Intuitions vs. Formulas

Paul Meehl was a strange and wonderful character, and one of the most versatile psychologists of the twentieth century. Among the departments in which he had faculty appointments at the University of Minnesota were psychology, law, psychiatry, neurology, and philosophy. He also wrote on religion, political science, and learning in rats. A statistically sophisticated researcher and a fierce critic of empty claims in clinical psychology, Meehl was also a practicing psychoanalyst. He wrote thoughtful essays on the philosophical foundations of psychological research that I almost memorized while I was a graduate student. I never met Meehl, but he was one of my heroes from the time I read his *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*.

In the slim volume that he later called “my disturbing little book,” Meehl reviewed the results of 20 studies that had analyzed whether *clinical predictions* based on the subjective impressions of trained professionals were more accurate than *statistical* predictions made by combining a few scores or ratings according to a rule. In a typical study, trained counselors predicted the grades of freshmen at the end of the school year. The counselors interviewed each student for forty-five minutes. They also had access to high school grades, several aptitude tests, and a four-page personal statement. The statistical algorithm used only a fraction of this information: high school grades and one aptitude test. Nevertheless, the formula was more accurate than 11 of the 14 counselors. Meehl reported generally similar results across a variety of other forecast outcomes, including violations of parole, success in pilot training, and criminal recidivism.

Not surprisingly, Meehl’s book provoked shock and disbelief among clinical psychologists, and the controversy it started has engendered a stream of research that is still flowing today, more than fifty yephy⍰Љ diars after its publication. The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred, but the score in the contest between algorithms and humans has not changed. About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgment. No exception has been convincingly documented.

The range of predicted outcomes has expanded to cover medical variables such as the longevity of cancer patients, the length of hospital stays, the diagnosis of cardiac disease, and the susceptibility of babies to sudden infant death syndrome; economic measures such as the prospects of success for new businesses, the evaluation of credit risks by banks, and the future career satisfaction of workers; questions of interest to government agencies, including assessments of the suitability of foster parents, the odds of recidivism among juvenile offenders, and the likelihood of other forms of violent behavior; and miscellaneous outcomes such as the evaluation of scientific presentations, the winners of football games, and the future prices of Bordeaux wine. Each of these domains entails a significant degree of uncertainty and unpredictability. We describe them as "low-validity environments." In every case, the accuracy of experts was matched or exceeded by a simple algorithm.

As Meehl pointed out with justified pride thirty years after the publication of his book, "There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one."

The Princeton economist and wine lover Orley Ashenfelter has offered a compelling demonstration of the power of simple statistics to outdo world-renowned experts. Ashenfelter wanted to predict the future value of fine Bordeaux wines from information available in the year they are made. The question is important because fine wines take years to reach their peak quality, and the prices of mature wines from the same vineyard vary dramatically across different vintages; bottles filled only twelve months apart can differ in value by a factor of 10 or more. An ability to forecast future prices is of substantial value, because investors buy wine, like art, in the anticipation that its value will appreciate.

It is generally agreed that the effect of vintage can be due only to variations in the weather during the grape-growing season. The best wines are produced when the summer is warm and dry, which makes the Bordeaux wine industry a likely beneficiary of global warming. The industry is also helped by wet springs, which increase quantity without much effect on quality. Ashenfelter converted that conventional knowledge into a statistical formula that predicts the price of a wine—

for a particular property and at a particular age—by three features of the weather: the average temperature over the summer growing season, the amount of rain at harvest-time, and the total rainfall during the previous winter. His formula provides accurate price forecasts years and even decades into the future. Indeed, his formula forecasts future prices much more accurately than the current prices of young wines do. This new example of a "Meehl pattern" challenges the abilities of the experts whose opinions help shape the early price. It also challenges economic theory, according to which prices should reflect all the available information, including the weather. Ashenfelter's formula is extremely accurate—the correlation between his predictions and actual prices is above .90.

Why are experts e yinferior to algorithms? One reason, which Meehl suspected, is that experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity. Simple combinations of features are better. Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not. According to Meehl, there are few circumstances under which it is a good idea to substitute judgment for a formula. In a famous thought experiment, he described a formula that predicts whether a particular person will go to the movies tonight and noted that it is proper to disregard the formula if information is received that the individual broke a leg today. The name "broken-leg rule" has stuck. The point, of course, is that broken legs are very rare—as well as decisive.

Another reason for the inferiority of expert judgment is that humans are incorrigibly inconsistent in making summary judgments of complex information. When asked to evaluate the same information twice, they frequently give different answers. The extent of the inconsistency is often a matter of real concern. Experienced radiologists who evaluate chest X-rays as "normal" or "abnormal" contradict themselves 20% of the time when they see the same picture on separate occasions. A study of 101 independent auditors who were asked to evaluate the reliability of internal corporate audits revealed a similar degree of inconsistency. A review of 41 separate studies of the reliability of judgments made by auditors, pathologists, psychologists, organizational managers, and other professionals suggests that this level of inconsistency is typical, even when a case is reevaluated within a few minutes. Unreliable judgments cannot be valid predictors of anything.

The widespread inconsistency is probably due to the extreme context dependency of System 1. We know from studies of priming that unnoticed stimuli in our environment have a substantial influence on our thoughts and actions. These influences fluctuate from moment to moment. The brief pleasure of a cool breeze on a hot day may make you slightly more positive and optimistic about whatever you are evaluating at the time. The prospects of a convict being granted parole may change significantly during the time that elapses between successive food breaks in the parole judges' schedule. Because you have little direct knowledge of what goes on in your mind, you will never know that you might have made a different judgment or reached a different decision under very slightly different circumstances. Formulas do not suffer from such problems. Given the same input, they always return the same answer. When predictability is poor—which it is in most of the studies reviewed by Meehl and his followers—inconsistency is destructive of any predictive validity.

The research suggests a surprising conclusion: to maximize predictive accuracy, final decisions should be left to formulas, especially in low-validity environments. In admission decisions for medical schools, for example, the final determination is often made by the faculty members who interview the candidate. The evidence is fragmentary, but there are solid grounds for a conjecture: conducting an interview is likely to diminish the accuracy of a selection procedure, if the interviewers also make the final admission decisions. Because interviewers are overconfident in their intuitions, they will assign too much weight to their personal impressions and too little weight to other sources of information, lowering validity. Similarly, the experts who evaluate the quas plity of immature wine to predict its future have a source of information that almost certainly makes things worse rather than better: they can taste the wine. In addition, of course, even if they have a good understanding of the effects of the weather on wine quality, they will not be able to maintain the consistency of a formula.

The most important development in the field since Meehl's original work is Robyn Dawes's famous article "The Robust Beauty of Improper Linear Models in Decision Making." The dominant statistical practice in the social sciences is to assign weights to the different predictors by following an algorithm, called multiple regression, that is now built into conventional software. The logic of multiple regression is unassailable: it finds the optimal formula for putting together a weighted combination of the predictors. However, Dawes observed that the complex statistical algorithm adds little or no value. One can do just as well by selecting a set of scores that have some validity for predicting the outcome and adjusting the values to make them comparable (by using standard scores or ranks). A formula that combines these predictors with equal weights is likely to be just as accurate in predicting new cases as the multiple-regression formula that was optimal in the original sample. More recent research went further: formulas that assign equal weights to all the predictors are often superior, because they are not affected by accidents of sampling.

The surprising success of equal-weighting schemes has an important practical implication: it is possible to develop useful algorithms without any prior statistical research. Simple equally weighted formulas based on existing statistics or on common sense are often very good predictors of significant outcomes. In a memorable example, Dawes showed that marital stability is well predicted by a formula:

frequency of lovemaking minus frequency of quarrels

You don't want your result to be a negative number.

The important conclusion from this research is that an algorithm that is constructed on the back of an envelope is often good enough to compete with an optimally weighted formula, and certainly good enough to outdo expert judgment. This logic can be applied in many domains, ranging from the selection of stocks by portfolio managers to the choices of medical treatments by doctors or patients.

A classic application of this approach is a simple algorithm that has saved the lives of hundreds of thousands of infants. Obstetricians had always known that an infant who is not breathing normally within a few minutes of birth is at high risk of brain damage or death. Until the anesthesiologist Virginia Apgar intervened in 1953, physicians and midwives used their clinical judgment to determine whether a baby was in distress. Different practitioners focused on different cues. Some watched for breathing problems while others monitored how soon the baby cried. Without a standardized procedure, danger signs were often missed, and many newborn infants died.

One day over breakfast, a medical resident asked how Dr. Apgar would make a systematic assessment of a newborn. “That’s easy,” she replied. “You would do it like this.” Apgar jotted down five variables (heart rate, respiration, reflex, muscle tone, and color) and three scores (0, 1, or 2, depending on the robustness of each sign). Realizing that she might have made a breakequthrough that any delivery room could implement, Apgar began rating infants by this rule one minute after they were born. A baby with a total score of 8 or above was likely to be pink, squirming, crying, grimacing, with a pulse of 100 or more—in good shape. A baby with a score of 4 or below was probably bluish, flaccid, passive, with a slow or weak pulse—in need of immediate intervention. Applying Apgar’s score, the staff in delivery rooms finally had consistent standards for determining which babies were in trouble, and the formula is credited for an important contribution to reducing infant mortality. The Apgar test is still used every day in every delivery room. Atul Gawande’s recent *A Checklist Manifesto* provides many other examples of the virtues of checklists and simple rules.

The Hostility to Algorithms

From the very outset, clinical psychologists responded to Meehl’s ideas with hostility and disbelief. Clearly, they were in the grip of an illusion of skill in terms of their ability to make long-term predictions. On reflection, it is easy to see how the illusion came about and easy to sympathize with the clinicians’ rejection of Meehl’s research.

The statistical evidence of clinical inferiority contradicts clinicians’ everyday experience of the quality of their judgments. Psychologists who work with patients have many hunches during each therapy session, anticipating how the patient will respond to an intervention, guessing what will happen next. Many of these hunches are confirmed, illustrating the reality of clinical skill.

The problem is that the correct judgments involve short-term predictions in the context of the therapeutic interview, a skill in which therapists may have years of practice. The tasks at which they fail typically require long-term predictions about the patient's future. These are much more difficult, even the best formulas do only modestly well, and they are also tasks that the clinicians have never had the opportunity to learn properly—they would have to wait years for feedback, instead of receiving the instantaneous feedback of the clinical session. However, the line between what clinicians can do well and what they cannot do at all well is not obvious, and certainly not obvious to them. They know they are skilled, but they don't necessarily know the boundaries of their skill. Not surprisingly, then, the idea that a mechanical combination of a few variables could outperform the subtle complexity of human judgment strikes experienced clinicians as obviously wrong.

The debate about the virtues of clinical and statistical prediction has always had a moral dimension. The statistical method, Meehl wrote, was criticized by experienced clinicians as "mechanical, atomistic, additive, cut and dried, artificial, unreal, arbitrary, incomplete, dead, pedantic, fractionated, trivial, forced, static, superficial, rigid, sterile, academic, pseudoscientific and blind." The clinical method, on the other hand, was lauded by its proponents as "dynamic, global, meaningful, holistic, subtle, sympathetic, configural, patterned, organized, rich, deep, genuine, sensitive, sophisticated, real, living, concrete, natural, true to life, and understanding."

This is an attitude we can all recognize. When a human competes with a machine, whether it is John Henry a-hammerin' on the mountain or the chess genius Garry Kasparov facing off against the computer Deep Blue, our sympathies lie with our fellow human. The aversion to algorithms making decisions that affect humans is rooted in the strong preference that many people have for the ormnatural over the synthetic or artificial. Asked whether they would rather eat an organic or a commercially grown apple, most people prefer the "all natural" one. Even after being informed that the two apples taste the same, have identical nutritional value, and are equally healthful, a majority still prefer the organic fruit. Even the producers of beer have found that they can increase sales by putting "All Natural" or "No Preservatives" on the label.

The deep resistance to the demystification of expertise is illustrated by the reaction of the European wine community to Ashenfelter's formula for predicting the price of Bordeaux wines. Ashenfelter's formula answered a prayer: one might thus have expected that wine lovers everywhere would be grateful to him for demonstrably improving their ability to identify the wines that later would be good. Not so. The response in French wine circles, wrote *The New York Times*, ranged "somewhere between violent and hysterical." Ashenfelter reports that one oenophile called his findings "ludicrous and absurd." Another scoffed, "It is like judging movies without actually seeing them."

The prejudice against algorithms is magnified when the decisions are consequential. Meehl remarked, "I do not quite know how to alleviate the horror some clinicians seem to experience when they envisage a treatable case being denied treatment because a 'blind, mechanical' equation misclassifies him." In contrast, Meehl and other proponents of algorithms have argued strongly that it is unethical to rely on intuitive judgments for important decisions if an algorithm is available that will make fewer mistakes. Their rational argument is compelling, but it runs against a stubborn psychological reality: for most people, the cause of a mistake matters. The story of a child dying because an algorithm made a mistake is more poignant than the story of the same tragedy occurring as a result of human error, and the difference in emotional intensity is readily translated into a moral preference.

Fortunately, the hostility to algorithms will probably soften as their role in everyday life continues to expand. Looking for books or music we might enjoy, we appreciate recommendations generated by soft ware. We take it for granted that decisions about credit limits are made without the direct intervention of any human judgment. We are increasingly exposed to guidelines that have the form of simple algorithms, such as the ratio of good and bad cholesterol levels we should strive to attain. The public is now well aware that formulas may do better than humans in some critical decisions in the world of sports: how much a professional team should pay for particular rookie players, or when to punt on fourth down. The expanding list of tasks that are assigned to algorithms should eventually reduce the discomfort that most people feel when they first encounter the pattern of results that Meehl described in his disturbing little book.

Learning from Meehl

In 1955, as a twenty-one-year-old lieutenant in the Israeli Defense Forces, I was assigned to set up an interview system for the entire army. If you wonder why such a responsibility would be forced upon someone so young, bear in mind that the state of Israel itself was only seven years old at the time; all its institutions were under construction, and someone had to build them. Odd as it sounds today, my bachelor's degree in psychology probably qualified me as the best-trained psychologist in the army. My direct supervisor, a brilliant researcher, had a degree in chemistry.

An idilnterview routine was already in place when I was given my mission. Every soldier drafted into the army completed a battery of psychometric tests, and each man considered for combat duty was interviewed for an assessment of personality. The goal was to assign the recruit a score of general fitness for combat and to find the best match of his personality among various branches: infantry, artillery, armor, and so on. The interviewers were themselves young draftees, selected for this assignment by virtue of their high intelligence and interest in dealing with people. Most were women, who were at the time exempt from combat duty. Trained for a few weeks in how to conduct a fifteen- to twenty-minute interview, they were encouraged to cover a range of topics and to form a general impression of how well the recruit would do in the army.

Unfortunately, follow-up evaluations had already indicated that this interview procedure was almost useless for predicting the future success of recruits. I was instructed to design an interview that would be more useful but would not take more time. I was also told to try out the new interview and to evaluate its accuracy. From the perspective of a serious professional, I was no more qualified for the task than I was to build a bridge across the Amazon.

Fortunately, I had read Paul Meehl's "little book," which had appeared just a year earlier. I was convinced by his argument that simple, statistical rules are superior to intuitive "clinical" judgments. I concluded that the then current interview had failed at least in part because it allowed the interviewers to do what they found most interesting, which was to learn about the dynamics of the interviewee's mental life. Instead, we should use the limited time at our disposal to obtain as much specific information as possible about the interviewee's life in his normal environment. Another lesson I learned from Meehl was that we should abandon the procedure in which the interviewers' global evaluations of the recruit determined the final decision. Meehl's book suggested that such evaluations should not be trusted and that statistical summaries of separately evaluated attributes would achieve higher validity.

I decided on a procedure in which the interviewers would evaluate several relevant personality traits and score each separately. The final score of fitness for combat duty would be computed according to a standard formula, with no further input from the interviewers. I made up a list of six characteristics that appeared relevant to performance in a combat unit, including "responsibility," "sociability," and "masculine pride." I then composed, for each trait, a series of factual questions about the individual's life before his enlistment, including the number of different jobs he had held, how regular and punctual he had been in his work or studies, the frequency of his interactions with friends, and his interest and participation in sports, among others. The idea was to evaluate as objectively as possible how well the recruit had done on each dimension.

By focusing on standardized, factual questions, I hoped to combat the halo effect, where favorable first impressions influence later judgments. As a further precaution against halos, I instructed the interviewers to go through the six traits in a fixed sequence, rating each trait on a five-point scale before going on to the next. And that was that. I informed the interviewers that they need not concern themselves with the recruit's future adjustment to the military. Their only task was to elicit relevant facts about his past and to use that information to score each personality dimension. "Your function is to provide reliable measurements," I told them. "Leave the predicok tive validity to me," by which I meant the formula that I was going to devise to combine their specific ratings.

The interviewers came close to mutiny. These bright young people were displeased to be ordered, by someone hardly older than themselves, to switch off their intuition and focus entirely on boring factual questions. One of them complained, "You are turning us into robots!" So I compromised. "Carry out the interview exactly as instructed," I told them, "and when you are done, have your wish: close your eyes, try to imagine the recruit as a soldier, and assign him a score on a scale of 1 to 5."

Several hundred interviews were conducted by this new method, and a few months later we collected evaluations of the soldiers' performance from the commanding officers of the units to which they had been assigned. The results made us happy. As Meehl's book had suggested, the new interview procedure was a substantial improvement over the old one. The sum of our six ratings predicted soldiers' performance much more accurately than the global evaluations of the previous interviewing method, although far from perfectly. We had progressed from "completely useless" to "moderately useful."

The big surprise to me was that the intuitive judgment that the interviewers summoned up in the "close your eyes" exercise also did very well, indeed just as well as the sum of the six specific ratings. I learned from this finding a lesson that I have never forgotten: intuition adds value even in the justly derided selection interview, but only after a disciplined collection of objective information and disciplined scoring of separate traits. I set a formula that gave the "close your eyes" evaluation the same weight as the sum of the six trait ratings. A more general lesson that I learned from this episode was do not simply trust intuitive judgment—your own or that of others—but do not dismiss it, either.

Some forty-five years later, after I won a Nobel Prize in economics, I was for a short time a minor celebrity in Israel. On one of my visits, someone had the idea of escorting me around my old army base, which still housed the unit that interviews new recruits. I was introduced to the commanding officer of the Psychological Unit, and she described their current interviewing practices, which had not changed much from the system I had designed; there was, it turned out, a considerable amount of research indicating that the interviews still worked well. As she came to the end of her description of how the interviews are conducted, the officer added, "And then we tell them, 'Close your eyes.'"

Do It Yourself

The message of this chapter is readily applicable to tasks other than making manpower decisions for an army. Implementing interview procedures in the spirit of Meehl and Dawes requires relatively little effort but substantial discipline. Suppose that you need to hire a sales representative for your firm. If you are serious about hiring the best possible person for the job, this is what you should do. First, select a few traits that are prerequisites for success in this position (technical proficiency, engaging personality, reliability, and so on). Don't overdo it —six dimensions is a good number. The traits you choose should be as independent as possible from each other, and you should feel that you can assess them reliably by asking a few factual questions. Next, make a list of those questions for each trait and think about how you will score it, say on a 1–5 scale. You should have an idea of what you will caleigl "very weak" or "very strong."

These preparations should take you half an hour or so, a small investment that can make a significant difference in the quality of the people you hire. To avoid halo effects, you must collect the information on one trait at a time, scoring each before you move on to the next one. Do not skip around. To evaluate each candidate, add up the six scores. Because you are in charge of the final decision, you should not do a "close your eyes." Firmly resolve that you will hire the candidate whose final score is the highest, even if there is another one whom you like better—try to resist your wish to invent broken legs to change the ranking. A vast amount of research offers a promise: you are much more likely to find the best candidate if you use this procedure than if you do what people normally do in such situations, which is to go into the interview unprepared and to make choices by an overall intuitive judgment such as "I looked into his eyes and liked what I saw."

Speaking of Judges vs. Formulas

> "Whenever we can replace human judgment by a formula, we should at least consider it."

> "He thinks his judgments are complex and subtle, but a simple combination of scores could probably do better."

> "Let's decide in advance what weight to give to the data we have on the candidates' past performance. Otherwise we will give too much weight to our impression from the interviews."

Expert Intuition: When Can We Trust It?

Professional controversies bring out the worst in academics. Scientific journals occasionally publish exchanges, often beginning with someone's critique of another's research, followed by a reply and a rejoinder. I have always thought that these exchanges are a waste of time. Especially when the original critique is sharply worded, the reply and the rejoinder are often exercises in what I have called sarcasm for beginners and advanced sarcasm. The replies rarely concede anything to a biting critique, and it is almost unheard of for a rejoinder to admit that the original critique was misguided or erroneous in any way. On a few occasions I have responded to criticisms that I thought were grossly misleading, because a failure to respond can be interpreted as conceding error, but I have never found the hostile exchanges instructive. In search of another way to deal with disagreements, I have engaged in a few "adversarial collaborations," in which scholars who disagree on the science agree to write a jointly authored paper on their differences, and sometimes conduct research together. In especially tense situations, the research is moderated by an arbiter.

My most satisfying and productive adversarial collaboration was with Gary Klein, the intellectual leader of an association of scholars and practitioners who do not like the kind of work I do. They call themselves students of Naturalistic Decision Making, or NDM, and mostly work in organizations where the"0%ЉЪ ty often study how experts work. The N DMers adamantly reject the focus on biases in the heuristics and biases approach. They criticize this model as overly concerned with failures and driven by artificial experiments rather than by the study of real people doing things that matter. They are deeply skeptical about the value of using rigid algorithms to replace human judgment, and Paul Meehl is not among their heroes. Gary Klein has eloquently articulated this position over many years.

This is hardly the basis for a beautiful friendship, but there is more to the story. I had never believed that intuition is always misguided. I had also been a fan of Klein's studies of expertise in firefighters since I first saw a draft of a paper he wrote in the 1970s, and was impressed by his book *Sources of Power*, much of which analyzes how experienced professionals develop intuitive skills. I invited him to join in an effort to map the boundary that separates the marvels of intuition from its flaws.

He was intrigued by the idea and we went ahead with the project—with no certainty that it would succeed. We set out to answer a specific question: When can you trust an experienced professional who claims to have an intuition? It was obvious that Klein would be more disposed to be trusting, and I would be more skeptical. But could we agree on principles for answering the general question?

Over seven or eight years we had many discussions, resolved many disagreements, almost blew up more than once, wrote many draft s, became friends, and eventually published a joint article with a title that tells the story: "Conditions for Intuitive Expertise: A Failure to Disagree." Indeed, we did not encounter real issues on which we disagreed—but we did not really agree.

Marvels and Flaws

Malcolm Gladwell's bestseller *Blink* appeared while Klein and I were working on the project, and it was reassuring to find ourselves in agreement about it. Gladwell's book opens with the memorable story of art experts faced with an object that is described as a magnificent example of a kouros, a sculpture of a striding boy. Several of the experts had strong visceral reactions: they felt in their gut that the statue was a fake but were not able to articulate what it was about it that made them uneasy. Everyone who read the book—millions did—remembers that story as a triumph of intuition. The experts agreed that they knew the sculpture was a fake without knowing how they knew—the very definition of intuition. The story appears to imply that a systematic search for the cue that guided the experts would have failed, but Klein and I both rejected that conclusion. From our point of view, such an inquiry was needed, and if it had been conducted properly (which Klein knows how to do), it would probably have succeeded.

Although many readers of the kouros example were surely drawn to an almost magical view of expert intuition, Gladwell himself does not hold that position. In a later chapter he describes a massive failure of intuition: Americans elected President Harding, whose only qualification for the position was that he perfectly looked the part. Square jawed and tall, he was the perfect image of a strong and decisive leader. People voted for someone who looked strong and decisive without any other reason to believe that he was. An intuitive prediction of how Harding would perform as president arose from substituting one question for another. A reader of this book should expect such an intuition to be held with confidence.

Intuition as Recognition

The early experiences that shaped Klein's views of intuition were starkly different from mine. My thinking was formed by observing the illusion of validity in myself and by reading Paul Meehl's demonstrations of the inferiority of clinical prediction. In contrast, Klein's views were shaped by his early studies of fireground commanders (the leaders of firefighting teams). He followed them as they fought fires and later interviewed the leader about his thoughts as he made decisions. As Klein described it in our joint article, he and his collaborators

> investigated how the commanders could make good decisions without comparing options. The initial hypothesis was that commanders would restrict their analysis to only a pair of options, but that hypothesis proved to be incorrect. In fact, the commanders usually generated only a single option, and that was all they needed. They could draw on the repertoire of patterns that they had compiled during more than a decade of both real and virtual experience to identify a plausible option, which they considered first. They evaluated this option by mentally simulating it to see if it would work in the situation they were facing…. If the course of action they were considering seemed appropriate, they would implement it. If it had shortcomings, they would modify it. If they could not easily modify it, they would turn to the next most plausible option and run through the same procedure until an acceptable course of action was found.

Klein elaborated this description into a theory of decision making that he called the recognition-primed decision (RPD) model, which applies to firefighters but also describes expertise in other domains, including chess. The process involves both System 1 and System 2. In the first phase, a tentative plan comes to mind by an automatic function of associative memory—System 1. The next phase is a deliberate process in which the plan is mentally simulated to check if it will work—an operation of System 2. The model of intuitive decision making as pattern recognition develops ideas presented some time ago by

Herbert Simon, perhaps the only scholar who is recognized and admired as a hero and founding figure by all the competing clans and tribes in the study of decision making. I quoted Herbert Simon's definition of intuition in the introduction, but it will make more sense when I repeat it now: "The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition."

This strong statement reduces the apparent magic of intuition to the everyday experience of memory. We marvel at the story of the firefighter who has a sudden urge to escape a burning house just before it collapses, because the firefighter knows the danger intuitively, "without knowing how he knows." However, we also do not know how we immediately know that a person we see as we enter a room is our friend Peter. The moral of Simon's remark is that the mystery of knowing without knowing is not a distinctive feature of intuition; it is the norm of mental life.

Acquiring Skill

How does the information that supports intuition get "stored in memory"? Certain types of intuitions are acquired very quickly. We have inherited from our ancestors a great facility to learn when to be afraid. Indeed, one experience is often sufficient to establish a long-term aversion and fear. Many of us have the visceral memory of a single dubious dish tto hat still leaves us vaguely reluctant to return to a restaurant. All of us tense up when we approach a spot in which an unpleasant event occurred, even when there is no reason to expect it to happen again. For me, one such place is the ramp leading to the San Francisco airport, where years ago a driver in the throes of road rage followed me from the freeway, rolled down his window, and hurled obscenities at me. I never knew what caused his hatred, but I remember his voice whenever I reach that point on my way to the airport.

My memory of the airport incident is conscious and it fully explains the emotion that comes with it. On many occasions, however, you may feel uneasy in a particular place or when someone uses a particular turn of phrase without having a conscious memory of the triggering event. In hindsight, you will label that unease an intuition if it is followed by a bad experience. This mode of emotional learning is closely related to what happened in Pavlov's famous conditioning experiments, in which the dogs learned to recognize the sound of the bell as a signal that food was coming. What Pavlov's dogs learned can be described as a learned hope. Learned fears are even more easily acquired.

Fear can also be learned—quite easily, in fact—by words rather than by experience. The fireman who had the "sixth sense" of danger had certainly had many occasions to discuss and think about types of fires he was not involved in, and to rehearse in his mind what the cues might be and how he should react. As I remember from experience, a young platoon commander with no experience of combat will tense up while leading troops through a narrowing ravine, because he was taught to identify the terrain as favoring an ambush. Little repetition is needed for learning.

Emotional learning may be quick, but what we consider as "expertise" usually takes a long time to develop. The acquisition of expertise in complex tasks such as high-level chess, professional basketball, or firefighting is intricate and slow because expertise in a domain is not a single skill but rather a large collection of miniskills. Chess is a good example. An expert player can understand a complex position at a glance, but it takes years to develop that level of ability. Studies of chess masters have shown that at least 10,000 hours of dedicated practice (about 6 years of playing chess 5 hours a day) are required to attain the highest levels of performance. During those hours of intense concentration, a serious chess player becomes familiar with thousands of configurations, each consisting of an arrangement of related pieces that can threaten or defend each other.

Learning high-level chess can be compared to learning to read. A first grader works hard at recognizing individual letters and assembling them into syllables and words, but a good adult reader perceives entire clauses. An expert reader has also acquired the ability to assemble familiar elements in a new pattern and can quickly "recognize" and correctly pronounce a word that she has never seen before. In chess, recurrent patterns of interacting pieces play the role of letters, and a chess position is a long word or a sentence.

A skilled reader who sees it for the first time will be able to read the opening stanza of Lewis Carroll's "Jabberwocky" with perfect rhythm and intonation, as well as pleasure:

’Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Acquiring expertise in chess is harder and slower than learning to read because there are many more letters in the “alphabet” of chess and because the “words” consist of many letters. After thousands of hours of practice, however, chess masters are able to read a chess situation at a glance. The few moves that come to their mind are almost always strong and sometimes creative. They can deal with a “word” they have never encountered, and they can find a new way to interpret a familiar one.

The Environment of Skill

Klein and I quickly found that we agreed both on the nature of intuitive skill and on how it is acquired. We still needed to agree on our key question: When can you trust a self-confident professional who claims to have an intuition?

We eventually concluded that our disagreement was due in part to the fact that we had different experts in mind. Klein had spent much time with fireground commanders, clinical nurses, and other professionals who have real expertise. I had spent more time thinking about clinicians, stock pickers, and political scientists trying to make unsupportable long-term forecasts. Not surprisingly, his default attitude was trust and respect; mine was skepticism. He was more willing to trust experts who claim an intuition because, as he told me, true experts know the limits of their knowledge. I argued that there are many pseudo-experts who have no idea that they do not know what they are doing (the illusion of validity), and that as a general proposition subjective confidence is commonly too high and often uninformative.

Earlier I traced people’s confidence in a belief to two related impressions: cognitive ease and coherence. We are confident when the story we tell ourselves comes easily to mind, with no contradiction and no competing scenario. But ease and coherence do not guarantee that a belief held with confidence is true. The associative machine is set to suppress doubt and to evoke ideas and information that are compatible with the currently dominant story. A mind that follows WY SIATI will achieve high confidence much too easily by ignoring what it does not know. It is therefore not surprising that many of us are prone to have high confidence in unfounded intuitions. Klein and I eventually agreed on an important principle: the confidence that people have in their intuitions is not a reliable guide to their validity. In other words, do not trust anyone—including yourself—to tell you how much you should trust their judgment.

If subjective confidence is not to be trusted, how can we evaluate the probable validity of an intuitive judgment? When do judgments reflect true expertise? When do they display an illusion of validity? The answer comes from the two basic conditions for acquiring a skill:

- an environment that is sufficiently regular to be predictable
- an opportunity to learn these regularities through prolonged practice

When both these conditions are satisfied, intuitions are likely to be skilled. Chess is an extreme example of a regular environment, but bridge and poker also provide robust statistical regularities that can support skill. Physicians, nurses, athletes, and firefighters also face complex but fundamentally orderly situations. The accurate intuitions that Gary Klein has described are due to highly valid cues that es the expert's System 1 has learned to use, even if System 2 has not learned to name them. In contrast, stock pickers and political scientists who make long-term forecasts operate in a zero-validity environment. Their failures reflect the basic unpredictability of the events that they try to forecast.

Some environments are worse than irregular. Robin Hogarth described "wicked" environments, in which professionals are likely to learn the wrong lessons from experience. He borrows from Lewis Thomas the example of a physician in the early twentieth century who often had intuitions about patients who were about to develop typhoid. Unfortunately, he tested his hunch by palpating the patient's tongue, without washing his hands between patients. When patient after patient became ill, the physician developed a sense of clinical infallibility. His predictions were accurate—but not because he was exercising professional intuition!

Meehl's clinicians were not inept and their failure was not due to lack of talent. They performed poorly because they were assigned tasks that did not have a simple solution. The clinicians' predicament was less extreme than the zero-validity environment of long-term political forecasting, but they operated in low-validity situations that did not allow high accuracy. We know this to be the case because the best statistical algorithms, although more accurate than human judges, were never very accurate. Indeed, the studies by Meehl and his followers never produced a "smoking gun" demonstration, a case in which clinicians completely missed a highly valid cue that the algorithm detected. An extreme failure of this kind is unlikely because human learning is normally efficient. If a strong predictive cue exists, human observers will find it, given a decent opportunity to do so. Statistical algorithms greatly outdo humans in noisy environments for two reasons: they are more likely than human judges to detect weakly valid cues and much more likely to maintain a modest level of accuracy by using such cues consistently.

It is wrong to blame anyone for failing to forecast accurately in an unpredictable world. However, it seems fair to blame professionals for believing they can succeed in an impossible task. Claims for correct intuitions in an unpredictable situation are self-delusional at best, sometimes worse. In the absence of valid cues, intuitive "hits" are due either to luck or to lies. If you find this conclusion surprising, you still have a lingering belief that intuition is magic. Remember this rule: intuition cannot be trusted in the absence of stable regularities in the environment.

Feedback and Practice

Some regularities in the environment are easier to discover and apply than others. Think of how you developed your style of using the brakes on your car. As you were mastering the skill of taking curves, you gradually learned when to let go of the accelerator and when and how hard to use the brakes. Curves differ, and the variability you experienced while learning ensures that you are now ready to brake at the right time and strength for any curve you encounter. The conditions for learning this skill are ideal, because you receive immediate and unambiguous feedback every time you go around a bend: the mild reward of a comfortable turn or the mild punishment of some difficulty in handling the car if you brake either too hard or not quite hard enough. The situations that face a harbor pilot maneuvering large ships are no less regular, but skill is much more difficult to acquire by sheer experience because of the long delay between actions and their manoticeable outcomes. Whether professionals have a chance to develop intuitive expertise depends essentially on the quality and speed of feedback, as well as on sufficient opportunity to practice.

Expertise is not a single skill; it is a collection of skills, and the same professional may be highly expert in some of the tasks in her domain while remaining a novice in others. By the time chess players become experts, they have "seen everything" (or almost everything), but chess is an exception in this regard. Surgeons can be much more proficient in some operations than in others. Furthermore, some aspects of any professional's tasks are much easier to learn than others. Psychotherapists have many opportunities to observe the immediate reactions of patients to what they say. The feedback enables them to develop the intuitive skill to find the words and the tone that will calm anger, forge confidence, or focus the patient's attention. On the other hand, therapists do not have a chance to identify which general treatment approach is most suitable for different patients. The feedback they receive from their patients' long-term outcomes is sparse, delayed, or (usually) nonexistent, and in any case too ambiguous to support learning from experience.

Among medical specialties, anesthesiologists benefit from good feedback, because the effects of their actions are likely to be quickly evident. In contrast, radiologists obtain little information about the accuracy of the diagnoses they make and about the pathologies they fail to detect. Anesthesiologists are therefore in a better position to develop useful intuitive skills. If an anesthesiologist says, "I have a feeling something is wrong," everyone in the operating room should be prepared for an emergency.

Here again, as in the case of subjective confidence, the experts may not know the limits of their expertise. An experienced psychotherapist knows that she is skilled in working out what is going on in her patient's mind and that she has good intuitions about what the patient will say next. It is tempting for her to conclude that she can also anticipate how well the patient will do next year, but this conclusion is not equally justified. Short-term anticipation and long-term forecasting are different tasks, and the therapist has had adequate opportunity to learn one but not the other. Similarly, a financial expert may have skills in many aspects of his trade but not in picking stocks, and an expert in the Middle East knows many things but not the future. The clinical psychologist, the stock picker, and the pundit do have intuitive skills in some of their tasks, but they have not learned to identify the situations and the tasks in which intuition will betray them. The unrecognized limits of professional skill help explain why experts are often overconfident.

Evaluating Validity

At the end of our journey, Gary Klein and I agreed on a general answer to our initial question: When can you trust an experienced professional who claims to have an intuition? Our conclusion was that for the most part it is possible to distinguish intuitions that are likely to be valid from those that are likely to be bogus. As in the judgment of whether a work of art is genuine or a fake, you will usually do better by focusing on its provenance than by looking at the piece itself. If the environment is sufficiently regular and if the judge has had a chance to learn its regularities, the associative machinery will recognize situations and generate quick and accurate predictions and decisions. You can trust someone's intuitions if these conditions are met.

Unfortunately, associativentu memory also generates subjectively compelling intuitions that are false. Anyone who has watched the chess progress of a talented youngster knows well that skill does not become perfect all at once, and that on the way to near perfection some mistakes are made with great confidence. When evaluating expert intuition you should always consider whether there was an adequate opportunity to learn the cues, even in a regular environment.

In a less regular, or low-validity, environment, the heuristics of judgment are invoked. System 1 is often able to produce quick answers to difficult questions by substitution, creating coherence where there is none. The question that is answered is not the one that was intended, but the answer is produced quickly and may be sufficiently plausible to pass the lax and lenient review of System 2. You may want to forecast the commercial future of a company, for example, and believe that this is what you are judging, while in fact your evaluation is dominated by your impressions of the energy and competence of its current executives. Because substitution occurs automatically, you often do not know the origin of a judgment that you (your System 2) endorse and adopt. If it is the only one that comes to mind, it may be subjectively undistinguishable from valid judgments that you make with expert confidence. This is why subjective confidence is not a good diagnostic of accuracy: judgments that answer the wrong question can also be made with high confidence.

You may be asking, Why didn't Gary Klein and I come up immediately with the idea of evaluating an expert's intuition by assessing the regularity of the environment and the expert's learning history—mostly setting aside the expert's confidence? And what did we think the answer could be? These are good questions because the contours of the solution were apparent from the beginning. We knew at the outset that fireground commanders and pediatric nurses would end up on one side of the boundary of valid intuitions and that the specialties studied by Meehl would be on the other, along with stock pickers and pundits.

It is difficult to reconstruct what it was that took us years, long hours of discussion, endless exchanges of draft s and hundreds of e-mails negotiating over words, and more than once almost giving up. But this is what always happens when a project ends reasonably well: once you understand the main conclusion, it seems it was always obvious.

As the title of our article suggests, Klein and I disagreed less than we had expected and accepted joint solutions of almost all the substantive issues that were raised. However, we also found that our early differences were more than an intellectual disagreement. We had different attitudes, emotions, and tastes, and those changed remarkably little over the years. This is most obvious in the facts that we find amusing and interesting. Klein still winces when the word *bias* is mentioned, and he still enjoys stories in which algorithms or formal procedures lead to obviously absurd decisions. I tend to view the occasional failures of algorithms as opportunities to improve them. On the other hand, I find more pleasure than Klein does in the come-uppance of arrogant experts who claim intuitive powers in zero-validity situations. In the long run, however, finding as much intellectual agreement as we did is surely more important than the persistent emotional differences that remained.

Speaking of Expert Intuition

“How much expertise does she have in this particular task? How much practice has she had?”

“Does he really believe that the environment of start-ups is sufficiently regular to justify an intuition that goes against the base rates?”

“She is very confident in her decision, but subjective confidence is a poor index of the accuracy of a judgment.”

“Did he really have an opportunity to learn? How quick and how clear was the feedback he received on his judgments?”

The Outside View

A few years after my collaboration with Amos began, I convinced some officials in the Israeli Ministry of Education of the need for a curriculum to teach judgment and decision making in high schools. The team that I assembled to design the curriculum and write a textbook for it included several experienced teachers, some of my psychology students, and Seymour Fox, then dean of the Hebrew University's School of Education, who was an expert in curriculum development.

After meeting every Friday afternoon for about a year, we had constructed a detailed outline of the syllabus, had written a couple of chapters, and had run a few sample lessons in the classroom. We all felt that we had made good progress. One day, as we were discussing procedures for estimating uncertain quantities, the idea of conducting an exercise occurred to me. I asked everyone to write down an estimate of how long it would take us to submit a finished draft of the textbook to the Ministry of Education. I was following a procedure that we already planned to incorporate into our curriculum: the proper way to elicit information from a group is not by starting with a public discussion but by confidentially collecting each person's judgment. This procedure makes better use of the knowledge available to members of the group than the common practice of open discussion. I collected the estimates and jotted the results on the blackboard. They were narrowly centered around two years; the low end was one and a half, the high end two and a half years.

Then I had another idea. I turned to Seymour, our curriculum expert, and asked whether he could think of other teams similar to ours that had developed a curriculum from scratch. This was a time when several pedagogical innovations like "new math" had been introduced, and Seymour said he could think of quite a few. I then asked whether he knew the history of these teams in some detail, and it turned out that he was familiar with several. I asked him to think of these teams when they had made as much progress as we had. How long, from that point, did it take them to finish their textbook projects?

He fell silent. When he finally spoke, it seemed to me that he was blushing, embarrassed by his own answer: "You know, I never realized this before, but in fact not all the teams at a stage comparable to ours ever did complete their task. A substantial fraction of the teams ended up failing to finish the job."

This was worrisome; we had never considered the possibility that we might fail. My anxiety rising, I asked how large he estimated that fraction was. Rw I⍰ sidering t20;About 40%," he answered. By now, a pall of gloom was falling over the room. The next question was obvious: "Those who finished," I asked. "How long did it take them?" "I cannot think of any group that finished in less than seven years," he replied, "nor any that took more than ten."

I grasped at a straw: "When you compare our skills and resources to those of the other groups, how good are we? How would you rank us in comparison with these teams?" Seymour did not hesitate long this time. "We're below average," he said, "but not by much." This came as a complete surprise to all of us—including Seymour, whose prior estimate had been well within the optimistic consensus of the group. Until I prompted him, there was no connection in his mind between his knowledge of the history of other teams and his forecast of our future.

Our state of mind when we heard Seymour is not well described by stating what we "knew." Surely all of us "knew" that a minimum of seven years and a 40% chance of failure was a more plausible forecast of the fate of our project than the numbers we had written on our slips of paper a few minutes earlier. But we did not acknowledge what we knew. The new forecast still seemed unreal, because we could not imagine how it could take so long to finish a project that looked so manageable. No crystal ball was available to tell us the strange sequence of unlikely events that were in our future. All we could see was a reasonable plan that should produce a book in about two years, conflicting with statistics indicating that other teams had failed or had taken an absurdly long time to complete their mission. What we had heard was base-rate information, from which we should have inferred a causal story: if so many teams failed, and if those that succeeded took so long, writing a curriculum was surely much harder than we had thought. But such an inference would have conflicted with our direct experience of the good progress we had been making. The statistics that Seymour provided were treated as base rates normally are—noted and promptly set aside.

We should have quit that day. None of us was willing to invest six more years of work in a project with a 40% chance of failure. Although we must have sensed that persevering was not reasonable, the warning did not provide an immediately compelling reason to quit. After a few minutes of desultory debate, we gathered ourselves together and carried on as if nothing had happened. The book was eventually completed eight(!) years later. By that time I was no longer living in Israel and had long since ceased to be part of the team, which completed the task after many unpredictable vicissitudes. The initial enthusiasm for the idea in the Ministry of Education had waned by the time the text was delivered and it was never used.

This embarrassing episode remains one of the most instructive experiences of my professional life. I eventually learned three lessons from it. The first was immediately apparent: I had stumbled onto a distinction between two profoundly different approaches to forecasting, which Amos and I later labeled the inside view and the outside view. The second lesson was that our initial forecasts of about two years for the completion of the project exhibited a planning fallacy. Our estimates were closer to a best-case scenario than to a realistic assessment. I was slower to accept the third lesson, which I call irrational perseverance: the folly we displayed that day in failing to abandon the project. Facing a choice, we gave up rationality rather than give up the enterprise.

Drawn to the Inside View

On that long-ago Friday, our curriculum expert made two judgments about the same problem and arrived at very different answers. The *inside view* is the one that all of us, including Seymour, spontaneously adopted to assess the future of our project. We focused on our specific circumstances and searched for evidence in our own experiences. We had a sketchy plan: we knew how many chapters we were going to write, and we had an idea of how long it had taken us to write the two that we had already done. The more cautious among us probably added a few months to their estimate as a margin of error.

Extrapolating was a mistake. We were forecasting based on the information in front of us—WYSIATI—but the chapters we wrote first were probably easier than others, and our commitment to the project was probably then at its peak. But the main problem was that we failed to allow for what Donald Rumsfeld famously called the “unknown unknowns.” There was no way for us to foresee, that day, the succession of events that would cause the project to drag out for so long. The divorces, the illnesses, the crises of coordination with bureaucracies that delayed the work could not be anticipated. Such events not only cause the writing of chapters to slow down, they also produce long periods during which little or no progress is made at all. The same must have been true, of course, for the other teams that Seymour knew about. The members of those teams were also unable to imagine the events that would cause them to spend seven years to finish, or ultimately fail to finish, a project that they evidently had thought was very feasible. Like us, they did not know the odds they were facing. There are many ways for any plan to fail, and although most of them are too improbable to be anticipated, the likelihood that *something* will go wrong in a big project is high.

The second question I asked Seymour directed his attention away from us and toward a class of similar cases. Seymour estimated the base rate of success in that reference class: 40% failure and seven to ten years for completion. His informal survey was surely not up to scientific standards of evidence, but it provided a reasonable basis for a baseline prediction: the prediction you make about a case if you know nothing except the category to which it belongs. As we saw earlier, the *baseline prediction* should be the anchor for further adjustments. If you are asked to guess the height of a woman about whom you know only that she lives in New York City, your baseline prediction is your best guess of the average height of women in the city. If you are now given case-specific information, for example that the woman's son is the starting center of his high school basketball team, you will adjust your estimate away from the mean in the appropriate direction. Seymour's comparison of our team to others suggested that the forecast of our outcome was slightly worse than the baseline prediction, which was already grim.

The spectacular accuracy of the outside-view forecast in our problem was surely a fluke and should not count as evidence for the validity of the *outside view*. The argument for the outside view should be made on general grounds: if the reference class is properly chosen, the outside view will give an indication of where the ballpark is, and it may suggest, as it did in our case, that the inside-view forecasts are not even close to it.

For a psychologist, the discrepancy between Seymour's two judgments is striking. He had in his head all the knowledge required to estimate the statistics of an appropriate reference class, but he reached his initial estimate without ever using that knowledge.

Seymour's forecast from his insidethaa view was not an adjustment from the baseline prediction, which had not come to his mind. It was based on the particular circumstances of our efforts. Like the participants in the Tom W experiment, Seymour knew the relevant base rate but did not think of applying it.

Unlike Seymour, the rest of us did not have access to the outside view and could not have produced a reasonable baseline prediction. It is noteworthy, however, that we did not feel we needed information about other teams to make our guesses. My request for the outside view surprised all of us, including me! This is a common pattern: people who have information about an individual case rarely feel the need to know the statistics of the class to which the case belongs.

When we were eventually exposed to the outside view, we collectively ignored it. We can recognize what happened to us; it is similar to the experiment that suggested the futility of teaching psychology. When they made predictions about individual cases about which they had a little information (a brief and bland interview), Nisbett and Borgida's students completely neglected the global results they had just learned. "Pallid" statistical information is routinely discarded when it is incompatible with one's personal impressions of a case. In the competition with the inside view, the outside view doesn't stand a chance.

The preference for the inside view sometimes carries moral overtones. I once asked my cousin, a distinguished lawyer, a question about a reference class: "What is the probability of the defendant winning in cases like this one?" His sharp answer that "every case is unique" was accompanied by a look that made it clear he found my question inappropriate and superficial. A proud emphasis on the uniqueness of cases is also common in medicine, in spite of recent advances in evidence-based medicine that point the other way. Medical statistics and baseline predictions come up with increasing frequency in conversations between patients and physicians. However, the remaining ambivalence about the outside view in the medical profession is expressed in concerns about the impersonality of procedures that are guided by statistics and checklists.

The Planning Fallacy

In light of both the outside-view forecast and the eventual outcome, the original estimates we made that Friday afternoon appear almost delusional. This should not come as a surprise: overly optimistic forecasts of the outcome of projects are found everywhere. Amos and I coined the term *planning fallacy* to describe plans and forecasts that

- are unrealistically close to best-case scenarios
- could be improved by consulting the statistics of similar cases

Examples of the planning fallacy abound in the experiences of individuals, governments, and businesses. The list of horror stories is endless.

- In July 1997, the proposed new Scottish Parliament building in Edinburgh was estimated to cost up to £40 million. By June 1999, the budget for the building was £109 million. In April 2000, legislators imposed a £195 million “cap on costs.” By November 2001, they demanded an estimate of “final cost,” which was set at £241 million. That estimated final cost rose twice in 2002, ending the year at £294.6 million. It rose three times more in 2003, reaching £375.8 million by June. The building was finally comanspleted in 2004 at an ultimate cost of roughly £431 million.
- A 2005 study examined rail projects undertaken worldwide between 1969 and 1998. In more than 90% of the cases, the number of passengers projected to use the system was overestimated. Even though these passenger shortfalls were widely publicized, forecasts did not improve over those thirty years; on average, planners overestimated how many people would use the new rail projects by 106%, and the average cost overrun was 45%. As more evidence accumulated, the experts did not become more reliant on it.
- In 2002, a survey of American homeowners who had remodeled their kitchens found that, on average, they had expected the job to cost $18,658; in fact, they ended up paying an average of $38,769.

The optimism of planners and decision makers is not the only cause of overruns. Contractors of kitchen renovations and of weapon systems readily admit (though not to their clients) that they routinely make most of their profit on additions to the original plan. The failures of forecasting in these cases reflect the customers' inability to imagine how much their wishes will escalate over time. They end up paying much more than they would if they had made a realistic plan and stuck to it.

Errors in the initial budget are not always innocent. The authors of unrealistic plans are often driven by the desire to get the plan approved —whether by their superiors or by a client—supported by the knowledge that projects are rarely abandoned unfinished merely because of overruns in costs or completion times. In such cases, the greatest responsibility for avoiding the planning fallacy lies with the decision makers who approve the plan. If they do not recognize the need for an outside view, they commit a planning fallacy.

Mitigating the Planning Fallacy

The diagnosis of and the remedy for the planning fallacy have not changed since that Friday afternoon, but the implementation of the idea has come a long way. The renowned Danish planning expert Bent Flyvbjerg, now at Oxford University, offered a forceful summary:

> The prevalent tendency to underweight or ignore distributional information is perhaps the major source of error in forecasting. Planners should therefore make every effort to frame the forecasting problem so as to facilitate utilizing all the distributional information that is available.

This may be considered the single most important piece of advice regarding how to increase accuracy in forecasting through improved methods. Using such distributional information from other ventures similar to that being forecasted is called taking an "outside view" and is the cure to the planning fallacy.

The treatment for the planning fallacy has now acquired a technical name, *reference class forecasting*, and Flyvbjerg has applied it to transportation projects in several countries. The outside view is implemented by using a large database, which provides information on both plans and outcomes for hundreds of projects all over the world, and can be used to provide statistical information about the likely overruns of cost and time, and about the likely underperformance of projects of different types.

The forecasting method that Flyvbjerg applies is similar to the practices recommended for overcoming base-rate neglect:

1. Identify an appropriate reference class (kitchen renovations, large railway projects, etc.).
2. Obtain the statistics of the reference class (in terms of cost per mile of railway, or of the percentage by which expenditures exceeded budget). Use the statistics to generate a baseline prediction.
3. Use specific information about the case to adjust the baseline prediction, if there are particular reasons to expect the optimistic bias to be more or less pronounced in this project than in others of the same type.

Flyvbjerg's analyses are intended to guide the authorities that commission public projects, by providing the statistics of overruns in similar projects. Decision makers need a realistic assessment of the costs and benefits of a proposal before making the final decision to approve it. They may also wish to estimate the budget reserve that they need in anticipation of overruns, although such precautions often become self-fulfilling prophecies. As one official told Flyvbjerg, "A budget reserve is to contractors as red meat is to lions, and they will devour it."

Organizations face the challenge of controlling the tendency of executives competing for resources to present overly optimistic plans. A well-run organization will reward planners for precise execution and penalize them for failing to anticipate difficulties, and for failing to allow for difficulties that they could not have anticipated—the unknown unknowns.

Decisions and Errors

That Friday afternoon occurred more than thirty years ago. I often thought about it and mentioned it in lectures several times each year. Some of my friends got bored with the story, but I kept drawing new lessons from it. Almost fifteen years after I first reported on the planning fallacy with Amos, I returned to the topic with Dan Lovallo. Together we sketched a theory of decision making in which the optimistic bias is a significant source of risk taking. In the standard rational model of economics, people take risks because the odds are favorable—they accept some probability of a costly failure because the probability of success is sufficient. We proposed an alternative idea.

When forecasting the outcomes of risky projects, executives too easily fall victim to the planning fallacy. In its grip, they make decisions based on delusional optimism rather than on a rational weighting of gains, losses, and probabilities. They overestimate benefits and underestimate costs. They spin scenarios of success while overlooking the potential for mistakes and miscalculations. As a result, they pursue initiatives that are unlikely to come in on budget or on time or to deliver the expected returns—or even to be completed.

In this view, people often (but not always) take on risky projects because they are overly optimistic about the odds they face. I will return to this idea several times in this book—it probably contributes to an explanation of why people litigate, why they start wars, and why they open small businesses.

Failing a Test

For many years, I thought that the main point of the curriculum story was what I had learned about my friend Seymour: that his best guess about the future of our project was not informed by what he knew about similar projects. I came off quite well in my telling of the story, ir In which I had the role of clever questioner and astute psychologist. I only recently realized that I had actually played the roles of chief dunce and inept leader.

The project was my initiative, and it was therefore my responsibility to ensure that it made sense and that major problems were properly discussed by the team, but I failed that test. My problem was no longer the planning fallacy. I was cured of that fallacy as soon as I heard Seymour's statistical summary. If pressed, I would have said that our earlier estimates had been absurdly optimistic. If pressed further, I would have admitted that we had started the project on faulty premises and that we should at least consider seriously the option of declaring defeat and going home. But nobody pressed me and there was no discussion; we tacitly agreed to go on without an explicit forecast of how long the effort would last. This was easy to do because we had not made such a forecast to begin with. If we had had a reasonable baseline prediction when we started, we would not have gone into it, but we had already invested a great deal of effort—an instance of the sunk-cost fallacy, which we will look at more closely in the next part of the book. It would have been embarrassing for us—especially for me—to give up at that point, and there seemed to be no immediate reason to do so. It is easier to change directions in a crisis, but this was not a crisis, only some new facts about people we did not know. The outside view was much easier to ignore than bad news in our own effort. I can best describe our state as a form of lethargy—an unwillingness to think about what had happened. So we carried on. There was no further attempt at rational planning for the rest of the time I spent as a member of the team—a particularly troubling omission for a team dedicated to teaching rationality. I hope I am wiser today, and I have acquired a habit of looking for the outside view. But it will never be the natural thing to do.

Speaking of the Outside View

> "He's taking an inside view. He should forget about his own case and look for what happened in other cases."

> "She is the victim of a planning fallacy. She's assuming a best-case scenario, but there are too many different ways for the plan to fail, and she cannot foresee them all."

> "Suppose you did not know a thing about this particular legal case, only that it involves a malpractice claim by an individual against a surgeon. What would be your baseline prediction? How many of these cases succeed in court? How many settle? What are the

amounts? Is the case we are discussing stronger or weaker than similar claims?"

"We are making an additional investment because we do not want to admit failure. This is an instance of the sunk-cost fallacy."

Frames and Reality

Italy and France competed in the 2006 final of the World Cup. The next two sentences both describe the outcome: “Italy won.” “France lost.” Do those statements have the same meaning? The answer depends entirely on what you mean by *meaning*.

For the purpose of logical reasoning, the two descriptions of the outcome of the match are interchangeable because they designate the same state of the world. As philosophers say, their truth conditions are identical: if one of these sentences is true, then the other is true as well. This is how Econs understand things. Their beliefs and preferences are reality-bound. In particular, the objects of their choices are states of the world, which are not affected by the words chosen to describe them.

There is another sense of *meaning*, in which “Italy won” and “France lost” do not have the same meaning at all. In this sense, the meaning of a sentence is what happens in your associative machinery while you understand it. The two sentences evoke markedly different associations. “Italy won” evokes thoughts of the Italian team and what it did to win. “France lost” evokes thoughts of the French team and what it did that caused it to lose, including the memorable head butt of an Italian player by the French star Zidane. In terms of the associations they bring to mind—how System 1 reacts to them—the two sentences really “mean” different things. The fact that logically equivalent statements evoke different reactions makes it impossible for Humans to be as reliably rational as Econs.

Emotional Framing

Amos and I applied the label of framing effects to the unjustified influences of formulation on beliefs an Con d preferences. This is one of the examples we used:

> Would you accept a gamble that offers a 10% chance to win $95 and a 90% chance to lose $5?

> Would you pay $5 to participate in a lottery that offers a 10% chance to win $100 and a 90% chance to win nothing?

First, take a moment to convince yourself that the two problems are identical. In both of them you must decide whether to accept an uncertain prospect that will leave you either richer by $95 or poorer by $5. Someone whose preferences are reality-bound would give the same answer to both questions, but such individuals are rare. In fact, one version attracts many more positive answers: the second. A bad outcome is much more acceptable if it is framed as the cost of a lottery ticket that did not win than if it is simply described as losing a gamble. We should not be surprised: *losses* evokes stronger negative feelings than *costs*. Choices are not reality-bound because System 1 is not reality-bound.

The problem we constructed was influenced by what we had learned from Richard Thaler, who told us that when he was a graduate student he had pinned on his board a card that said costs are not losses. In his early essay on consumer behavior, Thaler described the debate about whether gas stations would be allowed to charge different prices for purchases paid with cash or on credit. The credit-card lobby pushed hard to make differential pricing illegal, but it had a fallback position: the difference, if allowed, would be labeled a cash discount, not a credit surcharge. Their psychology was sound: people will more readily forgo a discount than pay a surcharge. The two may be economically equivalent, but they are not emotionally equivalent.

In an elegant experiment, a team of neuroscientists at University College London combined a study of framing effects with recordings of activity in different areas of the brain. In order to provide reliable measures of the brain response, the experiment consisted of many trials. Figure 14 illustrates the two stages of one of these trials.

First, the subject is asked to imagine that she received an amount of money, in this example £50.

The subject is then asked to choose between a sure outcome and a gamble on a wheel of chance. If the wheel stops on white she “receives” the entire amount; if it stops on black she gets nothing. The sure outcome is simply the expected value of the gamble, in this case a gain of £20.

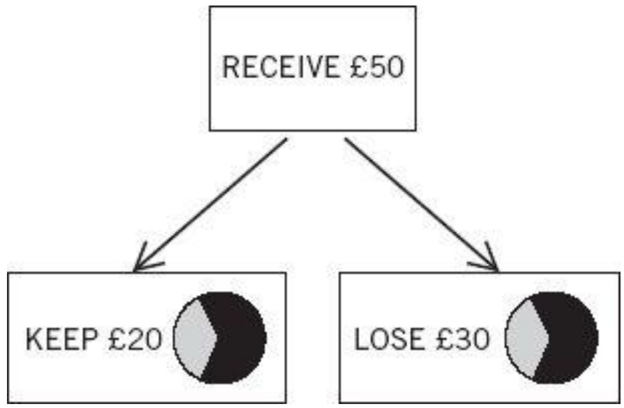

Figure 14

As shown, the same sure outcome can be framed in two different ways: as KEEP £20 or as LOSE £30. The objective outcomes are precisely identical in the two frames, and a reality-bound Econ would respond to both in the same way—selecting either the sure thing or the gamble regardless of the frame—but we already know that the Human mind is not bound to reality. Tendencies to approach or avoid are evoked by the words, and we expect System 1 to be biased in favor of the sure option when it is designated as KEEP and against that same option when it is designated as LOSE.

The experiment consisted of many trials, and each participant encountere Bon p>

The activity of the brain was recorded as the subjects made each decision. Later, the trials were separated into two categories:

1 Trials on which the subject's choice conformed to the frame

- preferred the sure thing in the KEEP version
- preferred the gamble in the LOSS version

2 Trials in which the choice did not conform to the frame.

The remarkable results illustrate the potential of the new discipline of neuroeconomics—the study of what a person's brain does while he makes decisions. Neuroscientists have run thousands of such experiments, and they have learned to expect particular regions of the brain to "light up"—indicating increased flow of oxygen, which suggests heightened neural activity—depending on the nature of the task. Different regions are active when the individual attends to a visual object, imagines kicking a ball, recognizes a face, or thinks of a house. Other regions light up when the individual is emotionally aroused, is in conflict, or concentrates on solving a problem. Although neuroscientists carefully avoid the language of “this part of the brain does such and such…,” they have learned a great deal about the “personalities” of different brain regions, and the contribution of analyses of brain activity to psychological interpretation has greatly improved. The framing study yielded three main findings:

- A region that is commonly associated with emotional arousal (the amygdala) was most likely to be active when subjects’ choices conformed to the frame. This is just as we would expect if the emotionally loaded words KEEP and LOSE produce an immediate tendency to approach the sure thing (when it is framed as a gain) or avoid it (when it is framed as a loss). The amygdala is accessed very rapidly by emotional stimuli—and it is a likely suspect for involvement in System 1.
- A brain region known to be associated with conflict and self-control (the anterior cingulate) was more active when subjects did not do what comes naturally—when they chose the sure thing in spite of its being labeled LOSE. Resisting the inclination of System 1 apparently involves conflict.
- The most “rational” subjects—those who were the least susceptible to framing effects—showed enhanced activity in a frontal area of the brain that is implicated in combining emotion and reasoning to guide decisions. Remarkably, the “rational” individuals were not those who showed the strongest neural evidence of conflict. It appears that these elite participants were (often, not always) reality-bound with little conflict.

By joining observations of actual choices with a mapping of neural activity, this study provides a good illustration of how the emotion evoked by a word can “leak” into the final choice.

An experiment that Amos carried out with colleagues at Harvard Medical School is the classic example of emotional framing. Physician participants were given statistics about the outcomes of two treatments for lung cancer: surgery and radiation. The five-year survival rates clearly favor surgery, but in the short term surgery is riskier than radiation. Half the participants read statistics about survival rates, the others received the same information in terms of mortality rates. The two descriptions of the short-term outcomes of surgery were:

> The one-month survival rate is 90%.
> There is 10% mortality in the first month.

You already know the results: surgery was much more popular in the former frame (84% of physicians chose it) than in the latter (where 50% favored radiation). The logical equivalence of the two descriptions is transparent, and a reality-bound decision maker would make the same choice regardless of which version she saw. But System 1, as we have gotten to know it, is rarely indifferent to emotional words: mortality is bad, survival is good, and 90% survival sounds encouraging whereas 10% mortality is frightening. An important finding of the study is that physicians were just as susceptible to the framing effect as medically unsophisticated people (hospital patients and graduate students in a business school). Medical training is, evidently, no defense against the power of framing.

The KEEP–LOSE study and the survival–mortality experiment differed in one important respect. The participants in the brain-imaging study had many trials in which they encountered the different frames. They had an opportunity to recognize the distracting effects of the frames and to simplify their task by adopting a common frame, perhaps by translating the LOSE amount into its KEEP equivalent. It would take an intelligent person (and an alert System 2) to learn to do this, and the few participants who managed the feat were probably among the "rational" agents that the experimenters identified. In contrast, the physicians who read the statistics about the two therapies in the survival frame had no reason to suspect that they would have made a different choice if they had heard the same statistics framed in terms of mortality. Reframing is effortful and System 2 is normally lazy. Unless there is an obvious reason to do otherwise, most of us passively accept decision problems as they are framed and therefore rarely have an opportunity to discover the extent to which our preferences are *frame-bound* rather than *reality-bound*.

Empty Intuitions

Amos and I introduced our discussion of framing by an example that has become known as the “Asian disease problem”:

> Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> If program A is adopted, 200 people will be saved.
>
> If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.

A substantial majority of respondents choose program A: they prefer the certain option over the gamble.

The outcomes of the programs are framed differently in a second version:

> If program A' is adopted, 400 people will die.
>
> If program B' is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die.

Look closely and compare the two versions: the consequences of programs A and A' are identical; so are the consequences of programs B and B'. In the second frame, however, a large majority of people choose the gamble.

The different choices in the two frames fit prospect theory, in which choices between gambles and sure things are resolved differently, depending on whether the outcomes are good or bad. Decision makers tend to prefer the sure thing over the gamble (they are risk averse) when the outcomes are good. They tend to reject the sure thing and accept the gamble (they are risk seeking) when both outcomes are negative. These conclusions were well established for choices about gambles and sure things in the domain of money. The disease problem shows that the same rule applies when the outcomes are measured in lives saved or lost. In this context, as well, the framing experiment reveals that risk-averse and risk-seeking preferences are not reality-bound. Preferences between the same objective outcomes reverse with different formulations.

An experience that Amos shared with me adds a grim note to the story. Amos was invited to give a speech to a group of public-health professionals—the people who make decisions about vaccines and other programs. He took the opportunity to present them with the Asian disease problem: half saw the “lives-saved” version, the others answered the “lives-lost” question. Like other people, these professionals were susceptible to the framing effects. It is somewhat worrying that the officials who make decisions that affect everyone’s health can be swayed by such a superficial manipulation—but we must get used to the idea that even important decisions are influenced, if not governed, by System 1.

Even more troubling is what happens when people are confronted with their inconsistency: “You chose to save 200 lives for sure in one formulation and you chose to gamble rather than accept 400 deaths in the other. Now that you know these choices were inconsistent, how do you decide?” The answer is usually embarrassed silence. The intuitions that determined the original choice came from System 1 and had no more moral basis than did the preference for keeping £20 or the aversion to losing £30. Saving lives with certainty is good, deaths are bad. Most people find that their System 2 has no moral intuitions of its own to answer the question.

I am grateful to the great economist Thomas Schelling for my favorite example of a framing effect, which he described in his book *Choice and Consequence*. Schelling’s book was written before our work on framing was published, and framing was not his main concern. He reported on his experience teaching a class at the Kennedy School at Harvard, in which Bon he linthe topic was child exemptions in the tax code. Schelling told his students that a standard exemption is allowed for each child, and that the amount of the exemption is independent of the taxpayer’s income. He asked their opinion of the following proposition:

> Should the child exemption be larger for the rich than for the poor?

Your own intuitions are very likely the same as those of Schelling's students: they found the idea of favoring the rich by a larger exemption completely unacceptable.

Schelling then pointed out that the tax law is arbitrary. It assumes a childless family as the default case and reduces the tax by the amount of the exemption for each child. The tax law could of course be rewritten with another default case: a family with two children. In this formulation, families with fewer than the default number of children would pay a surcharge. Schelling now asked his students to report their view of another proposition:

> Should the childless poor pay as large a surcharge as the childless rich?

Here again you probably agree with the students' reaction to this idea, which they rejected with as much vehemence as the first. But Schelling showed his class that they could not logically reject both proposals. Set the two formulations next to each other. The difference between the tax due by a childless family and by a family with two children is described as a reduction of tax in the first version and as an increase in the second. If in the first version you want the poor to receive the same (or greater) benefit as the rich for having children, then you must want the poor to pay at least the same penalty as the rich for being childless.

We can recognize System 1 at work. It delivers an immediate response to any question about rich and poor: when in doubt, favor the poor. The surprising aspect of Schelling's problem is that this apparently simple moral rule does not work reliably. It generates contradictory answers to the same problem, depending on how that problem is framed. And of course you already know the question that comes next. Now that you have seen that your reactions to the problem are influenced by the frame, what is your answer to the question: How should the tax code treat the children of the rich and the poor?

Here again, you will probably find yourself dumbfounded. You have moral intuitions about differences between the rich and the poor, but these intuitions depend on an arbitrary reference point, and they are not about the real problem. This problem—the question about actual states of the world—is how much tax individual families should pay, how to fill the cells in the matrix of the tax code. You have no compelling moral intuitions to guide you in solving that problem. Your moral feelings are attached to frames, to descriptions of reality rather than to reality itself. The message about the nature of framing is stark: framing should not be viewed as an intervention that masks or distorts an underlying preference. At least in this instance—and also in the problems of the Asian disease and of surgery versus radiation for lung cancer—there is no underlying preference that is masked or distorted by the frame. Our preferences are about framed problems, and our moral intuitions are about descriptions, not about substance.

Good Frames

Not all frames are equal, and s Bon nd t="4%" wome frames are clearly better than alternative ways to describe (or to think about) the same thing. Consider the following pair of problems:

> A woman has bought two $80 tickets to the theater. When she arrives at the theater, she opens her wallet and discovers that the tickets are missing. Will she buy two more tickets to see the play?

> A woman goes to the theater, intending to buy two tickets that cost $80 each. She arrives at the theater, opens her wallet, and discovers to her dismay that the $160 with which she was going to make the purchase is missing. She could use her credit card. Will she buy the tickets?

Respondents who see only one version of this problem reach different conclusions, depending on the frame. Most believe that the woman in the first story will go home without seeing the show if she has lost tickets, and most believe that she will charge tickets for the show if she has lost money.

The explanation should already be familiar—this problem involves mental accounting and the sunk-cost fallacy. The different frames evoke different mental accounts, and the significance of the loss depends on the account to which it is posted. When tickets to a particular show are lost, it is natural to post them to the account associated with that play. The cost appears to have doubled and may now be more than the experience is worth. In contrast, a loss of cash is charged to a “general revenue” account—the theater patron is slightly poorer than she had thought she was, and the question she is likely to ask herself is whether the small reduction in her disposable wealth will change her decision about paying for tickets. Most respondents thought it would not.

The version in which cash was lost leads to more reasonable decisions. It is a better frame because the loss, even if tickets were lost, is “sunk,” and sunk costs should be ignored. History is irrelevant and the only issue that matters is the set of options the theater patron has now, and their likely consequences. Whatever she lost, the relevant fact is that she is less wealthy than she was before she opened her wallet. If the person who lost tickets were to ask for my advice, this is what I would say: “Would you have bought tickets if you had lost the equivalent amount of cash? If yes, go ahead and buy new ones.” Broader frames and inclusive accounts generally lead to more rational decisions.

In the next example, two alternative frames evoke different mathematical intuitions, and one is much superior to the other. In an article titled “The MPG Illusion,” which appeared in *Science* magazine in 2008, the psychologists Richard Larrick and Jack Soll identified a case in which passive acceptance of a misleading frame has substantial costs and serious policy consequences. Most car buyers list gas mileage as one of the factors that determine their choice; they know that high-mileage cars have lower operating costs. But the frame that has traditionally been used in the United States—miles per gallon—provides very poor guidance to the decisions of both individuals and policy makers. Consider two car owners who seek to reduce their costs:

> Adam switches from a gas-guzzler of 12 mpg to a slightly less voracious guzzler that runs at 14 mpg.

> The environmentally virtuous Beth switches from a Bon ss es from 30 mpg car to one that runs at 40 mpg.

Suppose both drivers travel equal distances over a year. Who will save more gas by switching? You almost certainly share the widespread intuition that Beth’s action is more significant than Adam’s: she reduced mpg by 10 miles rather than 2, and by a third (from 30 to 40) rather than a sixth (from 12 to 14). Now engage your System 2 and work it out. If the two car owners both drive 10,000 miles, Adam will reduce his consumption from a scandalous 833 gallons to a still shocking 714 gallons, for a saving of 119 gallons. Beth's use of fuel will drop from 333 gallons to 250, saving only 83 gallons. The mpg frame is wrong, and it should be replaced by the gallons-per-mile frame (or liters-per–100 kilometers, which is used in most other countries). As Larrick and Soll point out, the misleading intuitions fostered by the mpg frame are likely to mislead policy makers as well as car buyers.

Under President Obama, Cass Sunstein served as administrator of the Office of Information and Regulatory Affairs. With Richard Thaler, Sunstein coauthored *Nudge*, which is the basic manual for applying behavioral economics to policy. It was no accident that the "fuel economy and environment" sticker that will be displayed on every new car starting in 2013 will for the first time in the United States include the gallons-per-mile information. Unfortunately, the correct formulation will be in small print, along with the more familiar mpg information in large print, but the move is in the right direction. The five-year interval between the publication of "The MPG Illusion" and the implementation of a partial correction is probably a speed record for a significant application of psychological science to public policy.

A directive about organ donation in case of accidental death is noted on an individual's driver license in many countries. The formulation of that directive is another case in which one frame is clearly superior to the other. Few people would argue that the decision of whether or not to donate one's organs is unimportant, but there is strong evidence that most people make their choice thoughtlessly. The evidence comes from a comparison of the rate of organ donation in European countries, which reveals startling differences between neighboring and culturally similar countries. An article published in 2003 noted that the rate of organ donation was close to 100% in Austria but only 12% in Germany, 86% in Sweden but only 4% in Denmark.

These enormous differences are a framing effect, which is caused by the format of the critical question. The high-donation countries have an opt out form, where individuals who wish not to donate must check an appropriate box. Unless they take this simple action, they are considered willing donors. The low-contribution countries have an opt-in form: you must check a box to become a donor. That is all. The best single predictor of whether or not people will donate their organs is the designation of the default option that will be adopted without having to check a box.

Unlike other framing effects that have been traced to features of System 1, the organ donation effect is best explained by the laziness of System 2. People will check the box if they have already decided what they wish to do. If they are unprepared for the question, they have to make the effort of thinking whether they want to check the box. I imagine an organ donation form in which people are required to solve a mathematical problem in the box that corresponds to their decision. One of the boxes contains the problem 2 + 2 = ? The problem in the other box is 13 × 37 = ? The rate of donations would surely be swayed.

When the role of formulation is acknowledged, a policy question arises: Which formulation should be adopted? In this case, the answer is straightforward. If you believe that a large supply of donated organs is good for society, you will not be neutral between a formulation that yields almost 100% donations and another formulation that elicits donations from 4% of drivers.

As we have seen again and again, an important choice is controlled by an utterly inconsequential feature of the situation. This is embarrassing—it is not how we would wish to make important decisions. Furthermore, it is not how we experience the workings of our mind, but the evidence for these cognitive illusions is undeniable.

Count that as a point against the rational-agent theory. A theory that is worthy of the name asserts that certain events are impossible—they will not happen if the theory is true. When an "impossible" event is observed, the theory is falsified. Theories can survive for a long time after conclusive evidence falsifies them, and the rational-agent model certainly survived the evidence we have seen, and much other evidence as well.

The case of organ donation shows that the debate about human rationality can have a large effect in the real world. A significant difference between believers in the rational-agent model and the skeptics who question it is that the believers simply take it for granted that the formulation of a choice cannot determine preferences on significant problems. They will not even be interested in investigating the problem—and so we are often left with inferior outcomes.

Skeptics about rationality are not surprised. They are trained to be sensitive to the power of inconsequential factors as determinants of preference—my hope is that readers of this book have acquired this sensitivity.

Speaking of Frames and Reality

"They will feel better about what happened if they manage to frame the outcome in terms of how much money they kept rather than how much they lost."

"Let's reframe the problem by changing the reference point. Imagine we did not own it; how much would we think it is worth?"

"Charge the loss to your mental account of 'general revenue'—you will feel better!"

"They ask you to check the box to opt out of their mailing list. Their list would shrink if they asked you to check a box to opt in!"