# 1

# Thinking and Deciding

*Life is the art of drawing sufficient conclusions from insufficient premises.*

—Samuel Butler

## 1.1 Decision Making Is a Skill

Humans today evolved from ancestors hundreds of thousands of years ago who lived in small groups and spent most of their waking hours foraging for sustenance. When we weren't searching for something to eat or drink, we were looking for safe places to live, choosing mates, and protecting our offspring. Our success in accomplishing these "survival tasks" arose not due to distinctively acute senses or especially powerful physical capacities. We dominate this planet today because of our distinctive capacity for good decision making. This same skill has allowed us to leave the planet, for brief periods; but, of course, the skill has allowed us to develop technologies and weapons that could render the planet uninhabitable if we make a few really bad decisions. Human beings have an exceptional ability to choose appropriate means to achieve their ends.

This book is about decision making, but it is not about *what* to choose; rather, it is about *how* we choose. Most of the conclusions in this book follow from research conducted by psychologists, economists, and biologists about how people actually make choices and decisions—people ranging from medical and financial experts to college student participants in psychological

experiments. The important finding is that diverse people in very different situations often think about their decisions in the same way. We have a common set of cognitive skills that are reflected in similar decision habits. But we also bring with us a common set of limitations on our thinking skills that can make our choices far from optimal, limitations that are most obvious when we must make judgments and decisions that are not like those we were "selected" to make in the ancestral environments in which we evolved.

Our decision-making capacities are not simply "wired in," following some evolutionary design. Choosing wisely is a learned *skill,* which, like any other skill, can be improved with experience. An analogy can be drawn with swimming. When most of us enter the water for the first time, we do so with a set of muscular skills that we use to keep ourselves from drowning. We also have one important bias: We want to keep our heads above water. That bias leads us to assume a vertical position, which is one of the few possible ways to drown. Even if we know better, in moments of panic or confusion we attempt to keep our heads wholly free of the water, despite the obvious effort involved compared with that of lying flat in a "jellyfish float." The first step in helping people learn to swim, therefore, is to make them feel comfortable with their head under water. Anybody who has managed to overcome the head-up bias can survive for hours by simply lying face forward on the water with arms and legs dangling—and lifting the head only when it is necessary to breathe (provided, of course, the waves are not too strong or the water too cold). Ordinary skills can thus be modified to cope effectively with the situation by removing a pernicious bias.

This book describes and explains these self-defeating thinking habits, and then suggests other strategies that will improve the decision maker's skill. This approach reflects the spirit of Benjamin Franklin, whose letter of advice about a pressing decision to his friend Joseph Priestley (1772) began, "I cannot, for want of sufficient premises, advise you *what* to determine, but if you please, I will tell you *how.*" We will describe pernicious modes of thought in order to provide advice about how to improve choices. But we will not suggest what your goals, preferences, or aspirations ought to be when making these choices. The purpose of this book is not to improve tastes, or preferences, or ethics—nor to provide advice about how to implement decisions once they have been made. Likewise (unlike many other books written on this subject), this book does not offer advice about how to feel good about yourself. Rather, our purpose is to increase skill in thinking about decisions and choices. In addition, to better understand the decision process and to identify the situations in which our choices are less than optimal, we introduce a second perspective on decision making, namely analyses of the nature of rational decision processes by philosophers and mathematicians.

## 1.2 Thinking: Automatic and Controlled

What is thinking? Briefly, it is the creation of mental representations of what *is not* in the immediate environment. Seeing a green wall is not thinking; however, imagining what that wall would be like if it were repainted blue is. Noting that a patient is jaundiced is not thinking; hypothesizing that the patient may suffer from liver damage is. Noticing that a stock's price has dropped is not thinking, but inferring the causes of that drop and deciding to sell the stock is.

Sir Frederick Bartlett, whose work 50 years ago helped create much of what is now termed *cognitive psychology,* defined thinking as the *skill* of "filling gaps in evidence" (1958). Thinking is probably best conceived of as an *extension of perception*—an extension that allows us to fill in the gaps in the picture of the environment painted in our minds by our perceptual systems, and to infer causal relationships and other important "affordances" of those environments. (For example, Steven Pinker [1997] provides an instructive analysis of the assumptions that we *must* be using as "premises" to "infer" a mental model of our three-dimensional world based on our fragmentary two-dimensional visual percepts.)

To simplify, there are basically two types of thought processes: automatic and controlled. The terms themselves imply the difference. Pure association is the simplest type of automatic thinking. Something in the environment "brings an idea to mind," or one idea suggests another, or a memory. As the English philosopher John Locke (1632–1706) pointed out, much of our thinking is associational. At the other extreme is controlled thought, in which we deliberately hypothesize a class of objects or experiences and then view our experiences in terms of these hypothetical possibilities. Controlled thought is "what if" thinking. The French psychologist Jean Piaget (1896–1980) defined such thinking as "formal," in which "reality is viewed as secondary to possibility." Such formal thought is only one type of controlled thinking. Other types include visual imagination, creation, and scenario building.

To distinguish between these two broad categories of thinking, we can give an example. Many of our clinical colleagues who practice psychotherapy are convinced that *all* instances of child abuse, no matter how far in the distant past and no matter how safe the child is at the time of disclosure, should be reported, "because one thing we know about child abuse is that no child abusers stop on their own." How do they know that? They may have treated a number of child abusers, and of course none of those they have seen have stopped on their own. (Otherwise, our colleagues wouldn't be seeing them.) The image of what a child abuser is like is automatically

associated with the abusers they have seen. These known abusers did not "stop on their own," so they conclude that all child abusers do not. The conclusion is automatic.

These colleagues do in fact have experience with abusers. The problem is that their experience is limited to those who have *not* stopped on their own, and since their experience is in treatment settings, these abusers cannot *by definition* stop without therapy. Abusers who have stopped on their own without therapy do not enter it and would be unlikely to identify themselves. They are systematically "*un*available." Or consider clinical psychologists and psychiatrists in private practice who maintain that low self-esteem "causes" negative social and individual behavior. But they see only people who are in therapy. People who engage in negative behaviors and don't feel bad about such behaviors don't voluntarily seek out therapists. (And therapists in coercive settings, such as residential treatment programs for severe juvenile delinquents, do not report that their clients have low self-esteem; in fact, it is often the opposite.) Thus, most people seen in voluntary treatment settings have engaged in negative behaviors *and* have a negative self-image. Therapists conclude that the self-image problem is at the basis of the behavior. It can just as easily be concluded, however, that the self-image problem leads people to therapy, or even that the negative self-image is *valuable* to these people because otherwise they would not be motivated to change their behaviors.

Controlled thinking indicates that the logic of this conclusion is flawed. A critic pointing out the flaw in his or her colleagues' reasoning does not do so on the basis of what comes to mind (the clients he or she is seeing), but quite literally *pauses* to ask "what if?" Such thinking corresponds to Piaget's definition of *formal*. The sample of people who are observed (child abusers who have not stopped on their own) is regarded as one of two possible sets, and the psychotherapist does not have the people in the other set available for observation. The playing field is not level when such logical specification of all possibilities is pitted against automatic thought. In these examples and many others that follow, the logical conclusion of "don't know" is supported, much to the distress of some readers. But it is better to know what we don't know and to deliberately seek more evidence on conclusions that are important, when we don't know.

The prototype of automatic thinking is the thinking involved when we drive a car. We respond to stimuli *not* present in the environment—for example, the expectation that the light will be red before we get to the intersection. Our thought processes are so automatic that we are usually unaware of them. We "steer the car" to reach a desired position without being aware that what we are doing is turning the steering wheel a certain amount so that the car will respond as we desire. It is only when we are learning to drive

that we are aware of the thought processes involved, and in fact we have really learned to drive only when we cease being aware of them. While much of driving involves *motor programs* as opposed to *mental representations,* we nevertheless do "think." This thinking is so automatic, however, that we can carry on conversations at the same time, listen to music, or even create prose or music in other parts of our head. When automatic thinking occurs in less mundane areas, it is often termed *intuition* (e.g., we admire the intuitive wisdom of a respected physician, mechanic, or business leader).

In contrast, a prototype of controlled thought is scientific reasoning. While the original ideas may arise intuitively, they are subjected to rigorous investigation by consideration of *alternative explanations* of the phenomena the ideas seem to explain. (In fact, one way of characterizing Piaget's idea of formal thought is that it is scientific thinking applied to everyday situations.) *Plausible explanations* are considered, and most of them are systematically eliminated by observation, logical reasoning, or experimentation. (However, there are historical instances of ideas later regarded as correct being eliminated as a result of poor experimentation; Schroedinger's equations describing the behavior of the hydrogen atom are an example. The physicist Paul Dirac later commented that Schroedinger had paid too much attention to the experiments, and not enough to the intuition that his equations were "beautiful.")

Occasionally, the degree to which thinking is automatic rather than controlled is not clear until the process is examined carefully. The situation is made more complicated by the fact that any significant intellectual achievement is a mixture of both automatic and controlled thought processes. For example, business executives often claim their decisions are "intuitive," but when questioned reveal that they have systematically "thought through" the relevant alternatives quite deliberately before deciding which "intuition" to honor. At the other extreme, the thinking of chess grandmasters has been shown to be much more automatic than most of us novices believe it to be. When a grandmaster's visual search across the chess board is traced by an eye movement camera, it often shows that the grandmaster looks at the best move first. Then, the subsequent eye movement pattern indicates the grandmaster is checking out alternative possibilities—most often only to come back to the original and best one. Moreover, the grandmaster is not distinguished from the mere expert by the number of moves he or she "looks ahead"; the eye camera indicates that *both* experts and grandmasters look ahead only two or three moves, with a maximum of five. In addition, masters and grandmasters can look at a mid-game position in a typical chess match for 5 seconds and then reproduce it almost perfectly. But mere experts and novices cannot do that. (And no one who has been tested can do it for pieces randomly placed on the board, demonstrating that the ability is not due to a general skill for visual

memory per se.) The conclusion is that grandmasters have a superior understanding of the "meaning" of positions in sensible chess games, that in 5 seconds they can automatically encode entire patterns of pieces as being ones familiar to them, and that they know from experience (estimated to require at least 50,000 hours of practice for master-level players) what constitutes good and bad moves from such patterns. As Herbert Simon and William Chase (1973) summarized their findings, "The most important processes underlying chess mastery are . . . immediate visual-perceptive processes rather than the subsequent logical-deductive thinking processes." Such immediate processes are automatic, like the decision to brake to avoid a collision.

One fundamental point of this book is that we often think in automatic ways when making judgments and choices. These automatic thinking processes can be described by certain psychological rules (e.g., heuristics), and they can systematically lead us to make poorer judgments and choices than we would by thinking in a more controlled manner about our decisions. This is not to say that deliberate, controlled thought is always perfect, or even always better than intuitive thought. In fact, we hope the reader who finishes this book will have a heightened appreciation of the relative advantages of the two modes of thinking and when to trust one or the other.

## 1.3 The Computational Model of the Mind

There has been a modest revolution in the sciences of the mind during the past half-century. A new field has emerged, named *cognitive science,* with a new conceptual paradigm for theorizing about human thought and behavior (Gardner, 1985; Pinker, 1997). The computational model of the mind is based on the assumption that the essence of thinking can be captured by describing what the brain does as manipulating symbols. (Note that we say, "the *essence* of thinking." We do not mean to imply that the brain itself literally manipulates symbols.) The computational model is obviously inspired by an analogy between the computing machine and the computing brain, but it is important to remember that it is an analogy. The two devices, brains and computers, perform similar functions, relating input information to output information (or actions) in an amazingly flexible manner, but their internal structures are quite different (most obviously, electronic circuits and biological neurons operate quite differently).

The central concept in the notion of a computational model is the manipulation of symbolic information. Perhaps the classic example of a cognitive process is the performance of a mental arithmetic task. Suppose we ask you to solve the following addition problem "in your head": 434 + 87 = ???

If we asked you to think aloud, we might hear something like the following: "Okay, I gotta add those numbers up, uh . . . 4 + 7, that's 11 . . . write down the 1, and let's see, carry the 1 . . . ummmm . . . so 3 + 8 equals 11, again, but I gotta add the carry, so that's 12, and uhhhh . . . write down the 2 and I gotta carry a 1 again. Now 4, that's 4, but I have to add the carry, which was 1, so that's 5, write down the 5. So, that's 521. Does that look okay? Yeah, the answer is 521."

Another controlled, deliberate method that one of us (Dawes) uses is to "work down" from the highest multiples of 10, while making a list of "remainders" in "another part of the head." Thus, 434 + 87 is equal to 400, with 34 and 87 remaining. The 87, being larger, is attacked first as 100 minus 20, with a 7 left over. So we now have 400 + 100 – 20 = 480. We now attack the 34, which is larger than the other remainder of 7. It is basically 20 + 10, with a remainder of 4. Because we are already 20 short of 500, we reach it with a remainder of 10 + 4, to which we add the previous remainder of 7 to obtain 21. The answer is 521. While the second algorithm may appear complex upon first being stated, it has the advantage of avoiding silly errors that lead to large mistakes (e.g., as a result of not "aligning" what is to be "carried over"). But a little bit of practice can also lead to the type of speed that absolutely amazes people who don't know the method.

The point is that either of these computational strategies is a good illustration of what we mean by symbol processing: Information goes into your brain through the eyes (or another sense organ); it is converted to some kind of internal, symbolic code, that retains the essential information from the digits; and then we perform mental operations to compare, manipulate, and transform that information, including combining the information from the external problem with our knowledge of arithmetic facts and algorithms we have learned in school. When we believe we have achieved the goal we set for ourselves when we started thinking about the problem, we respond to report the answer. The "amazing flexibility" of thought processes is illustrated by the dramatic differences in the two sequences of thought, which solve the same problem and produce the same final response. (Without some measure of the interior cognitive processes, like the think-aloud reports, it would be impossible to distinguish between the two strategies. To a large extent, this is the primary task of cognitive psychological researchers—scientifically identifying the hidden thought processes that occur "under the hood," in our heads.)

It was tempting to try to create a theory of performance of cognitive tasks by summarizing the contents of think-aloud reports as a sequence of pieces of information (e.g., "the sum for the rightmost column is 11") and operations

on that information to create new information (e.g., "plus" means looking up the sum of two digits in your long-term memory of arithmetic facts). However, such a theoretical endeavor was unsuccessful until we had an appropriate theoretical language in which to express all these complex representations and operations.

The "cognitive revolution" in psychology really got under way (in the 1960s) when the first computer programming languages were applied to the task of summarizing and mimicking the mental operations of people performing intellectual tasks like chess playing, logical deduction, and mental arithmetic. For example, the studies of grandmasters' chess-playing skills we mentioned above were part of a research program at Carnegie Mellon University aimed at describing human cognitive skills (including novice and expert levels) precisely enough so that computational models could be written in computer programming languages to simulate and compete with human players. As Newell and Simon (1972) put it,

> The programmed computer and the human problem solver are both species belonging to the genus "Information Processing System." . . . When we seek to explain the behavior of human problem solvers (or computers for that matter), we discover that their flexibility—their programmability—is the key to understanding them. Their viability depends upon their being able to behave adaptively in a wide range of environments. . . . If we carefully factor out the influences of the task environments from influences of the underlying hardware components and organization, we reveal the true simplicity of the adaptive system. For, as we have seen, we need to postulate only a very simple information processing system to account for human problem solving in such tasks as chess, logic, and cryptarithmetic. The apparently complex behavior of the information processing system in a given environment is produced by the interaction of the demands of that environment with a few basic parameters of the system, particularly characteristics of its memories. (p. 870)

Many aspects of human thinking, including judgment and decision making, can be captured with computational models. The essential parts of these models are symbols (e.g., a theoretical representation of the idea of "yellow," or "pawn," or "11") and operations that compare, combine, and record (in memory) the symbols. Thus, in the chess-playing example, symbols represent the board; the pieces; the rules; and at more complex levels, goals and strategies to win. One of the fundamental and ongoing research projects in cognitive science is to conduct an analysis of the contents of these representations, to describe the natural "mentalese" in which we think and to relate it to the biological substrate in which it must be implemented (e.g., Pinker, 1997, 2007). For purposes of the present book, we can

rely on rudimentary descriptions of mental representations in order to characterize the "knowledge" part of cognitive models of decision processes.

The other half of the cognitive theory is a description of the elementary information processes that operate on the representations to store them, compare them, and transform them in productive thought. It is very important to recognize that most of these operations are unconscious. Although we are aware of (and can report on) some aspects of cognitive processing, mostly the symbolic products of hidden processes such as the digit ideas in mental arithmetic, most of the cognitive system is unconscious. So, the first insight from cognitive science is that we can think of intellectual achievements, like judging and deciding, as computation and that computation can be broken down into symbolic representations and operations on those representations. In addition, we emphasize that both automatic and controlled modes of thinking can be modeled as computations in this sense.

Another important insight from cognitive science concerns the nature of the mechanism (the brain) that performs the computations. Since about 1970, there has been increasing consensus on the nature of the "cognitive architecture" of the human mind. The early outlines of the cognitive system included three kinds of memory stores: sensory input buffers that hold and transform incoming sensory information over a span of a few seconds; a limited short-term working memory where most of conscious thinking occurs; and a capacious long-term memory where we store concepts, images, facts, and procedures. These models provided a good account of simple memory achievements, but were limited in their ability to describe more complex inference, judgment, and decision behaviors. Modern conceptions distinguish between several more processing modules and memory buffers, all linked to a central working memory (Figure 1.1, a good introduction to the modern computational approach is provided by John Anderson, 2000).

In the multi-module model, there are input and output modules, which encode information from each sensory system (relying on one or more memory buffers) and generate motor responses. A *Working Memory,* often analogized to the surface of a workbench on which projects (problems) are completed, is the central hub of the system, and it comprises a central executive processor, a goal stack that organizes processing, and at least two short-term memory buffers that hold visual and verbal information that is currently in use. The other major part of the system is a *Long-Term Memory* that contains all sorts of information including procedures for thinking and deciding. The details of this particular modular division of labor are justified by both behavioral results (e.g., systematic studies of mental arithmetic) and the results of hundreds of studies of brain functions. We will report on some of the more interesting results from neuroscientific analysis of decision processes in Chapter 13.

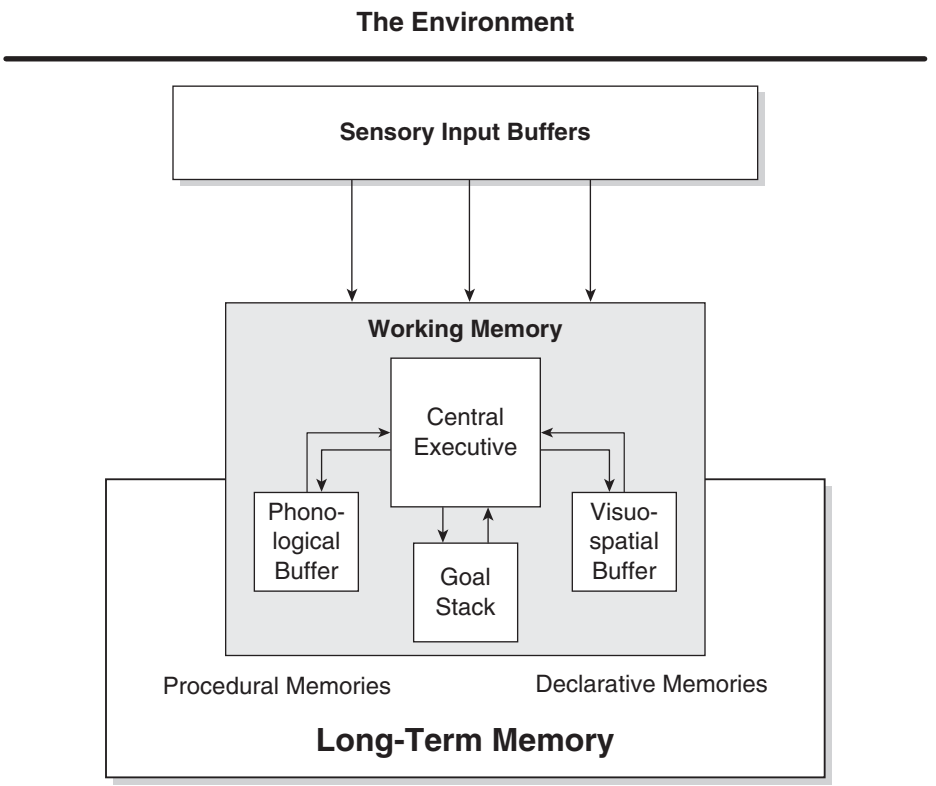


**Figure 1.1** An overview of the human information-processing system (with arrows indicating the flow of information and control from one part of the system to another)

Two properties of the memory stores will play major roles in our explanations for judgment and decision-making phenomena. First, the limited capacity of Working Memory will be used to explain some departures from optimal, rational performance. As Newell and Simon (1972) said (see quote above), "The apparently complex behavior of the information processing system in a given environment is produced by the interaction of the demands of that environment with a few basic parameters of the system, *particularly characteristics of its memories"* (p. 870, emphasis added). James March and Herbert Simon (1958) introduced the concept of *bounded rationality* in decision making, by which they meant approximately optimal behavior, where the primary explanation for departures from optimal is that we simply don't have the capacity to compute the optimal solutions because our working memory

imposes limits on how much information we can use. Second, we will often refer to the many facts and procedures that have been learned and stored in long-term memory. So, for example, we explain the differences between a grandmaster and a novice chess player with reference to stored knowledge about past chess games, good moves, and so forth, and with reference to special analytic skills (analogous to an educated person's knowledge of arithmetic algorithms), all stored almost permanently in long-term memory. (Remember we found that Working Memory differences could not explain the differences in chess skill as a function of expertise; research showed that novices and grandmasters had similar Working Memory capacities—both groups remembered the same number of chess pieces from a scrambled chess board. What the grandmasters seem to have that the novices lack is knowledge about chess stored in their long-term memories. This explains why experts remember so much more from a "meaningful" chess board.)

However, the sheer amount of information to consider—and the limits that Working Memory places on our ability to consider information—is not the only source of bounded rationality. For example, the type of automatic association discussed earlier can also provide impediments to rational thought in the simplest of situations (e.g., the automatic imputation of characteristics of child abusers seen in therapy to child abusers in general). "Information overload" is a sufficient condition for limited ("bounded") rationality, but it is not a necessary condition.

## 1.4 Through Darkest Psychoanalytic Theory and Behaviorism to Cognition

Most of the work discussed in this book has been done in the last half-century. Why? Because until the 1950s, psychology was dominated by two traditions: psychoanalytic theory and behaviorism. Neither of these traditions—which became preeminent in the early 1900s—treated thought as an important determinant of human behavior.

Unconscious needs and desires were the primary stuff of psychoanalytic theory; even defense mechanisms, by which these unconscious impulses could be channeled into socially acceptable—or neurotic—behaviors, were viewed as largely unconscious, and hence outside of the awareness of the individual. (People who claimed to be aware of their own defense mechanisms were said to be denying their problems through "intellectualization"; only the psychoanalyst could really understand them.)

Although dogmatic acceptance of psychoanalytic theory still lingers on in some settings, skepticism was enhanced by its failure to explain one of the

most important psychopathologies of the 20th century, Nazism. A strong implication of the theory was that the Nazi leaders, who engaged in monstrous activities, *had* to be suffering from the types of pathologies postulated by the theory. Moreover, these pathologies had to be related to pathologies and traumas of childhood, which—according to the theory—are crucial to the development of adult disorders. As Wordsworth said, "The child is father to the man." In fact, a 1943 United States Office of Strategic Services report, by Walter C. Langer, was devoted to an analysis of Adolf Hitler and a prediction of his future actions based on his "psychosexual perversion," which was later found not to exist. Supposedly incapable of normal sexual intercourse, Hitler was believed to achieve sexual release through urinating and defecating on his mistress. Moreover, Langer (1943/1972) wrote that Hitler survived World War I by granting homosexual favors to his officers. There is no historical evidence of any such behaviors. In fact, applying Hitler's philosophy of the insignificance of the individual human life to his own life as well as to others, he served without hesitation in the particularly dangerous position of a battlefield messenger, declining promotion to a safer position.

Psychoanalytic interpretations made no mention of Hitler's basic cognitive assumptions about the world, his thinking style, the ways in which he framed problems, or the heuristics he used for solving them. Instead, his behavior was predicted on the basis of his conflicted hatred of his brutal father and his unconscious identification of Germany with his mother. Except for making the somewhat obvious prediction that Hitler wouldn't succeed, this psychoanalytic approach didn't work. Moreover, careful study of the defendants at the Nuremberg war crimes trials—complete with Rorschach inkblot tests—failed to reveal any extraordinary psychosexual disorders or childhood problems. These men and women were ordinary people, much too ordinary. Years later, studying Adolf Eichmann, the SS officer who served as the director of the Central Office for Jewish Emigration and was responsible for the deaths of millions of Jews under the Nazi regime, the philosopher Hannah Arendt (1963) coined the phrase "the banality of evil."

In 1963, Stanley Milgram published his striking experiments on "destructive obedience." In them, he demonstrated that *a variety* of people would administer extremely painful and potentially lethal shocks to strangers as part of a psychological experiment, provided that they were urged to do so by an authority figure who "took responsibility," and that the victim was physically distant from them. (The shocks were not actually administered to the stranger, but the experimental participants were led to believe that they were.) In effect, Milgram did not ask, "How were the Nazis different from us?" but rather, "How are we like the Nazis?" He was able to answer the latter question better than others had answered the former. Subsequent research has

confirmed the general hypothesis of the banality of evil and the power of the immediate social situation to elicit remarkably cruel (or courageous) behavior from otherwise ordinary people (Ross & Nisbett, 1991; Zimbardo, 2007).

According to the behavioristic approach, in sharp contrast to psychoanalysis, the reinforcing properties of the rewards or punishments that follow a behavior determine whether the behavior will become habitual. Awareness is—as in the psychoanalytic tradition—unimportant; at most, it is an "epiphenomenon." B. F. Skinner, probably the most famous behaviorist of all time, put it this way: "The question is not whether machines think, but whether men do." Again, as with psychoanalytic theory, the failure of behaviorism can be attributed to its inability to account for important phenomena, rather than to any direct "disproofs." For example, there are no useful analyses of everyday speech and communication; intellectual achievements like "mental arithmetic" or chess playing; or behavior in modestly complicated gambling decisions from a behaviorist perspective. In fact, to address these phenomena, behaviorists have become so "cognitive" that it is difficult to separate them from psychologists who more comfortably march under the cognitive banner (Rachlin, 1989).

Accounts of even the most elementary learning processes seem to require more structure than is provided by basic behaviorism. For example, people and animals cannot be conditioned to avoid or fear just any food or danger. Children are distinctively nervous about snakes and spiders; rats (and children) are exceptionally sensitive to the pairing of smells and nausea (Garcia & Koelling, 1966; Mineka & Cook, 1993; Seligman, 1971). We are prepared (probably via some form of evolutionary selection) to learn certain associations, especially "causal" associations, and not others; the laws of behaviorist conditioning are not general across stimuli and responses. A related finding is that our conscious understanding of contingencies is a significant moderator, maybe even a necessary condition, for many forms of learning to occur. A number of ingenious experiments have demonstrated not only that awareness of "reinforcement contingencies" is important in determining whether behavior would be repeated, but also that in many areas—notably verbal behavior—such awareness was crucial (e.g., see Dulaney, 1968). This finding contradicted the general "law of effect," which maintains that the influence of consequences is automatic.

Ingenious experiments by Marvin Levine, Gordon Bower, Tom Trabasso, Jerome Bruner, and other early cognitive psychologists are one illustration of the necessity of postulating an active human mind in order to understand behavior (see Levine, 1975, for the history of this revolutionary research). The experiments involved a task termed *concept identification* in which participants are presented with stimuli that differ on many attributes—most

often geometric figures that vary in size, shape, color, and various pattern characteristics. The participant's task is to sort these stimuli into two categories and by so doing, identify the rule (or "concept") that the experimenter has used as the basis for classification. For example, the rule may be that red patterns are to be placed on the left and green ones on the right. Participants are simply told "correct" or "incorrect" when they sort each stimulus, and they are judged to have identified the concept when their sortings are consistently correct (10 correct responses in a row).

Behavioral analyses of responses to this task focused purely on the reinforcement (being told "right" or "wrong") for each choice. Awareness, to the degree that it exists, was assumed not to affect sorting. Early results appeared to support such analyses. For example, some participants were able to achieve perfect sorting without being able to verbalize the experimenter's rule (although it turned out that they could if pressed, their earlier reluctance apparently resulting from being unsure of themselves), and in some tasks participants did not achieve the perfect learning that would be predicted from intellectual insight (but the experimenter's rules themselves may have been ambiguous). Moreover, average success in concept identification *across participants* appeared to increase gradually, much like the learning of an athletic skill.

However, clever follow-up experiments demonstrated that learning in such tasks was in fact not gradual but "all or none," the type of learning predicted on the basis of an active hypothesis-testing mind that continually searches for the correct rule whenever the experimenter indicates that an incorrect sorting has been made. First, these investigators analyzed each participant's responses separately and determined the pattern of correct and incorrect responses *prior to the last error*. If learning was gradual, as predicted by most reinforcement theories, the probability of a correct sort, within a single participant's learning trials, should increase gradually from .50 (the chance probability of being "correct"). Instead, it was *stationary* at .50. The gradual increase found earlier was an artifact of averaging across participants who had identified the correct concept at different points of time in the experiment. Moreover, patterns of sorting after each error were indistinguishable irrespective of the point in the experiment at which the error occurred. By making an error, the participant indicated that he or she "didn't get the concept"; hence, performance was at the chance level prior to each error. An error indicated that the participant had not yet had the insight into the experimenter's rule.

Marvin Levine (1975) demonstrated that participants' conscious beliefs were virtually perfect predictors of their responses, particular error patterns, and time it took to learn. In an especially ingenious demonstration, he

showed that participants failed to learn very simple concepts (e.g., to sort all stimuli to the left), over hundreds of trials, if this concept was unexpected, or "absent from their hypothesis set." Bower and Trabasso (1968) devised a procedure they termed the *alternating reversal shift* procedure. Every *second* time the participant made an error, the rule was reversed. For example, participants who had initially been told "correct" when they placed red figures on the left and green ones on the right were told they were correct the second time they put a green figure on the left (or a red one on the right), and were subsequently told correct or incorrect according to this reversed rule—until they again made a second error, at which point the rule was reversed again. Except for participants lucky enough to identify the concept without making two errors, all participants would be "reinforced" a roughly equal number of times for placing red figures and green figures on the same side. If learning was a simple reinforcement process, participants should never identify the concept. But in fact they did. As a group, they identified the concept after being told they were incorrect (falsely called errors) roughly the same number of times as did those in comparison conditions where the rule was never reversed.

It is almost impossible to explain these results without postulating an active, hypothesis-testing mind mediating between the reinforcement provided by the experimenter and the behavior in the sorting task. Moreover, the mind we hypothesize is a limited mind. For example, participants who perfectly recalled all of their previous choices and the experimenter's responses to them would be totally confused by the alternating reversal shift procedure in the Bower and Trabasso experiments (and suspicious that the experimenter was doing something bizarre—because they were told they were wrong much less than half the time before identifying the concept). It is precisely such a limited, hypothesis-testing mind that this book is written about, and for.

Neither the psychoanalytic nor the behavioral tradition regarded people as decision makers who deliberately weighed the consequences of various courses of action and then chose from among them. Moreover, neither tradition has contributed useful explanations of decision-making behaviors. Most psychologists today accept the compelling assumption that ideas and beliefs cause behavior and that cognitive theories are the best route to understanding and improving important behaviors. If we want to understand why the juror said the defendant was a murderer, why the doctor diagnosed the patient with a blocked kidney duct, or why the pilot diverted to another airport for an unscheduled landing, the best way to proceed is to find out what they were thinking about before they made each of these decisions. This book uses such cognitive science concepts to better understand judgment and choice.

## 1.5 Quality of Choice: Rationality

If we aspire to give advice about how to make good decisions, we need to say something about what we mean by bad decisions. The quality of a decision cannot be determined unambiguously by its outcome. For example, most of us believe it would be foolish to accept an "even money" bet that the next time we throw a pair of dice we will roll "snake eyes." (The actual chance of throwing two ones, "snake eyes," is 1/36). Moreover, we would regard the person who accepted such a wager as a poor decision maker—even if he or she happened to roll snake eyes. On the other hand, if that person were in danger of physical harm or death at the hands of a loan shark, and that wager were the only way to raise enough money to avoid harm, then the person might not seem so foolish. What this example illustrates is that it is the potential outcomes, their probabilities, and their values to the decision maker *at the time the decision is made* that lead us to judge a particular choice to be wise or foolish. A general who is losing a war, for example, is much wiser to engage in a high-risk military venture than is a general who is winning a war. The failure of such a venture might not reflect unfavorably on the decision-making ability of the losing general; it is more "rational" for the losing general to take a risk.

So what is rationality? Often the term is used in a purely egocentric, evaluative sense: "Decisions I make are 'rational'; those of which I disapprove are not." Occasionally, we adopt a broader perspective, and judge rationality not just in terms of approval but in terms of the "best interests" *of the person making the decision*—although with "best interests" still defined egocentrically by *us*. As we said at the outset, good decisions are those that choose means, available in the circumstances, to achieve the decision maker's goals. Thus, for example, some of Adolf Hitler's decisions may be viewed as rational (and others as irrational), despite the fact that we disapprove of all of them.

In this book, *rationality* has a narrow technical meaning; it will nevertheless provide the criterion by which we will judge the wisdom of choices. A *rational choice* can be defined as one that meets four criteria:

1. It is based on the decision maker's current assets. Assets include not only money, but also physiological state, psychological capacities, social relationships, and feelings.
2. It is based on the possible consequences of the choice.
3. When these consequences are uncertain, their likelihood is evaluated according to the basic rules of probability theory.
4. It is a choice that is adaptive within the constraints of those probabilities and the values or satisfactions associated with each of the possible consequences of the choice.

Don't we make all our decisions like that? Definitely not. For example, Chapter 2 will detail how it is that we are affected not only by our present state but also by *how we got to it*—a clear violation of the first two criteria enunciated above. The past is over and cannot be changed, but we often let it influence our futures in an irrational manner. In Chapters 9 and 12, we will show how we are sensitive not just to the actual consequences of our decisions but also to the way in which we *frame* these consequences. Chapters 4 through 10 are devoted in large part to the cognitive heuristics (boundedly rational rules of thumb) we use to judge future likelihood—heuristics that systematically violate the rules of probability theory. Finally, Chapters 8 through 11 describe ways of making decisions that avoid the problems specified in the previous sections.

In fact, there are common decision-making procedures that have no direct relationship to these criteria of rationality. They include the following:

1. Habit, choosing what we have chosen before;
2. Conformity, making whatever choice (you think) most other people would make or imitating the choices of people you admire (Boyd and Richerson [1982] have pointed out that imitation of success can be adaptive in general, though not, for example, if it is imitation of the drug use of a particular rock star or professional athlete you admire for his or her professional achievements); and
3. Choosing on the basis of (your interpretation of) religious principles or cultural mandates.

The four criteria of rationality have a philosophical basis. If any are violated, the decision maker can reach contradictory conclusions about what to choose—even though the conclusions are based on the same preferences and the same knowledge. That is, the person violating these principles may decide that a course of action is simultaneously desirable and undesirable, or that choice A is preferable to choice B *and* that choice B is preferable to choice A. For example, a business executive who attends not just to the current assets of the company but also to the fact that they have been increasing or decreasing in the past could conclude that it is both wise and unwise to continue to finance a losing venture. A doctor whose probabilistic reasoning follows automatic thinking principles rather than the rules of probability could decide that a patient both should and should not have an operation; or a juror could decide that a defendant was both guilty and innocent. Because reality is not contradictory, contradictory thinking is irrational thinking. A proposition about reality cannot be both true and false.

## 1.6 The Invention of Modern Decision Theory

Where does this idea of rationality come from? It began in Renaissance Italy, for example, in the analysis of the practice of gambling by scholars such as Girolamo Cardano (1501–1576), a true Renaissance man who was simultaneously a mathematician, physician, accountant, and inveterate gambler. (He is also credited with inventing the combination lock.) In spite of his profound insights into risky decision making, he tended to lose, because his analyses of the numerical structure of random situations were accompanied by lousy arithmetic skills. The most recent impetus for the development of a rational decision theory, however, comes from a book published in 1947 entitled *Theory of Games and Economic Behavior* by mathematician John von Neumann and economist Oskar Morgenstern. (The first publication in 1944 omitted some of the most important analyses of decision making, so we cite the 1947 edition.) Von Neumann and Morgenstern provided a theory of decision making according to the principle of maximizing *expected utility*. The book does not discuss behavior per se; rather, it is a purely mathematical work that applies utility theory to optimal economic decisions. Its relevance to non-economic decisions was assured by basing the theoretical development on general *utility* (we prefer the term *personal value*), rather than solely on monetary outcomes.

This criterion of expected utility may most easily be understood by analyzing simple gambling situations. Because gambling situations are familiar and well-defined, we will rely on them heavily (as have most scholars in this area) to illustrate basic concepts, though we will try to provide a diverse collection of nonmonetary, everyday examples as well. Consider, for example, a choice between two gambles:

(a) With probability .20 win $45, otherwise nothing.

(b) With probability .25 win $30, otherwise nothing.

The *expected value* of each is equal to the probability of winning multiplied by the amount to be won. Thus, the expected value of gamble (a) is $.20 \times \$45 = \$9$, while that of gamble (b) is $.25 \times \$30 = \$7.50$. People need not, however, prefer gamble (a) simply because its expected value is higher. Depending upon their circumstances, they may find $30 to have more than four-fifths the *utility* of $45, in which case they would—according to the theory—choose gamble (b). For example, an individual may be out of money at the end of a week and simply desire to have enough money to eat until the following Monday. In that situation, the individual may find the difference in utility between $30 and $45 to be negligible compared with the difference between a one-fourth and a one-fifth chance of receiving any money at all.

Such a preference is represented in the von Neumann and Morgenstern theory by the conclusion that .25 times *that individual's utility* for $30 is greater than .20 times *that individual's utility* for $45. Let the utility of $30 be symbolized U($30) and the utility of $45 be symbolized U($45); then by simple algebra, .25 × U($30) > .20 × U($45), which is true if and only if U($30)/U($45) > .20/.25 (which is equal to 4/5).

In point of fact, most people when asked prefer gamble (a). But when faced with the choice between the following two gambles, most prefer (b′), the one with the $30 payoff:

(a′) With probability .80 win $45, otherwise nothing.

(b′) Win $30 for sure.

An individual who preferred (a) to (b) yet (b′) to (a′) would *violate* the von Neumann and Morgenstern principle of choosing according to expected utility. Using the same algebraic symbolism as before, a choice of (a) over (b) implies that .20 × U($45) > .25 × U($30), or U($45)/U($30) > .25/.20 = 5/4. But a choice of (b′) over (a′) implies that .80 U($45) < U($30), or U($45)/U($30) < 1/.80 = 5/4. So, there is a logical (algebraic) contradiction between the two choices. This means the theory not only specifies what is rational, but it can also be compared against human choices to test if people are rational.

Another possible violation of expected utility theory would occur if a person were willing to pay more for one gamble than another, yet preferred the other gamble when given a choice between the two. For example, such a person might prefer the sure $30 of alternative (b′), yet—realizing that (a′) has a higher expected value ($36 vs. $30)—be willing to pay more to play it than to play (b′). The theory equates the utility of each gamble with the utility of the maximal amount of money paid for playing each. The result is that by preferring the gamble for which he or she was willing to pay less, a person has implicitly indicated a preference for less money over more. Assuming any positive utility at all for money (a "no brainer" assumption), that is irrational—because the greatest amount of money is equal to the lesser amount plus some more. The conditions that lead to such contradictions will be discussed in Chapters 12 and 13.

What is important here, however, is not just that some choices can contradict expected utility theory, but that the four criteria of rationality listed above are *preconditions* for the development of expected utility theory. Thus, choices that violate expected utility theory can also violate very simple, fundamental, and plausible criteria for good decisions, criteria that almost all of us would say we would like to follow when we make important choices. Again, there is nothing in the theory that mandates what desires a decision

maker should wish to satisfy—that is, the theory does not prescribe *what* the utilities for various outcomes should be. But the theory does imply fairly strong relationships between some choices and other preferences.

Von Neumann and Morgenstern's work *Theory of Games and Economic Behavior* (1947) inspired a lot of interest in utility theory; many mathematically oriented researchers worked to draw out consequences of maximizing expected utility that were not present in the initial formulation. Others suggested that the basic formulation might be in error, but they did not advocate abandoning the four criteria of rationality; instead, often supported by examples that were intuitively compelling, they suggested that rational decision makers might choose according to some rational principle other than maximizing expected utility. These initial works focused on the *normative* question of how decision makers *should* choose. Soon, however, people became interested in the *descriptive* question of how decision makers—people, groups, organizations, and governments—*actually* choose. Do actual choices conform to the principle of maximizing expected utility?

The answer to this question appears to depend in large part on the field of the person asking it. Traditional economists, looking at the aggregate behavior of many individual decision makers in broad economic contexts, are satisfied that the principle of maximizing expected utility does describe what happens. As Gary Becker (1976), Nobel Prize–winning behavioral scientist, puts it, "All human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs from a variety of markets" (p. 14). Becker and many of his colleagues have taken this assertion seriously and have provided insightful analyses of nonfinancial, nonmarket behaviors including marriage, education, and murder.

There are good reasons to start with the optimistic hypothesis that the rational, expected utility theory and the descriptive—how people really behave—theories are the same. After all, our decision-making habits have been "designed" by millions of years of evolutionary selection and, if that weren't enough, have been shaped by a lifetime of adaptive learning experiences. Surely, truly maladaptive habits will have been eliminated by the pressures of evolution and learning and maybe, optimistically, only the rational tendencies are still intact.

In contrast, psychologists and behavioral economists studying the decision making of individuals and organizations tend to reach the opposite conclusion from that of traditional economists. Not only do the choices of individuals and social decision-making groups tend to violate the principle of maximizing expected utility; they are also often patently irrational. (Recall that irrationality as discussed here means that the chooser violates the rules of rational decision making and chooses contradictory courses of

action. We are not talking about the nature of the *goals* of the decision maker; we are talking about the failure to pursue those goals coherently, whatever those goals might be for the individual.) What is of more interest is that people are not just irrational, but irrational in *systematic* ways—ways related to their automatic or "bounded" thinking habits. Chapters 4 through 10 of this book are devoted to a discussion of these systematic irrationalities.

Those behavioral scientists who conclude that the rational model is not a good descriptive model have also criticized the apparent descriptive successes of the rational model reported by Becker and others. The catch is that by specifying the theory in terms of utility rather than concrete values (like dollars), it is almost always possible to *assume* that some sort of maximization principle works and then, ex post, to define utilities accordingly. This is analogous to the assertion that all people are "selfish," by definition, because they do what they "want" to do. (As James Buchanan [1978] points out, many aspects of standard economic theory tend to be "vacuously true" when phrased in terms of utilities, but demonstratively false if money is substituted for utility. In addition, Herbert Simon [1959], defending his more psychological approach, has pointed out some of the explanatory contortions that are necessary to make expected utility theory work descriptively.) However, the best arguments that these principles do not work descriptively come from demonstrations of out-and-out irrationality in light of our four criteria for rational individual decision making (see above).

This book reflects the mixture of approaches to judgment and decision making that has characterized this complex field since its beginnings—the rational, normative hypotheses (often accompanied by the optimistic notion that we approximate the rational in our actual behavior) versus the cognitive, descriptive hypotheses about how we really behave. Both the top-down normative view and the bottom-up descriptive approach are necessary to understand the ideal of adaptive rationality and the reality of human decision-making processes. Moreover, important insights into human nature result from knowing when we do behave rationally, adaptively. Perhaps most important of all, knowing when human behavior departs from the rational model is the first step in designing improvements in our essential thinking skills.

## References

Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York: Worth Publishers.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Arendt, H. (1963). *Eichmann in Jerusalem: A report on the banality of evil.* New York: Viking Press.

Bartlett, F. C. (1958). *Thinking: An experimental and social study*. New York: Basic Books.

Becker, G. (1976). *The economic approach to human behavior*. Chicago: University of Chicago Press.

Bigelow, J. (Ed.). (1887). *The complete works of Benjamin Franklin*. New York: Putnam.

Bower, G. H., & Trabasso, T. (1968). *Attention in learning*. New York: Wiley.

Boyd, R., & Richerson, P. J. (1982). Cultural transmission and the evolution of cooperative behavior. *Human Ecology, 10,* 325–351.

Buchanan, J. M. (1978). *Cost and choice: An inquiry in economic theory*. Chicago: University of Chicago Press.

Dulaney, D. E. (1968). Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In T. R. Dixon & D.R. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 98–109). Englewood Cliffs, NJ: Prentice Hall.

Garcia, J., & Koelling, R. A. (1966). The relation of cue to consequence in avoidance learning. *Psychonomic Science, 4,* 123–124.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Langer, W. C. (1972). *Adolf Hitler: The secret wartime report*. New York: Basic Books. (Published version of Langer's 1943 *Wartime Report to O.S.S.*)

Levine, M. (1975). *A cognitive theory of learning*. Hillsdale, NJ: Laurence Erlbaum.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Mineka, S., & Cook, M. (1993). Mechanisms involved in the observational conditions of fear. *Journal of Experimental Psychology, 122,* 23–38.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. New York: Viking.

Rachlin, H. (1989). *Judgment, decision, and choice*. New York: W. H. Freeman.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. New York: McGraw-Hill.

Seligman, M. E. (1971). Phobias and preparedness. *Behavior Therapy, 2,* 307–320.

Simon, H. A. (1959). Theories of decision making in economics and behavioral science. *American Economic Review, 49,* 253–280.

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist, 61,* 394–403.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Zimbardo, P. (2007). *The Lucifer effect: Understanding how good people turn evil*. New York: Random House.