's brain (in lateral intraparietal cortex, or LIP) appear closely to track the weight of evidence (e.g., Gold & Shadlen, 2002), rather than, for example, whether, or which, a decision is about to be made (although the causal relationship between such accumulation mechanisms and choices has been questioned; Katz, Yates, Pillow, & Huk, 2016). More broadly, a subfield of research mapping neural activity to mechanism for evidence accumulation and decision-making in a range of perceptual and motor tasks has become yielded promising results, as part of the general viewpoint that the brain is carrying out approximate Bayesian computations (e.g., De Lafuente, Jazayeri, & Shadlen, 2015; Knill & Pouget, 2004; Pouget, Beck, Ma, & Latham, 2013).

7.5 Sequential decision-making

Simple decisions, such as whether a signal is present or absent, or whether a random dot pattern is flowing left or right, are appealing starting points for experimentation and modeling. But, of course, the brain faces decisions of vastly greater complexity, on a variety of dimensions. Often the options between which we must decide are themselves highly complex (e.g., when choosing a house, an artwork, piece of music, or a possible friend), and the process of evaluating sensory and linguistic evidence may be arbitrarily complex (e.g., in recognizing, making sense of, and evaluating, an object, scene, artwork or person). Here, though, we focus on specific, and well-studied, aspect of decision complexity: the question of how to choose sequences of actions, or to develop policies for how to act. This is crucial, because individual actions typically have no well-defined value, except in the light of subsequent actions. Saving, rather than spending, money now may be advantageous—but not if the decision-maker will squander the accumulated savings in a gambling spree. Similarly, a squirrel caching food for winter will be benefit only if it is likely to retrieve it later; studying for an exam only makes sense if you intend to take it; reaching for a glass of water makes sense only if you intended to grasp it, and so on. Quite generally, our actions, whether life plans or individual motor actions, make sense only if the individual component actions are part of a larger coherent framework. This creates particular problems in learning which actions to take, because the stream of rewards or punishments received by the agent will depend on combinations of many actions, and it will typically be difficult to determine which individual action should be modified in order to improve decision-making in the future. Here, we consider some interesting special cases, which have received considerable attention both in machine learning and in the cognitive and brain sciences. Below, we begin by outline the key mathematical ideas in the abstract—later we will consider some of the many ways they can be applied in models of cognition and behavior.

7.5.1 Sequential decision-making problems

Problems related to taking sequences of actions have been studied most extensively in the literature on **planning** and **reinforcement learning**. The basic model of sequential decision-making is the discrete **Markov Decision Process (MDP)** (Puterman, 1994), which assumes that there exists a discrete set of states of the environment $\mathcal{S}$, a discrete state of actions an agent can take in that environment $\mathcal{A}$; that there is a transition function that defines distributions over next states given a previous state and action $T(s, a, s') = p(s_{t+1} = s' \mid s_t = s, a_t = a)$; and that there is a one-step reward function that maps state-action combinations to positive or negative real numbers $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. So in a simple card game, the states of the environment might just be the distribution of cards across players; the actions

of the agent might be picking up another card (twist) or doing nothing (stick); the transition function would determine the new state of the cards (which depends on which card is drawn from the pack); and the reward function might be the sum of the "points" that a person's hand of cards represents (possibly, as in Pontoon, with the crucial complication that if the sum is over some threshold, the person gets zero points). So the challenge of the player is to decide when to stop drawing cards, to maximize the likely points outcome. In the context of MDPs, an agent's behavior is conceptualized as a stimulus-action mapping or **policy**, formalized as a function mapping states to actions, $\pi : \mathcal{S} \rightarrow \mathcal{A}$. Thus, in a card game, a policy specifies, for each distribution of the cards, which action to take (e.g., actions might include: pick up a card from the table, throw away a card, do nothing, and so on, depending on the game being played).[12]

The first problem that arises in sequential decision-making is that of prediction (also called **policy evaluation**): given a policy and an initial state, how much reward would be obtained in the long-term by following the policy? Specifically, suppose we start from an initial state s_0, repeatedly take the action dictated by a policy $a_t = \pi(s_t)$, calculate the reward $r_t = R(s_t, a_t)$, and sample a new state from the transition function $s_{t+1} \sim T(s_t, a_t, \cdot)$. This generates a trajectory or **roll out** $\langle s_0, a_0, r_0, s_1, a_1, r_1, ... \rangle$. What is the long-term reward generated by such a trajectory? While there are different ways of defining what long-term reward means, one standard approach that has nice mathematical properties is to use the **expected cumulative discounted infinite sum of rewards** (i.e., value) associated with a state, given by

$$V^{\pi}(s_0) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{7.5}$$

where $\gamma \in (0, 1]$ is a discount rate that ensures that the infinite sum converges and the expectation is over the states and rewards generated by following the policy π and sampling according to the transition function[13]. The function denoted by V^{π} is typically called the **value function** for a policy π (it's worth noting that this is a technical term in reinforcement learning, distinct from intuitive notions of value or uses in economic studies of decision-making). An important property of discounted infinite horizon MDPs is that the value function can be compactly written as a set of stationary recursive equations over states $s \in \mathcal{S}$, known as **Bellman's equations** (Bellman, 1957):

$$V^{\pi}(s) = R(s, a) + \gamma \sum_{s'} T(s, a, s') V^{\pi}(s'), \tag{7.6}$$

where $a = \pi(s)$. Intuitively, Bellman's equation expresses the idea that the value at a state depends on the immediate reward given by $R(s, a)$ and discounted expected value from the next state onwards given by $\gamma \sum_{s'} T(s, a, s') V^{\pi}(s')$.

It is worth emphasizing that the distinction between reward and value in MDP model enriches the standard notion of expected utility in an important and cognitively relevant way. Specifically, the reward function models how a decision-maker assigns *intrinsic* utility to states of the world, such as food for a

[12]Here, we treat the environment as completely observable—anything that we don't know (the order of the cards on the table, the cards of other players) is simply viewed as a source of randomness affecting, which will affect the results an agent's actions. But this general approach can be extended to deal with **partially observable MDPs (POMDPs)** (Kaelbling, Littman, & Cassandra, 1998), where the agent has to deal with uncertainty about the true state of the environment, as indeed is typically the case in most cognitively realistic scenarios.

[13]Note that a discount rate is sufficient but not necessary for the value function to be finite. In cases where the structure of an MDP guarantees that values for a policy are finite, the discount rate can equal 1. Additionally, one way to interpret the discount rate is as a constant probability of the task represented by an MDP continuing versus transitioning to a 0-rewarding state forever (i.e., 1 minus the probability of terminating). An alternative interpretation in economics is that this style of **exponentional discounting** is, under certain assumptions, required to avoid what is known as **dynamic inconsistency**, which occurs when a preference between future two options changes purely as a function of time passing. In practice, experiments with human and indeed animals suggest that the psychology of balancing immediate and future rewards is far more complex (Loewenstein & Prelec, 1992), but we leave aside such issues here.

hungry animal[14]. The value function, on the other hand, corresponds to utility that is *derived* from a combination of reward, the environment, and future behavior. As we discuss more in Section 7.5.4, this makes the MDP model especially useful for modeling learning and computation in sequential decision-making settings.

The second problem that arises in sequential decision-making is **optimal control** (also called **policy optimization**): Given an MDP, what policy would maximize value? Finding a policy with the maximal value function is often what is meant by "solving" an MDP. However, this raises a new question: Since both policies and value functions are functions over states, what exactly do we mean by the maximal function? One that has the highest value at a specific set of states? Any states? All states? Fortunately, one of the attractive mathematical properties of infinite discounted MDPs is that there is a unique optimal value function that has the highest value over all states (there may not be a unique deterministic optimal policy though, since actions could be tied in value). Furthermore, the optimal value function can also be concisely expressed as a set of recursive **Bellman optimality equations**:

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s'} T(s,a,s') V^*(s') \right\} \tag{7.7}$$

The intuition here is that the value of a state is determined by the value we can achieve if we choose the best action—and this action will generate some immediate reward, and put us in a new state (according to the probabilistic transition function), which will have its own value. Thus we can recursively link together the value of current and future states.

The Bellman optimality equations express **state values**, but often we are also interested in the (closely related) value of an action taken at a certain state assuming that we will act optimally from then on. This quantity is often referred to as the **Q-value** (as in "quality") and corresponds to:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') V^*(s'). \tag{7.8}$$

Once we have Q-values in hand (or some way to quickly compute them from R, T, and V^*), any policy that is greedy with respect to the optimal Q-values (i.e., at each stage chooses the action with the highest, or tied highest, Q-value) is an optimal policy:

$$\pi^*(s) = \arg\max_a Q^*(s,a). \tag{7.9}$$

To summarize, MDPs provide a way to model basic sequential decision-making tasks, and one standard approach to modeling long-term rewards or value is the expected, cumulative, discounted infinite sum model. This model allows us to concisely define two computational problems: prediction, in which one wants to evaluate a policy at various states, and optimal control, in which one wants to find a policy that maximizes value. Having the Bellman equations is a good start, of course—but we actually need to solve the equations in an efficient way to evaluate policies and determine which policy is optimal. In the following sections, we discuss algorithms from planning and reinforcement learning that can solve these problems under different starting assumptions.

7.5.2 Prediction and control with a known model

Given a known reward $R(s,a)$ and transition model $T(s,a,s')$, several algorithms exist for prediction and control. Control with a known reward and transition model is often referred to as **planning**.

[14]Whether a notion of intrinsic utility is always applicable is by no means clear, especially for humans.

A broad class of sequential decision-making algorithms is based on **dynamic programming**. In dynamic programming, we assume access to the full state space and calculate the value function via backward induction, repeatedly backing up the values of future states onto potential predecessor states until the values of all states converge. Specifically, starting from an initial value function V_0, we calculate the $k+1$-th value at a state from the k-th value function (applying the Bellman optimality equation, above). In the case of policy evaluation, this is:

$$V^{\pi}_{k+1}(s) = R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{\pi}_{k}(s'), \tag{7.10}$$

where $a = \pi(s)$. The dynamic programming algorithm for optimal control finds the optimal value function and is known as **value iteration**. It works using backward induction in a similar manner by updating V^{*}_{k+1} according to:

$$V^{*}_{k+1}(s) = \max_{a} \left\{ R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{*}_{k}(s') \right\}. \tag{7.11}$$

Note that value iteration does not require us to explicitly represent the policy when computing the optimal value function.

Dynamic programming and value iteration form the theoretical foundation for a wide range of other sequential decision-making algorithms, including temporal difference learning algorithms (discussed in the next section). Moreover, there is also a connection between value iteration and **heuristic search** algorithms, which are used for planning when R and T are known but the state space $\mathcal{S}$ is too large to completely enumerate. In typical heuristic search algorithms, we assume we are given a set of initial states $\mathcal{S}_0 \subset \mathcal{S}$ and can construct a state transition graph by examining successor states according to T. One approach is to alternate between expanding the transition graph and solving for the optimal solution in that graph, using the solution to guide the next round of expansion. In cases where graph construction is also guided by an admissible heuristic (that is, one that always underestimates the total cost from a state), this process can be analyzed as a form of *asynchronous value iteration* over a dynamically changing subset of states that is guaranteed to converge to an optimal policy for the initial states. This way of viewing heuristic search provides a unifying perspective on classic deterministic planning algorithms such as A* (Hart, Nilsson, & Raphael, 1968) as well as MDP planning algorithms like LAO* (Hansen & Zilberstein, 2001) and tree-search based algorithms (Kocsis & Szepesvári, 2006). For further details, see Ghallab, Nau, and Traverso (2016).

Aside from dynamic programming, another way to evaluate a policy given a known T and R is to solve the system of linear equations given by Equation 7.6. To see why this is the case, we can rewrite the policy evaluation equations in terms of vectors and matrices, where $\mathbf{r}$ denotes a vector of state rewards such that $\mathbf{r}_i = R(s_i, \pi(s_i))$, $\mathbf{T}$ denotes the state transition matrix conditioned on the policy $\mathbf{T}_{i,j} = T(s_i, \pi(s_i), s_j)$, and $\mathbf{v}$ is the vector of state values that we are solving for. In matrix notation, the system of equations specified by Equation 7.6 can be written as:

$$\mathbf{v} = \mathbf{r} + \gamma \mathbf{T} \mathbf{v}. \tag{7.12}$$

We can then simply rearrange the terms into standard matrix-vector form as follows:

$$[\mathbf{I} - \gamma \mathbf{T}]\, \mathbf{v} = \mathbf{r}, \tag{7.13}$$

where $\mathbf{I}$ is the identity matrix. Once in this form, standard linear algebra algorithms can be used to solve for the value function $V^{\pi} = \mathbf{v}$.

Besides value iteration, an alternative approach to planning in an MDP is to search through the space of policies directly. This idea underlies the **policy iteration** algorithm. Here, we begin with an initial policy π_0 and then repeatedly evaluate it and greedily improve the policy until the policy stops

changing. That is, we alternate been computing V^{π_k} for a policy π_k (e.g., using dynamic programming) and calculating an updated greedy policy:

$$\pi_{k+1}(s) = \arg\max_a R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{\pi_k}(s'). \tag{7.14}$$

This procedure can be iterated until it reaches a fixed point (note though that ties between actions must be broken in a consistent manner, otherwise the algorithm might cycle between equivalent policies and never converge). Reassuringly, and perhaps surprisingly, it can be shown that the resulting policy is globally optimal (Sutton & Barto, 2018).

The algorithms for sequential prediction and control reviewed in this section can be used when the reward and transition functions are known. However, it is often the case that we don't have complete information about the form of a sequential decision problem, and need to infer at least one of these quantities. We next turn to a family of algorithms for when one or both functions are unknown.

7.5.3 Prediction and control with an unknown model

How do we evaluate policies or find optimal policies when a model of the environment is not known? This is precisely the type of situation that reinforcement learning algorithms are designed for. Current approaches can be divided into **model free** approaches, which aim to estimate or optimize the value function without estimating $R(s,a)$ and $T(s,a,s')$ explicitly, and **model-based** approaches, which seek to estimate a model of the environment from which value can be calculated using methods such as those in the previous section (Sutton & Barto, 2018). Typically, model estimation largely reduces to the kind of unsupervised learning problems we have already discussed in detail in this book – estimating probability densities, inferring latent variables, and building graphical models – so the focus here will be on model-free approaches.

Model-free reinforcement learning algorithms typically define simple learning rules that solve problems of prediction and control without needing to build a model of the environment at all. The aims is simply to learn from experience which actions, and sequences of actions, are the most successful through what can be viewed as a highly sophisticated version of trial-and-error learning. One broad class of model-free algorithms are based on the idea of estimating value functions by **approximate dynamic programming**—that is, by performing a stochastic approximation to true dynamic programming when one can only draw samples from the environment by interacting with it. For example, **temporal difference** (TD) learning methods update a representation of a value function using s, a, s', r samples generated at each timestep. The update rule for TD prediction to estimate a value function V^π when following the policy π is:

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha(r + \gamma V^\pi(s') - V^\pi(s)), \tag{7.15}$$

where α is a learning rate. Why does this make sense? The key idea behind the TD update rule is that changes to value estimates are driven by **prediction errors** weighted by the learning rate. More formally, we can observe that at convergence, the value function should match the Bellman equation, so for any state, action, and reward, we expect:

$$V^\pi(s) = r + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V^\pi(s'). \tag{7.16}$$

If there is a mismatch between the left- and right-hand sides of the equation, then subtracting $V^\pi(s)$ from the right-hand side will give us an error signal, indicating whether $V^\pi(s)$ is too high or too low based on the values of the other states. Denoting this error signal δ, we have

$$\delta = r + \gamma \sum_{s' \in \mathcal{S}} T(s,s') V^\pi(s') - V(s) \tag{7.17}$$

which justifies the simple learning rule above, which can now be written:

$$V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha\delta. \tag{7.18}$$

However, this still requires us to know $T(s, a, s')$. We obtain the TD prediction rule (Equation 7.15) not be trying to infer this directly, but by treating the state s' as a *sample* from the distribution associated with $T(s, a, s')$ – a single sample Monte Carlo approximation to the expectation in Equation 7.17 (recall the discussion of Monte Carlo sampling from Chapter 6). Provided enough iterations of learning are performed, and the learning rate is decreased appropriately over time, the algorithm will converge to a correct estimate of $V^{\pi}(s)$.

In TD control, we want to estimate the optimal action-value function $Q^*(s, a)$ rather than a state-value function since the problem of control is to guide action selection. For example, **Q-learning** uses the update rule:

$$Q^*(s, a) \leftarrow Q^*(s, a) + \alpha(r + \gamma \max_{a'} Q^*(s', a') - Q^*(s, a)). \tag{7.19}$$

We can understand this update rule using a similar argument to the TD prediction update rule above. Specifically, the second term on the right-hand side represents a prediction error weighted by the learning rate. Convergence to the true $Q^*(s, a)$ will occur as the number of iterations increase, provided α is decreased appropriately over time. Additionally, a useful property of Q-learning is that it is **off-policy**: that is the estimate of $Q^*(s, a)$ is independent of the policy followed by the agent and can consequently be based on any sequences of states, actions, and rewards as long as there is sufficient coverage of the state/action space.

7.5.4 Reinforcement learning and cognitive science

Formalisms for planning and reinforcement learning are useful because they provide a unifying, normative framework for understanding adaptation in terms of estimating or maximizing value. In particular, because all correct reinforcement learning algorithms, by design, will converge to a well-defined value function, they inherit some a priori normative justification as potential models for biological learning. The choice of a *specific* algorithm (e.g., model-based versus model-free learning) reflect different assumptions about computational tradeoffs or mechanisms available.

From an historical perspective, the development of the field of reinforcement learning is an excellent example of how attempts to engineer and reverse-engineer intelligent systems can lead to fruitful exchange of scientific insights across different levels of analysis. The earliest reinforcement learning algorithms were psychological models that formally described behavioral patterns in Pavlovian conditioning (e.g., Bush & Mosteller, 1955; Rescorla & Wagner, 1972). It was later realized that these mechanisms could be recast in the normative framework of dynamic programming (Bellman, 1957) and TD-learning (Sutton & Barto, 1987). These basic ideas have served as the foundation for several decades of research on learning in sequential decision-making settings, culminating in a number of breakthroughs in artificial intelligence over the past decade (e.g., surpassing humans in games such as as Atari, chess, and Go, Mnih et al., 2015; Silver et al., 2016).

For cognitive scientists, the principles underlying reinforcement learning algorithms provide key insights into adaptation in humans and other species. Here, we review several research threads, starting with cognitively simple models that link TD prediction with Pavlovian conditioning and working our way to more complex models of task hierarchies and model-based planning.

Pavlovian conditioning and TD prediction In Pavlovian (or classical) conditioning, an organism learns an association between an unconditioned stimulus that is instrinsically rewarding (e.g., water for a thirsty dog) and a conditioned stimulus (e.g., the sound of a bell that reliably precedes water). In the reinforcement learning framework, the unconditioned stimulus corresponds to a state with a positive

reward (s_{UC}, $R(s_{UC}) > 0$) and the conditioned stimulus corresponds to a state without a reward but that reliably transitions to the unconditioned stimulus (s_C). The estimated value of the conditioned stimulus ($V(s_C)$) then corresponds to the associative strength between the unconditioned and conditioned stimulus while TD prediction (Equation 7.15) characterizes learning dynamics for building an appropriate association given experiences. Despite the simplicity of its learning rule, the basic TD algorithm can account for a wide variety of learning phenomena studied in classical conditioning (Sutton & Barto, 1987). Moreover, work in neuroscience paints a compelling picture of how TD learning is implemented in the brain: The reward-prediction error δ described by TD learning has been found to correspond to phasic (i.e., transient) activity of the midbrain dopamine neurons and provides a global signal for synaptic modification (Schultz, Dayan, & Montague, 1997; Glimcher, 2011). These results represent a remarkable convergence of findings at all three of Marr's levels of analysis that we discussed in Chapter 1: the problem of value estimation (computational), TD prediction/stochastic approximation (algorithmic), and phasic dopamine (implementation).

Operant conditioning, control, and model-based versus model-free learning Whereas classical conditioning involves forming value-based associations between states from sequences of observations, operant (or instrumental) conditioning involves forming value-based associations between different states and actions from trial-and-error (Thorndike, 1898). Specifically, in an operant conditioning experiment, an organism takes an action in a state (e.g., a rat pressing a lever when a light is on) and then an outcome occurs that may be rewarding or punishing (e.g., a food pellet appears). These kinds of scenarios, especially when they involve extended sequences of states, actions and rewards, are particularly suited for modeling within the reinforcement learning framework.

As we touched on earlier, one of the most important dichotomies in the space of possible reinforcement learning algorithms is between model-based and model-free learning. Recall that in model-based reinforcement learning, an organism learns a model of the environment (i.e., a transition function and reward function) and then uses this model to compute a value function. Model-based reinforcement learning has been taken to correspond to people engaging in deliberative reasoning about what action makes the most sense in their environment (e.g., Daw, Niv, & Dayan, 2005). By contrast, in model-free reinforcement learning, the organism learns the value function directly (e.g., using Q-learning). Crucially, from an algorithmic perspective, model-based learning is more flexible but also more cognitively demanding than model-free learning because it involves re-computing the value function as the estimate of the transitions and rewards is updated. Additionally, it interesting to note that the model-based/model-free learning distinction can be mapped onto the familiar psychological distinction between goal-directed and habitual behavior (Wood & Rünger, 2016), although this is not the only way to formalize the distinction (Dezfouli & Balleine, 2013; Miller, Shenhav, & Ludvig, 2019).

In theory, model-based and model-free learning mechanisms are computationally and conceptually distinct, but in real biological systems these two processes are difficult to completely disentangle (Doll, Simon, & Daw, 2012). Over the last two decades, considerable progress has been made in studying the algorithmic interaction between these different forms of learning and control and their neural basis. For example, the **two-step task** (Gläscher, Daw, Dayan, & O'Doherty, 2010) is a simple MDP consisting of two choice stages and an outcome stage with states whose rewards crucially drift over time. The transitions between states in the choice stages and into the outcome stage states are stochastic but predictable above chance, meaning that if participants learn that an outcome stage has the highest reward, then they can engage in model-based planning to reach that outcome state. However, participants could also simply fall back on single-step value estimates provided by a model-free strategy, which would be initially insensitive to new reward information. In critical trials when new reward information is encountered model-based and model-free learning lead to divergent value updates, thus providing opportunities to distinguish people's algorithmic strategies. Paradigms such as these are often used to study how model-based and model-free learning compete for control of behavior (Gläscher et al., 2010), how they embody different algorithmic and mechanistic tradeoffs (Otto, Gershman, Markman, & Daw, 2013; Daw & Dayan, 2014; Solway & Botvinick, 2015), and how they can interact cooperatively (Kool, Gershman, & Cushman,

2017; Kool, Cushman, & Gershman, 2018).

Distributional RL Standard reinforcement learning algorithms form point-estimates of the value function using state, action, reward samples, but more recent work on **distributional reinforcement learning** explores representing values explicitly as distributions over possible returns (Bellemare, Dabney, & Munos, 2017; Dabney, Rowland, Bellemare, & Munos, 2018; Bellemare, Dabney, & Rowland, 2023). At first blush, it is not obvious why representing a distribution of values would provide any benefits over just representing the expected value—after all, when selecting between actions with different value distributions, we will be computing and comparing expectations. Nonetheless, in practice, value-distributions have been shown to provide a richer target for approximation and thus facilitate representation learning (e.g., with neural networks), mitigate the effects of learning while the policy is changing, and can support a wider variety of downstream behaviors as well as generalization (Bellemare et al., 2017). The success of the distributional approch in deep reinforcement learning has motivated investigation of whether the brain encodes value distributions: Dabney et al. (2020) showed that different dopamine neurons appear to track different levels of value (thus, together, encoding a distribution of values) and thus show a range of positive and negative reward prediction errors during learning. These results enrich the classic picture of how the brain implements scalar reward prediction errors.

Reward design and shaping In the standard reinforcement learning problem, we are given a reward function and must find an optimal policy. But we can also go in the opposite direction: given a desired policy, find a reward function that, when maximized, results in an optimal policy that matches the desired policy. This is known as the problem of **reward design** (Singh, Lewis, & Barto, 2009; Sorg, Singh, & Lewis, 2010) and appears in a number of important settings. One example is **reward shaping**, in which we aim to augment an existing reward function in such a way that the optimal policy is preserved but that would also facilitate faster learning. For example, if we want to incentivize a reinforcement learning agent to reach a goal state, we might not just want to provide a single reward for reaching the goal since that would provide an extremely sparse signal to learn from. Rather, we would want to provide additional shaping rewards for the intermediate steps towards the goal to facilitate faster learning. An important result from reinforcement learning is the **shaping theorem** (Ng, Harada, & Russell, 1999), which provides necessary and sufficient conditions on shaping functions such that they do not change the optimal policy (specifically, that they take the form of so-called "potential functions"). The shaping theorem can be used to design reward functions for people that allow them to achieve a long term goal but receive more intermediate feedback (Lieder, Chen, Krueger, & Griffiths, 2019). However, it has also been found that when in the role of a teacher, people do not simply provide evaluative feedback consistent with the shaping theorem. For example, people will inadvertently incentivize reinforcement learning algorithms to follow **positive reward cycles** where the algorithm systematically deviates from the target behavior in order to receive reward for correcting that deviation, followed by a further deviation and correction (with further reward), potentially looping indefinitely (Ho, Cushman, Littman, & Austerweil, 2019).

Additionally, in the reinforcement learning framework, reward is the driving force behind all adaptation and learning, leading some researchers to propose that maximizing a reward signal is sufficient for explaining all intelligent behavior (Silver, Singh, Precup, & Sutton, 2021). The reward design perspective allows us to pose this thesis as a well defined question: Given some suitably well specified intelligent behavior, does there exist a reward function that produces the target behavior when maximized? Abel et al. (2021) analyzed this question for Markov reward functions in MDPs (rewards defined over s, a, s' tuples) and found that for behaviors defined in terms of sets of policies (a generalization of a single optimal policy) such reward functions can fail to exist. For example, in a gridworld with a state space corresponding to locations in the grid, the behavior "always go in the same direction" cannot be expressed by a Markov reward function. One important takeaway from these results is that it is not always obvious what the expressivity of certain classes of reward functions are with respect to given MDP. Such findings motivate ongoing research on learning and optimizing non-Markov reward functions (Vazquez-Chanlatte, Jha, Tiwari, Ho, & Seshia, 2018; Icarte, Klassen, Valenzano, & McIlraith, 2018).

Representations and reinforcement learning Combining latent state inference with reinforcement learning offers one approach to modeling interactions between learning, decision-making, and representations, but it is not the only one. How an organism encodes states or actions has consequences for other processes, such as exploration or internal decision-making algorithms themselves (Ho, Abel, Griffiths, & Littman, 2019). For example, consider that the distinction between model-based and model-free learning is just as much about representation as it is about algorithms: In model-based learning, the value function is computed using a learned representation of the transition function, whereas in model-free learning, the value function is learned directly, without a separate representation of the transition function (Sutton & Barto, 2018).

One prominent alternative to either pure model-based or model-free learning is the *successor representation* (Dayan, 1993; Gershman, 2018; Russek, Momennejad, Botvinick, Gershman, & Daw, 2017), in which states are encoded based on whether they predict visiting other states when following a policy (e.g., an optimal policy or random policy). The simplest version of the successor representation for a policy π assigns an expected, discounted, cumulative visitation count to each future state s^+ from a current state s. This can be expressed as a set of recursive equations much like those for the value function of a fixed policy (Equation 7.6):

$$M^\pi(s, s^+) = \mathbf{1}[s = s^+] + \gamma \sum_{s'} T(s, \pi(s), s') M^\pi(s', s^+) \tag{7.20}$$

where $\mathbf{1}[s = s^+]$ is an indicator function that plays a similar role as the reward function in the policy evaluation equations—it "counts" each visit to a state s^+ by evaluating to 1 when the current state s is s^+ and 0 otherwise. The similarity between Equations 7.6 and 7.20 has the important consequence that the value of a state can be expressed as a linear combination of the successor representation and state reward function:

$$V^\pi(s) = \sum_{s^+} M^\pi(s, s^+) R^\pi(s^+) \tag{7.21}$$

To see why Equation 7.21 is equivalent to Equation 7.6, we can replace M^π with its definition and do a bit of algebra:

$$\begin{aligned}
V^\pi(s) &= \sum_{s^+} M^\pi(s, s^+) R^\pi(s^+) \\
V^\pi(s) &= \sum_{s^+} \left[\mathbf{1}[s = s^+] + \gamma \sum_{s'} T(s, \pi(s), s') M^\pi(s', s^+) \right] R^\pi(s^+) \\
V^\pi(s) &= \sum_{s^+} \mathbf{1}[s = s^+] R^\pi(s^+) + \gamma \sum_{s^+} \sum_{s'} T(s, \pi(s), s') M^\pi(s', s^+) R^\pi(s^+) \\
V^\pi(s) &= R^\pi(s) + \gamma \sum_{s'} T(s, \pi(s), s') \sum_{s^+} M^\pi(s', s^+) R^\pi(s^+) \\
V^\pi(s) &= R^\pi(s) + \gamma \sum_{s'} T(s, \pi(s), s') V^\pi(s')
\end{aligned}$$

Conveniently, the same algorithms used for estimating value (e.g., dynamic programming or TD prediction) can be used to estimate M^π (Dayan, 1993). And as can be seen from Equation 7.21, the same successor representation M^π can be used to derive the value of the policy π under *any* state reward function by simple multiplication and summation. This property of the successor representation has motivated its use as a model of policy-based transfer learning in neuroscience (Momennejad et al., 2017) as well as in deep reinforcement learning (Barreto et al., 2017). The successor representation also connects to a broader set of ideas based on *predictive maps* in sequential decision-making, which has been used to model grid cells in the hippocampus and other aspects of neural implementations of cognitive maps (Stachenfeld, Botvinick, & Gershman, 2017; Behrens et al., 2018).

Actions can also be represented in different ways. For instance, the same behavior of reaching for a glass of water can be construed at an extremely fine-grained level (e.g., individual muscle contractions) or at a more abstract level (e.g., quenching one's thirst). This intuition motivates **hierarchical reinforcement learning**, in which actions are represented, selected, and learned about at different temporal scales. Although there are a number of formalisms for hierarchical reinforcement learning (Parr & Russell, 1998; Dietterich, 2000), the most widely used is known as the **options framework** (Sutton, Precup, & Singh, 1999). In its simplest form, the options framework distinguishes between actions at two levels of abstraction: ground actions and options. Ground actions are the familiar atomic actions given by the MDP; options assign ground actions to take over multiple ground states before exiting, and so are essentially policies or partial policies in the ground MDP. Formally, given an MDP and a collection of options $\mathcal{O} = \{o_1, o_2, ..., o_n\}$, one can define an **option semi-MDP** that includes options as temporally extended actions (the original ground actions may or may not also be included). Like regular MDPs, semi-MDPs have Bellman equations. For example, the optimal option-level Bellman equation is:

$$V^*_{\mathcal{O}}(s) = \max_{o \in \mathcal{O}} \sum_{s',r,t} T^o(s', r, t \mid s) \left[r + \gamma^t V^*_{\mathcal{O}}(s') \right], \tag{7.22}$$

where $T^o(s', r, t \mid s)$ is the distribution over exit states, cumulative intra-option rewards, and exit times induced by initializing option o at state s. With regards to reinforcement learning in humans and animals, we can ask at least two separate but related questions using the options framework: First, how are options used? Second, how are options discovered? On the question of using options, work on hierarchical reinforcement learning in humans has examined how people learn action values at multiple levels of abstraction (Eckstein & Collins, 2020), how they learn option values via model-free mechanisms (Cushman & Morris, 2015), and how intra-option prediction errors are realized neurally (Botvinick, Niv, & Barto, 2009; Ribas-Fernandes et al., 2011). On the question of discovering options, several proposals have been put forth, including those based on policy compression (Solway et al., 2014), Bayesian inference (Tomov, Yagati, Kumar, Yang, & Gershman, 2020), and resource rationality (Correa, Ho, Callaway, & Griffiths, 2020) (see Chapter 13). Nonetheless, the study of how and why people acquire certain hierarchical action representations—and how to best conceptualize their interaction with subgoals, subtasks, and other forms of abstraction—are currently active areas of research.

Attention and sequential decision-making What is represented and how can also be understood as a consequence of the interaction of decision-making with attentional mechanisms (Radulescu, Niv, & Ballard, 2019). Although there is considerable debate about to what extent attention is a useful construct (James, 1890; Hommel et al., 2019), for our purposes attention can be seen as the process of biasing or filtering information so as to facilitate efficient learning and computation during decision-making. Thus, if pure inference is about inducing patterns "beyond the data", attention involves "reducing the data" to be more managable. In the context of single-stage decision-making, models that combine selective attention with reinforcement learning can explain how learning is modulated and mapped onto anatomical substrates of attention (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Niv, 2019; Niv et al., 2015).

Recent work has also studied the role that cognitive control—a form of top-down or goal-directed attention (Miller & Cohen, 2001; Shenhav et al., 2017)—plays in planning. Recall that the planning algorithms described in Section 7.5.2 all operate on the assumption of a fixed task representation to optimize a policy. For instance, when using heuristic tree search to plan a chess move, one would simulate sequences of moves and counter moves using a model that instantiates the rules for moving the pieces and the conditions for winning the game. However, there are reasons to relax the assumption of a fixed planning model: First, when planning in the real world, there is often no given model, so that the cognitive system must regularly face the challenge of constructing models as required. Second, even in domains with a well-defined ground truth model, such as chess, many details will be irrelevant to planning an immediate action. Finally, classic findings in psychology on problem solving, analogical transfer, and insight suggests that people readily switch between different representations of problems to

solve them (Duncker, 1945; Ohlsson, 2012; Holyoak, 2012). Motivated by these considerations, Ho et al. (2022) propose and test a normative model of **value-guided task construal** that takes into account interactions between constructing a model (formalized as selecting a simplified MDP) and optimizing a policy in that model (e.g., using one of the algorithms in Section 7.5.2). The key idea is to treat model and policy selection as a two level optimization process: An outer loop selects a simplified model (a **construal**) that is used by an inner loop planning algorithm to compute an optimal policy. In its simplest form, the outer loop seeks to optimize the value of representation (VOR) over task construals:

$$\text{VOR}(c) = U(\pi_c) - C(c), \tag{7.23}$$

where $U(\pi_c) = V^{\pi_c}(s_0)$, the value of a policy computed under a construal c when evaluated on the true task from an initial state s_0, and $C(c)$ is a cost associated with the complexity of a construal. Attention comes into play when considering this cost term—specifically, the cost biases selection towards construals that require attending to fewer details (e.g., obstacles in a maze or pieces on a chess board). Across a series of experiments, Ho et al. found support for the idea that people flexibly form task representations consistent with this account. An important direction for future research is to explore the algorithms that enable people to form simplified representations as well as to understand the interaction of model construction with other cognitive mechanisms.

7.6 Active learning

So far, we have discussed cases in which the utility of the consequences of our actions is directly of interest. But in many areas of cognition, consequences may not be of primarily interest in themselves, but rather because they provide further information. This is the domain of *active learning*, where our actions are, at least in part, not in the service of an externally defined goal, but are driven by the objective of gathering relevant information as efficiently as possible. In scrutinizing a map, a page of a book, or a face, our eyes do not wander aimlessly but search for features of particular relevance or interest. So, our eyes will selectively sample the major towns, roads or harbors rather than uniform areas of sea; they will jump to the paragraphs that seem likely to contain new and interesting information; and alight on those facial features which are mostly likely to give away identity or emotional expression. Similarly, in deciding who to talk to, what to type into a search engine, what to read, we are often foraging for information, rather than attempt to achieve any concrete external goal. Indeed, large amounts of human activity, especially in the fields of education (when we are learning, say, history or science) and culture (going to movies, reading novels, listening to music), involve the acquisition and processing of information which is not in the service of any immediate task or goal. Our attention is limited and must be deployed wisely in a complex world with many attractions and distractions. Sometimes, of course, our focus is much narrower—we have a specific decision to make or course of action to pursue, and we want to gather information that will help us solve the challenge of the moment.

In any case, it should be clear that the process of choosing which information to gather (and, similarly, which information to pay attention to and which to ignore, once that information is gathered) is of central importance in the operation of just about every aspect of cognition. There is a continual cycle, in which our current state of knowledge directs our senses and our attention to actively gather new information; and this new information is then used to update our knowledge state; and we then search for yet more information guided by this update knowledge state, and so on. We are relentlessly active learners about our world, searching for useful and interesting information, rather than simply passively taking account of whatever data happens to float into view.

To take a mundane example, suppose we have lost our keys. We do not simply wait for useful evidence concerning their location to turn up. We active search for useful clues. We taps our pockets, peer into bags, and look under sofas, hoping to gather sensory information which will give us a, hopefully

decisive, lead. And in gathering clues at a crime scene, or designing scientific experiments we are, equally, attempting to choose sets of actions likely to lead to data that are diagnostic between the alternative theories under considerations (e.g., Lindley, 1956; Platt, 1964). Instead, we are actively attempt to find information that is useful, or generally interesting, as possible. Now what counts as useful or interesting will vary depending on our objectives (finding our keys, catching a criminal, pinning down the best scientfic theory). And in the absence of a specific goal, we find still information interesting, and other information dull. Indeed, much of leisure time is spent searching engaged in searching for and consuming information without any obvious immediate relevance to our life decisions (from watching movies, sports, reading history and fiction, listening to music and so on). But while the general question of what makes information interesting is hard and open-ended problem (Chater & Loewenstein, 2016), the same principle of active learning is at work. Our minds as searching for, and attending to, the interesting, and attempting to avoid information that is dull or useless.

There may seem something slightly paradoxical about the very idea that it is possible to actively choose which data we wish to receive. After all, before we have moved our eyes to a new location, or conduct of the scientific experiment, we do not know what data we will receive—otherwise the act of data gathering is entirely redundant. But if we do not know what data we will receive, how can we assess its potential value?

The answer, as so often in the Bayesian approach, stems from use of prior knowledge. So, before moving our eyes, or conducting the experiment, we can consider our prior probability distribution of possible sets of data (where *prior*, here, simply means *prior to the data gathering act being contemplated*. Suppose that the agent is able to assign a *value* to each possible data outcome; then the informational value of action which may result in such data can simply be defined as the expected value, where the expectation is relative to the prior distribution over the data. So, for example, our prior knowledge of the human face, alongside current low fidelity information the visual periphery, may be enough to narrow down certain locations in visual space as likely to be much more interesting than others, and hence as much more appropriate targets for eye movements. Thus, for example, in scanning images it is possible for eye movements to jump between "informative" elements, such as eyes and mouths; and to pay much less attention to patches of cheek or forehead, or background wall.

Suppose that our Bayesian agent has a range of possible actions $a \in \mathcal{A}$. In the light of each action, each state of the world $s \in \mathcal{S}$ will, and in the light of the current state of knowledge, $\mathcal{K}$, of the agent, have probability $P(s|a, \mathcal{K})$ (where K represents background knowledge), and utility (in contexts where this is well-defined) $U(s)$. The expected utility, given knowledge $\mathcal{K}$ and action a is written:

$$EU(a, \mathcal{K}) = \sum_{s \in \mathcal{S}} P(s|a, \mathcal{K})U(s) \tag{7.24}$$

The rational Bayesian agent will, of course, choose the action which achieves the maximum expected utility: that is, in the light of knowledge $\mathcal{K}$, it will choose an action $a^*(\mathcal{K})$, such that $EU(a^*, \mathcal{K}) \geq EU(a, \mathcal{K})$ for all a. For simplicity we assume $a^*(\mathcal{K})$ is the unique action that attains expected utility $EU(a^*, \mathcal{K})$. Now consider that the agent has the possibility of actively finding out some information, e.g., actively carrying out some observation or experiment, yielding some data $d \in \mathcal{D}$. These new data will be added to the agent's background knowledge, $\mathcal{K}$, creating the new background knowledge $\mathcal{K} \cup d$, and will lead to the agent to choose a new $a^*(\mathcal{K} \cup d)$ in the light of updated probabilities $P(s|a, \mathcal{K} \cup d)$. How much utility, $U(d)$, should the agent attach to acquiring this data? The agent must evaluate this quantity in the light of its updated probabilities $P(s|a, \mathcal{K} \cup d)$. Specifically, the key question is how much greater is its expected utility, in the light of these probabilities, if it chooses $a^*(\mathcal{K} \cup d)$ over and above its revised expected utility given its original choice of action, $a^*(\mathcal{K})$. The utility to the agent of d is therefore:

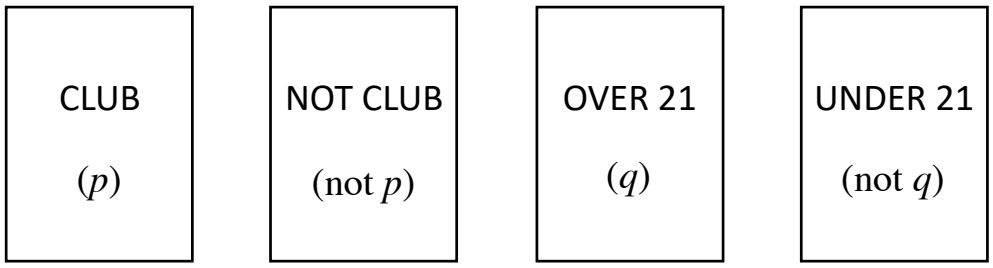
$$U(d) = \sum_{s \in \mathcal{S}} P(s|a^*(\mathcal{K} \cup d), \mathcal{K} \cup d)U(s) - \sum_{s \in \mathcal{S}} P(s|a^*(\mathcal{K}), \mathcal{K} \cup d)U(s) \tag{7.25}$$

Intuitively, we can reason as follows. Really useful data d changes our model of the world so that so that we switch from what previously looked like our best choice of action $A^*(\mathcal{K})$ to a new action $A^*(\mathcal{K} \cup d)$ which, according to our updated probabilities, has a much higher expected value. So, for example, data that tells us where to look for treasure, or that the fungus we were about to eat is actually a toadstool is valuable. Conversely, data that doesn't change our action (e.g., about denomination of the treasure or the precise weight of the fungus) is from this narrow viewpoint, entirely valueless.[15] Note that here, and in a wide variety of contexts, the *expected* value of acquiring new information can only be positive or zero (where the expectation comes from the subjective probabilities of the decision-making agent). That is, if a rational agent asks a question, that agent cannot consistently believe that its choices of actions in the light of the answer will, in expectation, be worse than if had received no answer (it will only change course of action in the light of new information if it believes the new action has at least as high an expected utility as its previously entertained course of action). This means that, from the point of view of a fully rational agent, in expectation, acquiring new information cannot be harmful. Notice though that from an external perspective, an agent learning new information can both be harmful and harmful in expectation. For example, if in a search for buried treasure, I have by pure chance decided to start digging in precisely the right spot, then the expected value (from an outside perspective) of further information (e.g., from old treasure maps, or historical reseach) may well be negative, because it can only lead me away from my current (lucky) guess.

We have so far been considering the utility of a piece of data. But, of course, the Bayesian agent does not know which data they will encounter *a priori*. Hence, the expected value of the observation or experiment, d, is the expectation of $U(d)$, in the light of knowledge $\mathcal{K}$ before the experiment has been carried out:

$$EU(d) = \sum_{d \in \mathcal{D}} P(d|\mathcal{K})U(d). \tag{7.26}$$

To get an intuition for how this works, consider a variant of a well-known psychology of reasoning task (Wason, 1966, 1968) in which people must actively select data in the light of a rule, such as *if a person is in the club, they must be at least 21 years old*, which has the form: if p then q. In the experimental task, the participant is given four cards, each with an age on one side, and whether they have entered the club or not on the other side. But we can only see the cards face up—the task is to say which cards would we like to turn over.



The answer to the question of which cards to turn over (i.e., which information to search for), depends of course on our utilities. These are often only vaguely specified in everyday life, and even in many experimental tasks. But these utilities will clearly depend on our objectives. Suppose, for example, that we are a police offer, checking for violations of the rule—and suppose we get high utility for finding law-breaking.

[15] The story can be more complex in a variety of ways. For example, new data may cause us to change of utilities, $U(s)$ as well as our probabilities; and sequential decision-making may be crucial. I may find that apparently valueless information about the pattern on a fungus becomes crucially valuable in the light of yet further information—e.g., the information contained in a field mushroom guidebook. We will ignore such complications here for simplicity.

Then we can clearly ignore people who are not at the club (we don't turn the not-p card); or who are over 21 (we don't turn the q card). But we do want actively to learn about people who are in the club—the expected payoff for doing so depends on our prior probability, in the light of our background knowledge, that they may be under-21 (and, of course, the utility we derive from finding any such rule-breakers). And we want to check the under-21's (the not-q card, in case they happen to be in the club). In most realistic scenarios, the expected payoff for turning this card is rather low, though—after all, there is vast numbers of people who are under 21, and the chance that one of them happens to be in the club is low. So the turning over the p will have the greatest expected utility, with some lesser expected utility for the *q-card*, and zero expected utility for the others (indeed, if we take account of the "effort" of making the enquiry, these options will have a negative expected utility, and hence won't be chosen). This fits with the experimental data (Cheng & Holyoak, 1985; Cosmides, 1989).[16]

But to see how utility is crucial, suppose instead we are a representative from the students union, checking that people who are over 21 (q card) are not being unfairly turned away (*not-p* card). In this role, we obtain utility not from finding violations of the rule (the p, not-q cases), but from finding exclusions cases not justified by the rule (not-p, q cases). To find such cases involves turning over (only) the not-p and emphq cards. So which cards we choose—i.e., which information we actively choose to investigate—depends not just on the rule, but on our goals; and these changes in card selections depending on the framing of the task are observed experimentally (Gigerenzer & Hug, 1992). Notice that these shifts are not predicted if, as in early discussions of the selection task, the problem of data selection is viewed as purely a matter of "logic," independent of the utilities of the decision maker.

While this style of analysis may be applicable when we are searching for information to achieve a specific goal (e.g., finding our lost keys, or detecting violators of a rule), much active learning has a more disinterested character. The goal of the agent is to find out the state of the world, independent of any immediate implications for action. Indeed, most aspects of perception and cognition seem to fall into this category. Whether browsing a newspaper, learning a language, or conducting a scientific investigation, we often have little if any sense of whether, or in what way, the information we are gathering might modify our actions. In such cases, a different approach, based on information theory can be applied (Lindley, 1956). In this set-up, there are no actions, merely mutually exclusive and exhaustive states of the world $s \in \mathcal{S}$, each with probability $P(s)$ (these states might, for example, correspond to alternative categories, scientific hypotheses, or suspects in a murder mystery; note that we drop conditioning on background knowledge K, for notational convenience). The amount of uncertainty we have about which state the world is in measured by Shannon's entropy:

$$H(s) = \sum_{s \in \mathcal{S}} P(s)) \log 1/P(s) \tag{7.27}$$

In the light of data, d, the so-called conditional entropy is now:

$$H(s|d) = \sum_{s \in \mathcal{S}} P(s|d) \log 1/P(s|d) \tag{7.28}$$

The information gain $IG(s, d)$, the reduction in entropy over s on learning data d, can be written:

$$IG(s, d) = H(s) - H(s|d) \tag{7.29}$$

information gain can be negative or positive. For example, we might be almost certain about some conclusion, and then learn information that increases are doubt. In this case, entropy will increase. But following the discussion above, from the point of view of the agent's own subjective probabilities, the *expectation* of information gain can only be positive or zero (there is something very odd about planning to

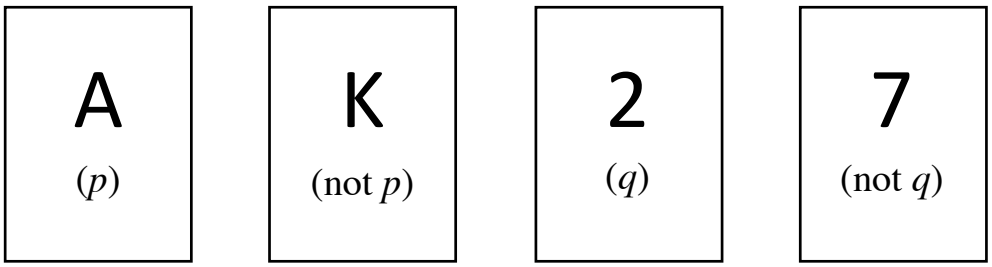
[16]The Bayesian approach differs, though, from some other theoretical viewpoints. For example, (Cosmides, 1989) postulates innate special-purpose "cheater-detection module" to explain this type of experimental data.

make an observation which one believes, on average, will leave one more ignorant than before). And this intuition is captured by the information-theoretic approach: the expected information gain, $EIG(s, d)$ averaged over possible d is therefore:

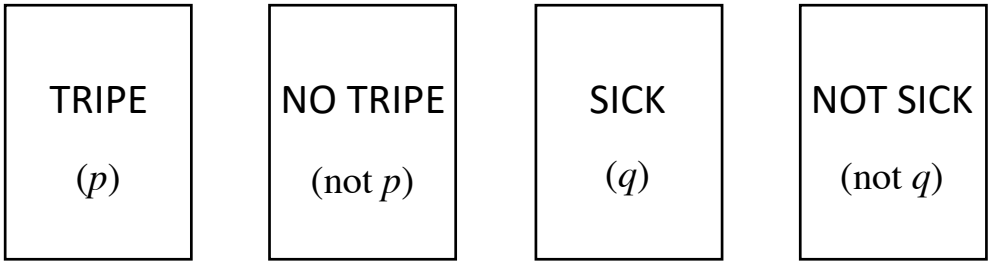
$$EIG(s, d) = \sum_{d \in \mathcal{D}} P(d)[H(s) - H(s|d)] \geq 0 \tag{7.30}$$

The inequality implies that, on average, any new observation or experiment is expected to yield positive or at worst zero information (and this follows from elementary information theory(Cover & Thomas, 1991)). Expected information gain, and closely related notions, have been used as a measure of the goodness of experiments, as a model of active learning in neural networks (Mackay, 1992), and to model cognitive phenomena ranging from how eye movements are directed during reading (e.g., Legge, Klitz, & Tjan, 1997).

Indeed, this approach has also been applied to a different variation of the four card selection task (Wason, 1966, 1968). Suppose that we consider an abstract rule (not one considering clubs, age-restrictions, or any other real-world context), such as *If a card has an A on one side, it has a 2 on the other*. And we are presented with the four cards:



The participant has no particular utilities associated with the rule (they aren't searching for rule-violators, or cases where people have been treated in a way that the rule does not justify). Instead, the task is simply a matter of gathering information concerning whether the rule is, or is not, true. In information-theoretic terms, then, we suppose we begin with some prior assumption (perhaps complete ignorance concerning the truth of the rule), and we wish to turn over the cards which are most likely to reduce our uncertainty, in expectation, as much as possible (Oaksford & Chater, 1994). There are, of course, many ways in which the details of a model of this set-up can be constructed. But to get an intuition about the inferences people might make, consider a real-world case. Suppose we are wondering, for example, whether eating tripe make people sick.



Intuitively, it is clear we should turn over the TRIPE (p) card—it will be very informative to discover whether the person was sick or not. It will also be natural to query the people who were sick (the q-card).

There are lots of ways of becoming sick—but it might turn out that they were recent tripe-eaters, which would give evidence in favor of the hypothesis. There is also a rather remote possibility of getting useful information from checking people who are not sick (the not-q-card), just in case they happen to have eaten tripe—because this would be evidence against the rule. But, as tripe-eating is so unusual, the probability of this outcome is very low, and most likely we will merely sample a healthy non-tripe eater, which will tell us little or nothing. So the tendency to actively investigate the cards should have the order $p > q >$ not-$q >$ not-p, which is observed empirically (Oaksford & Chater, 1994) (though see (Oberauer, Wilhelm IV, & Diaz, 1999), which finds that directly manipulating rarity can sometimes have at best weak impacts on card choices).

This analysis in terms of expected amount of information gained provides a rational analysis of active data selection in this task–which is particularly intriguing, as it has often been argue that the "logically" correct response to the task is purely to search for falsification of a rule (i.e., thus turning over just the p and not-q cards) and that turning over the q card at all is simply a mistake—a viewpoint which seems to fit with Popper's falsificationist philosophy of science (Popper, 1935/1990), rather than a Bayesian perspective on scientific inference(Howson & Urbach, 1993). The Bayesian active learning framework can also capture a lot of variations of the task, and the direction in which changing the probabilities of events p and q modifies card selection frequencies (Oaksford & Chater, 2003). But people's data selection is not perfectly calibrated with these probabilities—people seem to choose data as if, as is true of the vast majority of real-world rules, p and q are rare, by default, even when this does not hold in a specific experimental context. More generally, this viewpoint helps explain why people frequently adopt a **positive test strategy**(Navarro & Perfors, 2011), searching for instances that confirm a hypothesis of interest, in cases where searching for counterexamples has only a small chance of uncovering relevant evidence. Thus, at least in many contexts, the tendency to search for positive instances is not an example of **confirmation bias** but has a rational basis(Klayman & Ha, 1987).

Notice, though, that actively selecting information to gather as much information as possible is defined with respect to a specific set of hypotheses which we wish to test (e.g., whether a specific rule does, or does not, hold). But often, as noted above, our aims are much more open-ended—we may sometimes scan a newspaper to see how a particular event turned out, but often we are just wondering if anything 'interesting' has happened. Similarly, in science we are sometimes attempting to design an experiment to test between several specific hypotheses; but often our enquiries are far more exploratory. The question of how best to capture active learning in these open-ended contexts is important and unresolved—we only have the beginnings of a theory of what makes information 'interesting' (Chater & Loewenstein, 2016).

We have focused here on the problem of determining which data to sample. But at least as important is the parallel question: which computations to carry out, once the data has been sampled. Given presumably severe computational limits on the brain, one of the cognitive system's most important tasks is to carefully direct its computational resources. As with the problem of choosing which information to sample, this very idea has a slightly paradoxical flavor: how is it possible to determine how useful the results of a computation is likely to be, before we have carried it out? And, as before, the key is to be able to deploy prior information, to determine which computations are likely to be useful, and which are not. We explore this question in detail in Chapter 13, where we consider how the rational use of limited computational resources might explain some of the ways that human behavior deviates from Bayesian decision theory.

7.7 Forward and inverse models

We have seen that minimal consistency conditions on an agent yield the result that the agent can be described as behaving "as if" it maximizes expected utility, according to some set of probabilities and utilities (which can be inferred from the pattern of decisions that the agent makes). In cognitive science,

though, a key goal is to understand the computational processes that support inference and decision-making; hence Bayesian decision theory can be viewed not as a black box model, but as a specification of the processes that may underpin decision-making, whether in making simple psychophysics judgments or developing complex sequential policies for interacting with the world. Moreover, neuroscience has begun to provide preliminary evidence which suggests that Bayesian computations may indeed be instantiated in the brain. In this chapter, we have focussed framed our discussion in terms of "high level" aspects of decision making, involving conscious deliberation and choice. But note that Bayesian decision theory applies equally to the analysis of actions over which we exert little or no conscious control: e.g., basic perceptuo-motor processes such as picking up objects, maintaining our balance, or catching a ball. Such tasks involving integrating of large amounts of perceptual information and generating plans to control a highly complex motor system. The very fluency of perceptuo-motor control suggests that it may be well-approximated by an optimal Bayesian model and indeed, this approach has proved to be very productive (e.g., Kording & Wolpert, 2006)

One key aspect of successful motor control is building a so-called **forward model**, which maps from motor commands to sensory experiences generated by implementing those commands. Such a model is crucial, for example, in distinguishing the sensory consequences of one's own action (e.g., moving one's head or eyes) from the sensory consequences from change in the external world (e.g., that the room is being shaken by an earthquake). Bayesian inference can then invert the forward model, to create an inverse model, which infers which motor commands will lead either to desired sensory consequences (e.g., in planning an action) or which motor commands did lead to an observed action (e.g., in the interpretation of actions by others).

This logic can be applied not just to one's own actions, but in the interpretation of other people's behavior. Bayesian decision theory provides a forward model, mapping beliefs and desires into actions. But it is also interesting to consider the inverse problem: mapping from observers actions to inferred beliefs and desires. This process of inversion is, of course, the paradigm of Bayesian inference; but it is also of great importance in cognition. The ability to infer people's underlying mental states from their behaviour (including, of course, their linguistic behaviour) appears to be central to human communication (Baker, Saxe, & Tenenbaum, 2009; Grice, 1957), as well, perhaps as being important implicated in empathy, altruism, and coordinated social behaviour (Baron-Cohen, 1997; Houlihan, Kleiman-Weiner, Hewitt, Tenenbaum, & Saxe, 2023). Indeed, a long tradition of research in social psychology suggests that, to some degree, such inverse modelling is involved in the interpretation of our own actions, to infer our own beliefs and desires (Bem, 1972; Nisbett & Wilson, 1977). We take up the problems of **inverse decision-making** and **inverse planning** in Chapter 14; and will draw out links with rational theories of communication in Chapter 16.

7.8 The limits of reason

The premise of this chapter is that the maximization of expected utility or similar quantities can provide the basis for models of decision-making across a range of domains, from animal foraging, to motor control, learning, and high-level decision-making. This viewpoint may appear to clash with the research traditions in the fields of judgment and decision-making and behavioral economics, which appear to show that people frequently and systematically deviate from Bayesian decision theory — and indeed the basic consistency assumptions which are the foundation of the economic approach to decision-making are routinely and systematically violated (e.g., Kahneman & Tversky, 1984).

Some theorists have argued that the departures from rationality are so widespread that the Bayesian perspective on decision-making, and the rational analysis of behaviour more broadly, may be a theoretical blind alley; instead, it has been argue that collections of heuristics or layers of input-output rules might better explain behavior (e.g., Brooks, 1991; Gigerenzer & Todd, 1999; McFarland & Bösser, 1993).

We take the opposite view: that to abandon a rational theory of decision-making is to render human behavior utterly mysterious: indeed, it is to lose the distinction between *behavior* (e.g., picking up a cup; waving to a friend, or typing a message) and mere *movement* (e.g., falling over, having a reflex triggered by the doctor's hammer, or inadvertly leading on the computer keyboard). Bayesian decision theory helps explain behavior as purposeful activity: our actions are aligned with our preferences and our beliefs. Thus, we pick up the coffee because we believe it to contain coffee; and we want to drink it. The fine details of our motor actions can also be explained in the same terms: our aim of picking up the cup smoothly and efficiently, and to move it without spilling its contents, will help explain the specific way in which we move. More broadly, the Bayesian viewpoint explains how beliefs, preferences and actions can be linked together, across many scales (from individual movements, to actions, to momentary plans, to the direction of our entire life), in a way that is as coherent as possible. If we attempt, by contrast, to see behavior nothing more than a set of reflexes or a toolbox of special-purpose heuristics, it is difficult to understand the origin of the coherence of human behavior (e.g., Bratman, 1987).[17]

Throughout this book, our use of Bayesian modeling is as a guide to what the ideal solutions to specific inductive problems encountered by humans look like, which can in turn be used as a tool for making sense of human behavior. It would be unrealistic to expect that the approach will always quantitatively model the precise details of human decision-making. We suggest that the Bayesian approach is likely to be particularly effective in domains for which human performance has been shaped by powerful forces of natural selection and learning—e.g., motor control, sequencing actions, planning, common-sense reasoning and so on. It is likely to be much less well-adapted to solving numerical and verbally stated decision problems (e.g., concerning choices between gambles), with which we are unfamiliar.[18] Moreover, the brain cannot follow Bayesian decision theory exactly—in all but the simplest contexts, exact Bayesian calculations are computationally intractable, and can only be approximated, for example, by sampling methods (Chapter 6; Chater et al., 2020; Sanborn & Chater, 2016; Vul et al., 2014). But to understand the purposeful nature of human behavior remains inexplicable, it is essential that we understand human actions as approximating, a rational model, rather than being entirely unconstrained. That is, intelligent decisions can be the product of limited reasoning, but not of no reasoning whatever.

7.9 Summary

The focus of this book is the problem of induction: how is it possible to know the structure of the world from partial and noisy data? But any such learning is useless, from the point of the survival and reproduction of an organism, without the ability to translate knowledge into action: i.e., to combine knowledge with our values to *decide* what to do. Bayesian decision theory provides a solution to this problem, indicating how rational agents should act upon their beliefs. Even simple decisions can involve complex underlying processes of evidence accumulation, and these complexities are magnified when we consider sequences of interdependent decisions. Nonetheless, cognitive scientists have made substantial progress in identifying the mathematical principles underlying human decision-making, building upon and complementing the more general ideas of probabilistic modeling outlined in the previous chapters. As we now transition to the second part of the book, considering more elaborate models and more detailed applications to human cognition, the principles of Bayesian decision theory provide a foundation for linking belief and action.

[17] Of course, our thoughts and behaviors are by no means entirely coherent—but cognition appears to be directed at eliminating incoherence, where possible (e.g., Festinger, 1957; Thagard, 2002).

[18] We note, though, that understanding how people behave in these unfamiliar contexts, where their behavior may depart substantially from Bayesian decision theory, may be of great practical importance, in for example, understanding the tortuous choice processes people might use to select mortgages, pensions or courses of medical treatment (Newell, Lagnado, & Shanks, 2022).

References

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., & Singh, S. (2021). On the expressivity of markov reward. In (pp. 7799–7812).

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*(4), 503–546.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind.* MIT Press.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., Hasselt, H. P. van, & Silver, D. (2017). Successor features for transfer in reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30.*

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In *Proceedings of the International Conference on Machine Learning* (pp. 449–458).

Bellemare, M. G., Dabney, W., & Rowland, M. (2023). *Distributional reinforcement learning.* MIT Press.

Bellman, R. (1957). *Dynamic programming.* Princeton University Press.

Bem, D. J. (1972). Self-perception theory. In *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Elsevier.

Berg, R. van den, Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis.* Springer.

Binmore, K. (2008). *Rational decisions.* Princeton University Press.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

Bradley, R. (2017). *Decision theory with a human face.* Cambridge University Press.

Bratman, M. (1987). *Intention, plans, and practical reason.* University of Chicago Press.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1-3), 139–159.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Bush, R., & Mosteller, F. (1955). *Stochastic models of learning.* Wiley.

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, *126*, 137–154.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, *29*(5), 506–512.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Cooter, R., & Rappoport, P. (1984). Were the ordinalists wrong about welfare economics? *Journal of Economic literature*, *22*(2), 507–530.

Correa, C. G., Ho, M. K., Callaway, F., & Griffiths, T. L. (2020). Resource-rational task decomposition to minimize planning costs. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.*

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the Wason selection task. *Cognition*, *31*(3), 187–276.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* Wiley.

Crupi, V., Chater, N., & Tentori, K. (2013). New axioms for probability and likelihood ratio measures. *The British Journal for the Philosophy of Science*, *64*(1), 189-204.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*(45), 13817–13822.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675.

Dabney, W., Rowland, M., Bellemare, M., & Munos, R. (2018). Distributional reinforcement learning with quantile regression. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.*

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655).

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dawkins, R. (1978). *The selfish gene.* Oxford University Press.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

De Lafuente, V., Jazayeri, M., & Shadlen, M. N. (2015). Representation of accumulating evidence for a decision in two parietal areas. *Journal of Neuroscience*, *35*(10), 4306–4318.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*(12), e1003364.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*, 227–303.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Duncker, K. (1945). On problem-solving. *Psychological monographs*, *58*(5).

Eckstein, M. K., & Collins, A. G. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, *117*(47), 29381–29389.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666.

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, *38*(33), 7193–7200.

Ghallab, M., Nau, D., & Traverso, P. (2016). *Automated planning and acting.* Cambridge University Press.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* Oxford University Press, USA.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(supplement_3), 15647–15654.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions and reward. *Neuron*, *36*, 299-308.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574.

Good, I. J. (1950). *Probability and the weighing of evidence.* C. Griffin.

Good, I. J. (1979). A. M. Turing's statistical work in World War II. *Biometrika*, *66*, 393-396.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, *274*(5286), 427–430.

Hansen, E. A., & Zilberstein, S. (2001). LAO*: A heuristic search algorithm that finds solutions with loops. *Artificial Intelligence*, *129*(1-2), 35–62.

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, *4*(2), 100–107.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Ho, M. K., Abel, D., Griffiths, T. L., & Littman, M. L. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111-116.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520-549.

Holyoak, K. J. (2012). The oxford handbook of thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), (p. 234-259). Oxford University Press.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288–2303.

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220047.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach.* Open Court.

Icarte, R. T., Klassen, T., Valenzano, R., & McIlraith, S. (2018). Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning* (pp. 2107–2116).

James, W. (1890). *Principles of psychology.* Holt.

Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341-350.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211-228.

Knight, F. H. (1921). *Risk, uncertainty and profit.* Hart, Schaffner, & Marx.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. In *European conference on machine learning* (pp. 282–293).

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making* (pp. 153–178). Academic Press.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321–1333.

Kording, K., & Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kreps, D. M. (1988). *Notes on the theory of choice.* Westview Press.

Kreps, D. M. (1990). *Game theory and economic modelling.* Oxford University Press.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*(2), 227-260.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review*, *104*(3), 524-553.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Lieder, F., Chen, O. X., Krueger, P. M., & Griffiths, T. L. (2019). Cognitive prostheses for goal achievement. *Nature human behaviour*, *3*(10), 1096–1106.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, *107*(2), 573–597.

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models.* California Institute of Technology.

McFarland, D., & Bösser, T. (1993). *Intelligent behavior in animals and robots.* MIT Press.

McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems*, *2*, 339–364.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167-202.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292-311.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166-180.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120-134.

Neumann, J. von, & Morgenstern, O. (1944). *The theory of games and economic behavior.* Princeton University Press.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2022). *Straight choices: The psychology of decision making.* Psychology Press.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (mt). *Journal of Neuroscience*, *8*(6), 2201–2211.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th Annual International Conference on Machine Learning* (pp. 278–287).

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608-631.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, *10*(2), 289–318.

Oberauer, K., Wilhelm IV, O., & Diaz, R. R. (1999). Bayesian rationality for the wason selection task? a test of optimal data selection theory. *Thinking & Reasoning*, *5*(2), 115–144.

Ohlsson, S. (2012). The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. *The Journal of Problem Solving*, *5*(1), 7.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751–761.

Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems 10.*

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *science*, *146*(3642), 347–353.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864-901.

Popper, K. R. (1935/1990). *The logic of scientific discovery.* Unwin Hyman.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming.* Wiley.

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: the role of structure and attention. *Trends in Cognitive Sciences*, *23*(4), 278–292.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (p. 64-99). Appleton-Century-Crofts.

Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370 - 379.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*(9), e1005768.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Prentice Hall.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Savage, L. J. (1972). *The foundations of statistics (2nd edition).* Courier Corporation.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535.

Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2601–2606).

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology*, *10*(8), e1003779.

Sorg, J., Singh, S. P., & Lewis, R. L. (2010). Internal rewards mitigate agent boundedness. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 1007–1014).

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653.

Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Meeting of the Cognitive Science Society* (pp. 355–378).

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

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211.

Thagard, P. (2002). *Coherence in thought and action.* MIT Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, *16*(4).

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A*, *20*(7), 1419–1433.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550-592.

Vazquez-Chanlatte, M., Jha, S., Tiwari, A., Ho, M. K., & Seshia, S. (2018). Learning task specifications from demonstrations. In *Advances in Neural Information Processing Systems 31.*

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101-118.

Wald, A. (1947). *Sequential analysis.* Wiley.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology.* Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, *67*(1), 289–314.