14

TOM W'S SPECIALTY

Have a look at a simple puzzle:

> Tom W is a graduate student at the main university in your state. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W is now a student in each of these fields. Use 1 for the most likely, 9 for the least likely.
>
> business administration
> computer science
> engineering
> humanities and education
> law
> medicine
> library science
> physical and life sciences
> social science and social work

This question is easy, and you knew immediately that the relative size of enrollment in the different fields is the key to a solution. So far as you know, Tom W was picked at random from the graduate students at the university, like a single marble drawn from an urn. To decide whether a marble is more likely to be red or green, you need to know how many marbles of each color

there are in the urn. The proportion of marbles of a particular kind is called a *base rate*. Similarly, the base rate of humanities and education in this problem is the proportion of students of that field among all the graduate students. In the absence of specific information about Tom W, you will go by the base rates and guess that he is more likely to be enrolled in humanities and education than in computer science or library science, because there are more students overall in the humanities and education than in the other two fields. Using base-rate information is the obvious move when no other information is provided.

Next comes a task that has nothing to do with base rates.

> The following is a personality sketch of Tom W written during Tom's senior year in high school by a psychologist, on the basis of psychological tests of uncertain validity:
>
> > Tom W is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people, and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.
>
> Now please take a sheet of paper and rank the nine fields of specialization listed below by how similar the description of Tom W is to the typical graduate student in each of the following fields. Use 1 for the most likely and 9 for the least likely.

You will get more out of the chapter if you give the task a quick try; reading the report on Tom W is necessary to make your judgments about the various graduate specialties.

This question too is straightforward. It requires you to retrieve, or perhaps to construct, a stereotype of graduate students in the different fields. When the experiment was first conducted, in the early 1970s, the average ordering was as follows. Yours is probably not very different:

1. computer science
2. engineering
3. business administration
4. physical and life sciences
5. library science
6. law
7. medicine
8. humanities and education
9. social science and social work

You probably ranked computer science among the best fitting because of hints of nerdiness ("corny puns"). In fact, the description of Tom W was written to fit that stereotype. Another specialty that most people ranked high is engineering ("neat and tidy systems"). You probably thought that Tom W is not a good fit with your idea of social science and social work ("little feel and little sympathy for other people"). Professional stereotypes appear to have changed little in the nearly forty years since I designed the description of Tom W.

The task of ranking the nine careers is complex and certainly requires the discipline and sequential organization of which only System 2 is capable. However, the hints planted in the description (corny puns and others) were intended to activate an association with a stereotype, an automatic activity of System 1.

The instructions for this similarity task required a comparison of the description of Tom W to the stereotypes of the various fields of specialization. For the purposes of that task, the accuracy of the description—whether or not it is a true portrait of Tom W—is irrelevant. So is your knowledge of the base rates of the various fields. The similarity of an individual to the stereotype of a group is unaffected by the size of the group. Indeed, you could compare the description of Tom to an image of graduate students in library science even if there is no such department at the university.

If you examine Tom W again, you will see that he is a good fit to stereotypes of some small groups of students (computer scientists, librarians, engineers) and a much poorer fit to the largest groups (humanities and education, social science and social work). Indeed, the participants almost always ranked the two largest fields very low. Tom W was intentionally designed as an "anti-base-rate" character, a good fit to small fields and a poor fit to the most populated specialties.

PREDICTING BY REPRESENTATIVENESS

The third task in the sequence was administered to graduate students in psychology, and it is the critical one: rank the fields of specialization in order of the likelihood that Tom W is now a graduate student in each of these fields. The members of this prediction group knew the relevant statistical facts: they were familiar with the base rates of the different fields, and they knew that the source of Tom W's description was not highly trustworthy. However, we expected them to focus exclusively on the similarity of the description to the stereotypes—we called it representativeness—ignoring both the base rates and the doubts about the veracity of the description. They would then rank the small specialty—computer science—as highly probable, because that outcome gets the highest representativeness score.

Amos and I worked hard during the year we spent in Eugene, and I sometimes stayed in the office through the night. One of my tasks for such a night was to make up a description that would pit representativeness and base rates against each other. Tom W was the result of my efforts, and I completed the description in the early morning hours. The first person who showed up to work that morning was our colleague and friend Robyn Dawes, who was both a sophisticated statistician and a skeptic about the validity of intuitive judgment. If anyone would see the relevance of the base rate, it would have to be Robyn. I called Robyn over, gave him the question I had just typed, and asked him to guess Tom W's profession. I still remember his sly smile as he said tentatively, "computer scientist?" That was a happy moment—even the mighty had fallen. Of course, Robyn immediately recognized his mistake as soon as I mentioned "base rate," but he had not spontaneously thought of it. Although he knew as much as anyone about the role of base rates in prediction, he neglected them when presented with the description of an individual's personality. As expected, he substituted a judgment of representativeness for the probability he was asked to assess.

Amos and I then collected answers to the same question from 114 graduate students in psychology at three major universities, all of whom had taken several courses in statistics. They did not disappoint us. Their rankings of the nine fields by probability did not differ from ratings by similarity to the stereotype. Substitution was perfect in this case: there was no indication that the participants did anything else but judge representativeness. The question about probability (likelihood) was difficult, but the question about similarity was easier, and it was answered instead. This is a serious mistake, because judgments of similarity and probability are not constrained by the same logical rules. It is entirely acceptable for judgments of similarity to be unaffected by base rates and also by the possibility that the description was inaccurate, but anyone who ignores base rates and the quality of evidence in probability assessments will certainly make mistakes.

The concept "the probability that Tom W studies computer science" is not a simple one. Logicians and statisticians disagree about its meaning, and some would say it has no meaning at all. For many experts it is a measure of subjective degree of belief. There are some events you are sure of, for example, that the sun rose this morning, and others you consider impossible, such as the Pacific Ocean freezing all at once. Then there are many events, such as your next-door neighbor being a computer scientist, to which you assign an intermediate degree of belief—which is your probability of that event.

Logicians and statisticians have developed competing definitions of probability, all very precise. For laypeople, however, probability (a synonym of *likelihood* in everyday language) is a vague notion, related to uncertainty, propensity, plausibility, and surprise. The vagueness is not particular to this concept, nor is it especially troublesome. We know, more or less, what we mean when we use a word such as *democracy* or *beauty* and the people we are talking to understand, more or less, what we intended to say. In all the years I spent asking questions about the probability of events, no one ever raised a hand to ask me, "Sir, what do you mean by probability?" as they would have done if I had asked them to assess a strange concept such as globability. Everyone acted as if they knew how to answer my questions, although we all understood that it would be unfair to ask them for an explanation of what the word means.

People who are asked to assess probability are not stumped, because they do not try to judge probability as statisticians and philosophers use the word. A question about probability or likelihood activates a mental shotgun, evoking answers to easier questions. One of the easy answers is an automatic assessment of representativeness—routine in understanding language. The (false) statement that "Elvis Presley's parents wanted him to be a dentist" is mildly funny because the discrepancy between the images of Presley and a dentist is detected automatically. System 1 generates an impression of similarity without intending to do so. The representativeness heuristic is involved when someone says "She will win the election; you can see she is a winner" or "He won't go far as an academic; too many tattoos." We rely on representativeness when we judge the potential leadership of a candidate for office by the shape of his chin or the forcefulness of his speeches.

Although it is common, prediction by representativeness is not statisti-

cally optimal. Michael Lewis's bestselling *Moneyball* is a story about the inefficiency of this mode of prediction. Professional baseball scouts traditionally forecast the success of possible players in part by their build and look. The hero of Lewis's book is Billy Beane, the manager of the Oakland A's, who made the unpopular decision to overrule his scouts and to select players by the statistics of past performance. The players the A's picked were inexpensive, because other teams had rejected them for not looking the part. The team soon achieved excellent results at low cost.

THE SINS OF REPRESENTATIVENESS

Judging probability by representativeness has important virtues: the intuitive impressions that it produces are often—indeed, usually—more accurate than chance guesses would be.

- On most occasions, people who act friendly are in fact friendly.
- A professional athlete who is very tall and thin is much more likely to play basketball than football.
- People with a PhD are more likely to subscribe to *The New York Times* than people who ended their education after high school.
- Young men are more likely than elderly women to drive aggressively.

In all these cases and in many others, there is some truth to the stereotypes that govern judgments of representativeness, and predictions that follow this heuristic may be accurate. In other situations, the stereotypes are false and the representativeness heuristic will mislead, especially if it causes people to neglect base-rate information that points in another direction. Even when the heuristic has some validity, exclusive reliance on it is associated with grave sins against statistical logic.

One sin of representativeness is an excessive willingness to predict the occurrence of unlikely (low base-rate) events. Here is an example: you see a person reading *The New York Times* on the New York subway. Which of the following is a better bet about the reading stranger?

She has a PhD.
She does not have a college degree.

Representativeness would tell you to bet on the PhD, but this is not necessarily wise. You should seriously consider the second alternative, because many more nongraduates than PhDs ride in New York subways. And if you must guess whether a woman who is described as "a shy poetry lover" studies Chinese literature or business administration, you should opt for the latter option. Even if every female student of Chinese literature is shy and loves poetry, it is almost certain that there are more bashful poetry lovers in the much larger population of business students.

People without training in statistics are quite capable of using base rates in predictions under some conditions. In the first version of the Tom W problem, which provides no details about him, it is obvious to everyone that the probability of Tom W's being in a particular field is simply the base-rate frequency of enrollment in that field. However, concern for base rates evidently disappears as soon as Tom W's personality is described.

Amos and I originally believed, on the basis of our early evidence, that base-rate information will *always* be neglected when information about the specific instance is available, but that conclusion was too strong. Psychologists have conducted many experiments in which base-rate information is explicitly provided as part of the problem, and many of the participants are influenced by those base rates, although the information about the individual case is almost always weighted more than mere statistics. Norbert Schwarz and his colleagues showed that instructing people to "think like a statistician" enhanced the use of base-rate information, while the instruction to "think like a clinician" had the opposite effect.

An experiment that was conducted a few years ago with Harvard undergraduates yielded a finding that surprised me: enhanced activation of System 2 caused a significant improvement of predictive accuracy in the Tom W problem. The experiment combined the old problem with a modern variation of cognitive fluency. Half the students were told to puff out their cheeks during the task, while the others were told to frown. Frowning, as we have seen, generally increases the vigilance of System 2 and reduces both overconfidence and the reliance on intuition. The students who puffed out their cheeks (an emotionally neutral expression) replicated the original results: they relied exclusively on representativeness and ignored the base rates. As the authors had predicted, however, the frowners did show some sensitivity to the base rates. This is an instructive finding.

When an incorrect intuitive judgment is made, System 1 and System 2 should both be indicted. System 1 suggested the incorrect intuition, and System 2 endorsed it and expressed it in a judgment. However, there are two possible reasons for the failure of System 2—ignorance or laziness. Some people ignore base rates because they believe them to be irrelevant in the presence of individual information. Others make the same mistake because they are not focused on the task. If frowning makes a difference, laziness seems to be the proper explanation of base-rate neglect, at least among Harvard undergrads. Their System 2 "knows" that base rates are relevant even when they are not explicitly mentioned, but applies that knowledge only when it invests special effort in the task.

The second sin of representativeness is insensitivity to the quality of evidence. Recall the rule of System 1: WYSIATI. In the Tom W example, what activates your associative machinery is a description of Tom, which may or may not be an accurate portrayal. The statement that Tom W "has little feel and little sympathy for people" was probably enough to convince you (and most other readers) that he is very unlikely to be a student of social science or social work. But you were explicitly told that the description should not be trusted!

You surely understand in principle that worthless information should not be treated differently from a complete lack of information, but WYSIATI makes it very difficult to apply that principle. Unless you decide immediately to reject evidence (for example, by determining that you received it from a liar), your System 1 will automatically process the information available as if it were true. There is one thing you can do when you have doubts about the quality of the evidence: let your judgments of probability stay close to the base rate. Don't expect this exercise of discipline to be easy—it requires a significant effort of self-monitoring and self-control.

The correct answer to the Tom W puzzle is that you should stay very close to your prior beliefs, slightly reducing the initially high probabilities of well-populated fields (humanities and education; social science and social work) and slightly raising the low probabilities of rare specialties (library science, computer science). You are not exactly where you would be if you had known nothing at all about Tom W, but the little evidence you have is not trustworthy, so the base rates should dominate your estimates.

HOW TO DISCIPLINE INTUITION

Your probability that it will rain tomorrow is your subjective degree of belief, but you should not let yourself believe whatever comes to your mind.

To be useful, your beliefs should be constrained by the logic of probability. So if you believe that there is a 40% chance that it will rain sometime tomorrow, you must also believe that there is a 60% chance it will not rain tomorrow, and you must not believe that there is a 50% chance that it will rain tomorrow morning. And if you believe that there is a 30% chance that candidate X will be elected president, and an 80% chance that he will be reelected if he wins the first time, then you must believe that the chances that he will be elected twice in a row are 24%.

The relevant "rules" for cases such as the Tom W problem are provided by Bayesian statistics. This influential modern approach to statistics is named after an English minister of the eighteenth century, the Reverend Thomas Bayes, who is credited with the first major contribution to a large problem: the logic of how people should change their mind in the light of evidence. Bayes's rule specifies how prior beliefs (in the examples of this chapter, base rates) should be combined with the diagnosticity of the evidence, the degree to which it favors the hypothesis over the alternative. For example, if you believe that 3% of graduate students are enrolled in computer science (the base rate), and you also believe that the description of Tom W is 4 times more likely for a graduate student in that field than in other fields, then Bayes's rule says you must believe that the probability that Tom W is a computer scientist is now 11%. If the base rate had been 80%, the new degree of belief would be 94.1%. And so on.

The mathematical details are not relevant in this book. There are two ideas to keep in mind about Bayesian reasoning and how we tend to mess it up. The first is that base rates matter, even in the presence of evidence about the case at hand. This is often not intuitively obvious. The second is that intuitive impressions of the diagnosticity of evidence are often exaggerated. The combination of WYSIATI and associative coherence tends to make us believe in the stories we spin for ourselves. The essential keys to disciplined Bayesian reasoning can be simply summarized:

- Anchor your judgment of the probability of an outcome on a plausible base rate.
- Question the diagnosticity of your evidence.

Both ideas are straightforward. It came as a shock to me when I realized that I was never taught how to implement them, and that even now I find it unnatural to do so.

SPEAKING OF REPRESENTATIVENESS

"The lawn is well trimmed, the receptionist looks competent, and the furniture is attractive, but this doesn't mean it is a well-managed company. I hope the board does not go by representativeness."

"This start-up looks as if it could not fail, but the base rate of success in the industry is extremely low. How do we know this case is different?"

"They keep making the same mistake: predicting rare events from weak evidence. When the evidence is weak, one should stick with the base rates."

"I know this report is absolutely damning, and it may be based on solid evidence, but how sure are we? We must allow for that uncertainty in our thinking."

15

LINDA: LESS IS MORE

The best-known and most controversial of our experiments involved a fictitious lady called Linda. Amos and I made up the Linda problem to provide conclusive evidence of the role of heuristics in judgment and of their incompatibility with logic. This is how we described Linda:

> Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

The audiences who heard this description in the 1980s always laughed because they immediately knew that Linda had attended the University of California at Berkeley, which was famous at the time for its radical, politically engaged students. In one of our experiments we presented participants with a list of eight possible scenarios for Linda. As in the Tom W problem, some ranked the scenarios by representativeness, others by probability. The Linda problem is similar, but with a twist.

> Linda is a teacher in elementary school.
> Linda works in a bookstore and takes yoga classes.
> Linda is active in the feminist movement.
> Linda is a psychiatric social worker.
> Linda is a member of the League of Women Voters.

> *Linda is a bank teller.*
> *Linda is an insurance salesperson.*
> *Linda is a bank teller and is active in the feminist movement.*

The problem shows its age in several ways. The League of Women Voters is no longer as prominent as it was, and the idea of a feminist "movement" sounds quaint, a testimonial to the change in the status of women over the last thirty years. Even in the Facebook era, however, it is still easy to guess the almost perfect consensus of judgments: Linda is a very good fit for an active feminist, a fairly good fit for someone who works in a bookstore and takes yoga classes—and a very poor fit for a bank teller or an insurance salesperson.

Now focus on the critical items in the list: Does Linda look more like a bank teller, or more like a bank teller who is active in the feminist movement? Everyone agrees that Linda fits the idea of a "feminist bank teller" better than she fits the stereotype of bank tellers. The stereotypical bank teller is not a feminist activist, and adding that detail to the description makes for a more coherent story.

The twist comes in the judgments of likelihood, because there is a logical relation between the two scenarios. Think in terms of Venn diagrams. The set of feminist bank tellers is wholly included in the set of bank tellers, as every feminist bank teller is a bank teller. Therefore the probability that Linda is a feminist bank teller *must* be lower than the probability of her being a bank teller. When you specify a possible event in greater detail you can only lower its probability. The problem therefore sets up a conflict between the intuition of representativeness and the logic of probability.

Our initial experiment was between-subjects. Each participant saw a set of seven outcomes that included only one of the critical items ("bank teller" or "feminist bank teller"). Some ranked the outcomes by resemblance, others by likelihood. As in the case of Tom W, the average rankings by resemblance and by likelihood were identical; "feminist bank teller" ranked higher than "bank teller" in both.

Then we took the experiment further, using a within-subject design. We made up the questionnaire as you saw it, with "bank teller" in the sixth position in the list and "feminist bank teller" as the last item. We were convinced that subjects would notice the relation between the two outcomes, and that their rankings would be consistent with logic. Indeed, we were so certain of this that we did not think it worthwhile to conduct a special experiment. My assistant was running another experiment in the lab, and she asked the subjects to complete the new Linda questionnaire while signing out, just before they got paid.

About ten questionnaires had accumulated in a tray on my assistant's desk before I casually glanced at them and found that all the subjects had ranked "feminist bank teller" as more probable than "bank teller." I was so surprised that I still retain a "flashbulb memory" of the gray color of the metal desk and of where everyone was when I made that discovery. I quickly called Amos in great excitement to tell him what we had found: we had pitted logic against representativeness, and representativeness had won!

In the language of this book, we had observed a failure of System 2: our participants had a fair opportunity to detect the relevance of the logical rule, since both outcomes were included in the same ranking. They did not take advantage of that opportunity. When we extended the experiment, we found that 89% of the undergraduates in our sample violated the logic of probability. We were convinced that statistically sophisticated respondents would do better, so we administered the same questionnaire to doctoral students in the decision-science program of the Stanford Graduate School of Business, all of whom had taken several advanced courses in probability, statistics, and decision theory. We were surprised again: 85% of these respondents also ranked "feminist bank teller" as more likely than "bank teller."

In what we later described as "increasingly desperate" attempts to eliminate the error, we introduced large groups of people to Linda and asked them this simple question:

Which alternative is more probable?
Linda is a bank teller.
Linda is a bank teller and is active in the feminist movement.

This stark version of the problem made Linda famous in some circles, and it earned us years of controversy. About 85% to 90% of undergraduates at several major universities chose the second option, contrary to logic. Remarkably, the sinners seemed to have no shame. When I asked my large undergraduate class in some indignation, "Do you realize that you have violated an elementary logical rule?" someone in the back row shouted, "So what?" and a graduate student who made the same error explained herself by saying, "I thought you just asked for my opinion."

The word *fallacy* is used, in general, when people fail to apply a logical rule that is obviously relevant. Amos and I introduced the idea of a *conjunction fallacy*, which people commit when they judge a conjunction of two events (here, bank teller and feminist) to be more probable than one of the events (bank teller) in a direct comparison.

As in the Muller-Lyer illusion, the fallacy remains attractive even when you recognize it for what it is. The naturalist Stephen Jay Gould described his own struggle with the Linda problem. He knew the correct answer, of course, and yet, he wrote, "a little homunculus in my head continues to jump up and down, shouting at me—'but she can't just be a bank teller; read the description.'" The little homunculus is of course Gould's System 1 speaking to him in insistent tones. (The two-system terminology had not yet been introduced when he wrote.)

The correct answer to the short version of the Linda problem was the majority response in only one of our studies: 64% of a group of graduate students in the social sciences at Stanford and at Berkeley correctly judged "feminist bank teller" to be less probable than "bank teller." In the original version with eight outcomes (shown above), only 15% of a similar group of graduate students had made that choice. The difference is instructive. The longer version separated the two critical outcomes by an intervening item (insurance salesperson), and the readers judged each outcome independently, without comparing them. The shorter version, in contrast, required an explicit comparison that mobilized System 2 and allowed most of the statistically sophisticated students to avoid the fallacy. Unfortunately, we did not explore the reasoning of the substantial minority (36%) of this knowledgeable group who chose incorrectly.

The judgments of probability that our respondents offered, in both the Tom W and Linda problems, corresponded precisely to judgments of representativeness (similarity to stereotypes). Representativeness belongs to a cluster of closely related basic assessments that are likely to be generated together. The most representative outcomes combine with the personality description to produce the most coherent stories. The most coherent stories are not necessarily the most probable, but they are *plausible*, and the notions of coherence, plausibility, and probability are easily confused by the unwary.

The uncritical substitution of plausibility for probability has pernicious effects on judgments when scenarios are used as tools of forecasting. Consider these two scenarios, which were presented to different groups, with a request to evaluate their probability:

A massive flood somewhere in North America next year, in which more than 1,000 people drown

An earthquake in California sometime next year, causing a flood in which more than 1,000 people drown

The California earthquake scenario is more plausible than the North American scenario, although its probability is certainly smaller. As expected, probability judgments were higher for the richer and more detailed scenario, contrary to logic. This is a trap for forecasters and their clients: adding detail to scenarios makes them more persuasive, but less likely to come true.

To appreciate the role of plausibility, consider the following questions:

Which alternative is more probable?
Mark has hair.
Mark has blond hair.

and

Which alternative is more probable?
Jane is a teacher.
Jane is a teacher and walks to work.

The two questions have the same logical structure as the Linda problem, but they cause no fallacy, because the more detailed outcome is only more detailed—it is not more plausible, or more coherent, or a better story. The evaluation of plausibility and coherence does not suggest an answer to the probability question. In the absence of a competing intuition, logic prevails.

LESS IS MORE, SOMETIMES EVEN IN JOINT EVALUATION

Christopher Hsee, of the University of Chicago, asked people to price sets of dinnerware offered in a clearance sale in a local store, where dinnerware regularly runs between $30 and $60. There were three groups in his experiment. The display below was shown to one group; Hsee labels that *joint evaluation*, because it allows a comparison of the two sets. The other two groups were shown only one of the two sets; this is *single evaluation*. Joint evaluation is a within-subject experiment, and single evaluation is between-subjects.

	Set A: 40 pieces	Set B: 24 pieces
Dinner plates	8, all in good condition	8, all in good condition
Soup/salad bowls	8, all in good condition	8, all in good condition
Dessert plates	8, all in good condition	8, all in good condition
Cups	8, 2 of them broken	
Saucers	8, 7 of them broken	

Assuming that the dishes in the two sets are of equal quality, which is worth more? This question is easy. You can see that Set A contains all the dishes of Set B, and seven additional intact dishes, and it *must* be valued more. Indeed, the participants in Hsee's joint evaluation experiment were willing to pay a little more for Set A than for Set B: $32 versus $30.

The results reversed in single evaluation, where Set B was priced much higher than Set A: $33 versus $23. We know why this happened. Sets (including dinnerware sets!) are represented by norms and prototypes. You can sense immediately that the average value of the dishes is much lower for Set A than for Set B, because no one wants to pay for broken dishes. If the average dominates the evaluation, it is not surprising that Set B is valued more. Hsee called the resulting pattern *less is more*. By removing 16 items from Set A (7 of them intact), its value is improved.

Hsee's finding was replicated by the experimental economist John List in a real market for baseball cards. He auctioned sets of ten high-value cards, and identical sets to which three cards of modest value were added. As in the dinnerware experiment, the larger sets were valued more than the smaller ones in joint evaluation, but less in single evaluation. From the perspective of economic theory, this result is troubling: the economic value of a dinnerware set or of a collection of baseball cards is a sum-like variable. Adding a positively valued item to the set can only increase its value.

The Linda problem and the dinnerware problem have exactly the same structure. Probability, like economic value, is a sum-like variable, as illustrated by this example:

probability (Linda is a teller) = probability (Linda is feminist teller) +
probability (Linda is non-feminist teller)

This is also why, as in Hsee's dinnerware study, single evaluations of the Linda problem produce a less-is-more pattern. System 1 averages instead of adding, so when the non-feminist bank tellers are removed from the set, subjective probability increases. However, the sum-like nature of the variable is less obvious for probability than for money. As a result, joint evaluation eliminates the error only in Hsee's experiment, not in the Linda experiment.

Linda was not the only conjunction error that survived joint evaluation.

We found similar violations of logic in many other judgments. Participants in one of these studies were asked to rank four possible outcomes of the next Wimbledon tournament from most to least probable. Björn Borg was the dominant tennis player of the day when the study was conducted. These were the outcomes:

A. Borg will win the match.
B. Borg will lose the first set.
C. Borg will lose the first set but win the match.
D. Borg will win the first set but lose the match.

The critical items are B and C. B is the more inclusive event and its probability *must* be higher than that of an event it includes. Contrary to logic, but not to representativeness or plausibility, 72% assigned B a lower probability than C—another instance of less is more in a direct comparison. Here again, the scenario that was judged more probable was unquestionably more plausible, a more coherent fit with all that was known about the best tennis player in the world.

To head off the possible objection that the conjunction fallacy is due to a misinterpretation of probability, we constructed a problem that required probability judgments, but in which the events were not described in words, and the term *probability* did not appear at all. We told participants about a regular six-sided die with four green faces and two red faces, which would be rolled 20 times. They were shown three sequences of greens (G) and reds (R), and were asked to choose one. They would (hypothetically) win $25 if their chosen sequence showed up. The sequences were:

1. RGRRR
2. GRGRRR
3. GRRRRR

Because the die has twice as many green as red faces, the first sequence is quite unrepresentative—like Linda being a bank teller. The second sequence, which contains six tosses, is a better fit to what we would expect from this die, because it includes two G's. However, this sequence was constructed by adding a G to the beginning of the first sequence, so it can only be less likely than the first. This is the nonverbal equivalent to Linda being a feminist bank teller. As in the Linda study, representativeness dominated. Almost two-thirds of respondents preferred to bet on sequence 2 rather than on

sequence 1. When presented with arguments for the two choices, however, a large majority found the correct argument (favoring sequence 1) more convincing.

The next problem was a breakthrough, because we finally found a condition in which the incidence of the conjunction fallacy was much reduced. Two groups of subjects saw slightly different variants of the same problem:

A health survey was conducted in a sample of adult males in British Columbia, of all ages and occupations. Please give your best estimate of the following values:	A health survey was conducted in a sample of 100 adult males in British Columbia, of all ages and occupations. Please give your best estimate of the following values:
What percentage of the men surveyed have had one or more heart attacks?	How many of the 100 participants have had one or more heart attacks?
What percentage of the men surveyed are both over 55 years old and have had one or more heart attacks?	How many of the 100 participants both are over 55 years old and have had one or more heart attacks?

The incidence of errors was 65% in the group that saw the problem on the left, and only 25% in the group that saw the problem on the right.

Why is the question "How many of the 100 participants . . ." so much easier than "What percentage . . ."? A likely explanation is that the reference to 100 individuals brings a spatial representation to mind. Imagine that a large number of people are instructed to sort themselves into groups in a room: "Those whose names begin with the letters *A* to *L* are told to gather in the front left corner." They are then instructed to sort themselves further. The relation of inclusion is now obvious, and you can see that individuals whose name begins with C will be a subset of the crowd in the front left corner. In the medical survey question, heart attack victims end up in a corner of the room, and some of them are less than 55 years old. Not everyone will share this particular vivid imagery, but many subsequent experiments have shown that the frequency representation, as it is known, makes it easy to appreciate that one group is wholly included in the other. The solution to the puzzle appears to be that a question phrased as "how many?" makes you think of individuals, but the same question phrased as "what percentage?" does not.

What have we learned from these studies about the workings of System 2? One conclusion, which is not new, is that System 2 is not impressively alert. The undergraduates and graduate students who participated in our studies of the conjunction fallacy certainly "knew" the logic of Venn diagrams, but they did not apply it reliably even when all the relevant information was laid out in front of them. The absurdity of the less-is-more pattern was obvious in Hsee's dinnerware study and was easily recognized in the "how many?" representation, but it was not apparent to the thousands of people who have committed the conjunction fallacy in the original Linda problem and in others like it. In all these cases, the conjunction appeared plausible, and that sufficed for an endorsement of System 2.

The laziness of System 2 is part of the story. If their next vacation had depended on it, and if they had been given indefinite time and told to follow logic and not to answer until they were sure of their answer, I believe that most of our subjects would have avoided the conjunction fallacy. However, their vacation did not depend on a correct answer; they spent very little time on it, and were content to answer as if they had only been "asked for their opinion." The laziness of System 2 is an important fact of life, and the observation that representativeness can block the application of an obvious logical rule is also of some interest.

The remarkable aspect of the Linda story is the contrast to the broken-dishes study. The two problems have the same structure, but yield different results. People who see the dinnerware set that includes broken dishes put a very low price on it; their behavior reflects a rule of intuition. Others who see both sets at once apply the logical rule that more dishes can only add value. Intuition governs judgments in the between-subjects condition; logic rules in joint evaluation. In the Linda problem, in contrast, intuition often overcame logic even in joint evaluation, although we identified some conditions in which logic prevails.

Amos and I believed that the blatant violations of the logic of probability that we had observed in transparent problems were interesting and worth reporting to our colleagues. We also believed that the results strengthened our argument about the power of judgment heuristics, and that they would persuade doubters. And in this we were quite wrong. Instead, the Linda problem became a case study in the norms of controversy.

The Linda problem attracted a great deal of attention, but it also became a magnet for critics of our approach to judgment. As we had already done, researchers found combinations of instructions and hints that reduced the incidence of the fallacy; some argued that, in the context of the Linda problem, it is reasonable for subjects to understand the word "probability"

as if it means "plausibility." These arguments were sometimes extended to suggest that our entire enterprise was misguided: if one salient cognitive illusion could be weakened or explained away, others could be as well. This reasoning neglects the unique feature of the conjunction fallacy as a case of conflict between intuition and logic. The evidence that we had built up for heuristics from between-subjects experiment (including studies of Linda) was not challenged—it was simply not addressed, and its salience was diminished by the exclusive focus on the conjunction fallacy. The net effect of the Linda problem was an increase in the visibility of our work to the general public, and a small dent in the credibility of our approach among scholars in the field. This was not at all what we had expected.

If you visit a courtroom you will observe that lawyers apply two styles of criticism: to demolish a case they raise doubts about the strongest arguments that favor it; to discredit a witness, they focus on the weakest part of the testimony. The focus on weaknesses is also normal in political debates. I do not believe it is appropriate in scientific controversies, but I have come to accept as a fact of life that the norms of debate in the social sciences do not prohibit the political style of argument, especially when large issues are at stake—and the prevalence of bias in human judgment is a large issue.

Some years ago I had a friendly conversation with Ralph Hertwig, a persistent critic of the Linda problem, with whom I had collaborated in a vain attempt to settle our differences. I asked him why he and others had chosen to focus exclusively on the conjunction fallacy, rather than on other findings that provided stronger support for our position. He smiled as he answered, "It was more interesting," adding that the Linda problem had attracted so much attention that we had no reason to complain.

SPEAKING OF LESS IS MORE

"They constructed a very complicated scenario and insisted on calling it highly probable. It is not—it is only a plausible story."

"They added a cheap gift to the expensive product, and made the whole deal less attractive. Less is more in this case."

"In most situations, a direct comparison makes people more careful and more logical. But not always. Sometimes intuition beats logic even when the correct answer stares you in the face."

16

CAUSES TRUMP STATISTICS

Consider the following scenario and note your intuitive answer to the question.

> A cab was involved in a hit-and-run accident at night.
> Two cab companies, the Green and the Blue, operate in the city.
> You are given the following data:
>
> - 85% of the cabs in the city are Green and 15% are Blue.
> - A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green?

This is a standard problem of Bayesian inference. There are two items of information: a base rate and the imperfectly reliable testimony of a witness. In the absence of a witness, the probability of the guilty cab being Blue is 15%, which is the base rate of that outcome. If the two cab companies had been equally large, the base rate would be uninformative and you would consider only the reliability of the witness, concluding that the probability is 80%. The two sources of information can be combined by Bayes's rule.

The correct answer is 41%. However, you can probably guess what people do when faced with this problem: they ignore the base rate and go with the witness. The most common answer is 80%.

CAUSAL STEREOTYPES

Now consider a variation of the same story, in which only the presentation of the base rate has been altered.

> You are given the following data:
>
> - The two companies operate the same number of cabs, but Green cabs are involved in 85% of accidents.
> - The information about the witness is as in the previous version.

The two versions of the problem are mathematically indistinguishable, but they are psychologically quite different. People who read the first version do not know how to use the base rate and often ignore it. In contrast, people who see the second version give considerable weight to the base rate, and their average judgment is not too far from the Bayesian solution. Why?

In the first version, the base rate of Blue cabs is a statistical fact about the cabs in the city. A mind that is hungry for causal stories finds nothing to chew on: How does the number of Green and Blue cabs in the city cause this cab driver to hit and run?

In the second version, in contrast, the drivers of Green cabs cause more than 5 times as many accidents as the Blue cabs do. The conclusion is immediate: the Green drivers must be a collection of reckless madmen! You have now formed a stereotype of Green recklessness, which you apply to unknown individual drivers in the company. The stereotype is easily fitted into a causal story, because recklessness is a causally relevant fact about individual cabdrivers. In this version, there are two causal stories that need to be combined or reconciled. The first is the hit and run, which naturally evokes the idea that a reckless Green driver was responsible. The second is the witness's testimony, which strongly suggests the cab was Blue. The inferences from the two stories about the color of the car are contradictory and approximately cancel each other. The chances for the two colors are about equal (the Bayesian estimate is 41%, reflecting the fact that the base rate of Green cabs is a little more extreme than the reliability of the witness who reported a Blue cab).

The cab example illustrates two types of base rates. *Statistical base rates* are facts about a population to which a case belongs, but they are not relevant to the individual case. *Causal base rates* change your view of how the individual case came to be. The two types of base-rate information are treated differently:

- Statistical base rates are generally underweighted, and sometimes neglected altogether, when specific information about the case at hand is available.
- Causal base rates are treated as information about the individual case and are easily combined with other case-specific information.

The causal version of the cab problem had the form of a stereotype: Green drivers are dangerous. Stereotypes are statements about the group that are (at least tentatively) accepted as facts about every member. Here are two examples:

> Most of the graduates of this inner-city school go to college.
> Interest in cycling is widespread in France.

These statements are readily interpreted as setting up a propensity in individual members of the group, and they fit in a causal story. Many graduates of this particular inner-city school are eager and able to go to college, presumably because of some beneficial features of life in that school. There are forces in French culture and social life that cause many Frenchmen to take an interest in cycling. You will be reminded of these facts when you think about the likelihood that a particular graduate of the school will attend college, or when you wonder whether to bring up the Tour de France in a conversation with a Frenchman you just met.

Stereotyping is a bad word in our culture, but in my usage it is neutral. One of the basic characteristics of System 1 is that it represents categories as norms and prototypical exemplars. This is how we think of horses, refrigerators, and New York police officers; we hold in memory a representation of one or more "normal" members of each of these categories. When the categories are social, these representations are called stereotypes. Some stereotypes are perniciously wrong, and hostile stereotyping can have dreadful

consequences, but the psychological facts cannot be avoided: stereotypes, both correct and false, are how we think of categories.

You may note the irony. In the context of the cab problem, the neglect of base-rate information is a cognitive flaw, a failure of Bayesian reasoning, and the reliance on causal base rates is desirable. Stereotyping the Green drivers improves the accuracy of judgment. In other contexts, however, such as hiring or profiling, there is a strong social norm against stereotyping, which is also embedded in the law. This is as it should be. In sensitive social contexts, we do not want to draw possibly erroneous conclusions about the individual from the statistics of the group. We consider it morally desirable for base rates to be treated as statistical facts about the group rather than as presumptive facts about individuals. In other words, we reject causal base rates.

The social norm against stereotyping, including the opposition to profiling, has been highly beneficial in creating a more civilized and more equal society. It is useful to remember, however, that neglecting valid stereotypes inevitably results in suboptimal judgments. Resistance to stereotyping is a laudable moral position, but the simplistic idea that the resistance is costless is wrong. The costs are worth paying to achieve a better society, but denying that the costs exist, while satisfying to the soul and politically correct, is not scientifically defensible. Reliance on the affect heuristic is common in politically charged arguments. The positions we favor have no cost and those we oppose have no benefits. We should be able to do better.

CAUSAL SITUATIONS

Amos and I constructed the variants of the cab problem, but we did not invent the powerful notion of causal base rates; we borrowed it from the psychologist Icek Ajzen. In his experiment, Ajzen showed his participants brief vignettes describing some students who had taken an exam at Yale and asked the participants to judge the probability that each student had passed the test. The manipulation of causal base rates was straightforward: Ajzen told one group that the students they saw had been drawn from a class in which 75% passed the exam, and told another group that the same students had been in a class in which only 25% passed. This is a powerful manipulation, because the base rate of passing suggests the immediate inference that the test that only 25% passed must have been brutally difficult. The difficulty of a test is, of course, one of the causal factors that determine every student's outcome. As expected, Ajzen's subjects were highly sensitive to the causal base rates, and every student was judged more likely to pass in the high-success condition than in the high-failure rate.

Ajzen used an ingenious method to suggest a noncausal base rate. He told his subjects that the students they saw had been drawn from a sample, which itself was constructed by selecting students who had passed or failed the exam. For example, the information for the high-failure group read as follows:

> The investigator was mainly interested in the causes of failure and constructed a sample in which 75% had failed the examination.

Note the difference. This base rate is a purely statistical fact about the ensemble from which cases have been drawn. It has no bearing on the question asked, which is whether the individual student passed or failed the test. As expected, the explicitly stated base rates had some effects on judgment, but they had much less impact than the statistically equivalent causal base rates. System 1 can deal with stories in which the elements are causally linked, but it is weak in statistical reasoning. For a Bayesian thinker, of course, the versions are equivalent. It is tempting to conclude that we have reached a satisfactory conclusion: causal base rates are used; merely statistical facts are (more or less) neglected. The next study, one of my all-time favorites, shows that the situation is rather more complex.

CAN PSYCHOLOGY BE TAUGHT?

The reckless cabdrivers and the impossibly difficult exam illustrate two inferences that people can draw from causal base rates: a stereotypical trait that is attributed to an individual, and a significant feature of the situation that affects an individual's outcome. The participants in the experiments made the correct inferences and their judgments improved. Unfortunately, things do not always work out so well. The classic experiment I describe next shows that people will not draw from base-rate information an inference that conflicts with other beliefs. It also supports the uncomfortable conclusion that teaching psychology is mostly a waste of time.

The experiment was conducted a long time ago by the social psychologist Richard Nisbett and his student Eugene Borgida, at the University of Michigan. They told students about the renowned "helping experiment" that had been conducted a few years earlier at New York University. Partici-

pants in that experiment were led to individual booths and invited to speak over the intercom about their personal lives and problems. They were to talk in turn for about two minutes. Only one microphone was active at any one time. There were six participants in each group, one of whom was a stooge. The stooge spoke first, following a script prepared by the experimenters. He described his problems adjusting to New York and admitted with obvious embarrassment that he was prone to seizures, especially when stressed. All the participants then had a turn. When the microphone was again turned over to the stooge, he became agitated and incoherent, said he felt a seizure coming on, and asked for someone to help him. The last words heard from him were, "C-could somebody-er-er-help-er-uh-uh-uh [choking sounds]. I . . . I'm gonna die-er-er-er I'm . . . gonna die-er-er-I seizure I-er [chokes, then quiet]." At this point the microphone of the next participant automatically became active, and nothing more was heard from the possibly dying individual.

What do you think the participants in the experiment did? So far as the participants knew, one of them was having a seizure and had asked for help. However, there were several other people who could possibly respond, so perhaps one could stay safely in one's booth. These were the results: only four of the fifteen participants responded immediately to the appeal for help. Six never got out of their booth, and five others came out only well after the "seizure victim" apparently choked. The experiment shows that individuals feel relieved of responsibility when they know that others have heard the same request for help.

Did the results surprise you? Very probably. Most of us think of ourselves as decent people who would rush to help in such a situation, and we expect other decent people to do the same. The point of the experiment, of course, was to show that this expectation is wrong. Even normal, decent people do not rush to help when they expect others to take on the unpleasantness of dealing with a seizure. And that means you, too.

Are you willing to endorse the following statement? "When I read the procedure of the helping experiment I thought I would come to the stranger's help immediately, as I probably would if I found myself alone with a seizure victim. I was probably wrong. If I find myself in a situation in which other people have an opportunity to help, I might not step forward. The presence of others would reduce my sense of personal responsibility more than I initially thought." This is what a teacher of psychology would hope you would learn. Would you have made the same inferences by yourself?

The psychology professor who describes the helping experiment wants the students to view the low base rate as causal, just as in the case of the fictitious Yale exam. He wants them to infer, in both cases, that a surprisingly high rate of failure implies a very difficult test. The lesson students are meant to take away is that some potent feature of the situation, such as the diffusion of responsibility, induces normal and decent people such as them to behave in a surprisingly unhelpful way.

Changing one's mind about human nature is hard work, and changing one's mind for the worse about oneself is even harder. Nisbett and Borgida suspected that students would resist the work and the unpleasantness. Of course, the students would be able and willing to recite the details of the helping experiment on a test, and would even repeat the "official" interpretation in terms of diffusion of responsibility. But did their beliefs about human nature really change? To find out, Nisbett and Borgida showed them videos of brief interviews allegedly conducted with two people who had participated in the New York study. The interviews were short and bland. The interviewees appeared to be nice, normal, decent people. They described their hobbies, their spare-time activities, and their plans for the future, which were entirely conventional. After watching the video of an interview, the students guessed how quickly that particular person had come to the aid of the stricken stranger.

To apply Bayesian reasoning to the task the students were assigned, you should first ask yourself what you would have guessed about the two individuals if you had not seen their interviews. This question is answered by consulting the base rate. We have been told that only 4 of the 15 participants in the experiment rushed to help after the first request. The probability that an unidentified participant had been immediately helpful is therefore 27%. Thus your prior belief about any unspecified participant should be that he did not rush to help. Next, Bayesian logic requires you to adjust your judgment in light of any relevant information about the individual. However, the videos were carefully designed to be uninformative; they provided no reason to suspect that the individuals would be either more or less helpful than a randomly chosen student. In the absence of useful new information, the Bayesian solution is to stay with the base rates.

Nisbett and Borgida asked two groups of students to watch the videos and predict the behavior of the two individuals. The students in the first group were told only about the procedure of the helping experiment, not about its results. Their predictions reflected their views of human nature and their understanding of the situation. As you might expect, they predicted that both individuals would immediately rush to the victim's aid. The second group of students knew both the procedure of the experiment and its results. The comparison of the predictions of the two groups provides an answer to a significant question: Did students learn from the results of the helping experiment anything that significantly changed their way of thinking? The answer is straightforward: they learned nothing at all. Their predictions about the two individuals were indistinguishable from the predictions made by students who had not been exposed to the statistical results of the experiment. They knew the base rate in the group from which the individuals had been drawn, but they remained convinced that the people they saw on the video had been quick to help the stricken stranger.

For teachers of psychology, the implications of this study are disheartening. When we teach our students about the behavior of people in the helping experiment, we expect them to learn something they had not known before; we wish to change how they think about people's behavior in a particular situation. This goal was not accomplished in the Nisbett-Borgida study, and there is no reason to believe that the results would have been different if they had chosen another surprising psychological experiment. Indeed, Nisbett and Borgida reported similar findings in teaching another study, in which mild social pressure caused people to accept much more painful electric shocks than most of us (and them) would have expected. Students who do not develop a new appreciation for the power of social setting have learned nothing of value from the experiment. The predictions they make about random strangers, or about their own behavior, indicate that they have not changed their view of how they would have behaved. In the words of Nisbett and Borgida, students "quietly exempt themselves" (and their friends and acquaintances) from the conclusions of experiments that surprise them. Teachers of psychology should not despair, however, because Nisbett and Borgida report a way to make their students appreciate the point of the helping experiment. They took a new group of students and taught them the procedure of the experiment but did not tell them the group results. They showed the two videos and simply told their students that the two individuals they had just seen had not helped the stranger, then asked them to guess the global results. The outcome was dramatic: the students' guesses were extremely accurate.

To teach students any psychology they did not know before, you must surprise them. But which surprise will do? Nisbett and Borgida found that when they presented their students with a surprising statistical fact, the students managed to learn nothing at all. But when the students were surprised by individual cases—two nice people who had not helped—they immediately made the generalization and inferred that helping is more difficult than they had thought. Nisbett and Borgida summarize the results in a memorable sentence:

> Subjects' unwillingness to deduce the particular from the general was matched only by their willingness to infer the general from the particular.

This is a profoundly important conclusion. People who are taught surprising statistical facts about human behavior may be impressed to the point of telling their friends about what they have heard, but this does not mean that their understanding of the world has really changed. The test of learning psychology is whether your understanding of situations you encounter has changed, not whether you have learned a new fact. There is a deep gap between our thinking about statistics and our thinking about individual cases. Statistical results with a causal interpretation have a stronger effect on our thinking than noncausal information. But even compelling causal statistics will not change long-held beliefs or beliefs rooted in personal experience. On the other hand, surprising individual cases have a powerful impact and are a more effective tool for teaching psychology because the incongruity must be resolved and embedded in a causal story. That is why this book contains questions that are addressed personally to the reader. You are more likely to learn something by finding surprises in your own behavior than by hearing surprising facts about people in general.

SPEAKING OF CAUSES AND STATISTICS

"We can't assume that they will really learn anything from mere statistics. Let's show them one or two representative individual cases to influence their System 1."

"No need to worry about this statistical information being ignored. On the contrary, it will immediately be used to feed a stereotype."

17

REGRESSION TO THE MEAN

I had one of the most satisfying eureka experiences of my career while teaching flight instructors in the Israeli Air Force about the psychology of effective training. I was telling them about an important principle of skill training: rewards for improved performance work better than punishment of mistakes. This proposition is supported by much evidence from research on pigeons, rats, humans, and other animals.

When I finished my enthusiastic speech, one of the most seasoned instructors in the group raised his hand and made a short speech of his own. He began by conceding that rewarding improved performance might be good for the birds, but he denied that it was optimal for flight cadets. This is what he said: "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver. The next time they try the same maneuver they usually do worse. On the other hand, I have often screamed into a cadet's earphone for bad execution, and in general he does better on his next try. So please don't tell us that reward works and punishment does not, because the opposite is the case."

This was a joyous moment of insight, when I saw in a new light a principle of statistics that I had been teaching for years. The instructor was right—but he was also completely wrong! His observation was astute and correct: occasions on which he praised a performance were likely to be followed by a disappointing performance, and punishments were typically followed by an improvement. But the inference he had drawn about the efficacy of reward and punishment was completely off the mark. What he had observed is known as *regression to the mean*, which in that case was due to random fluctuations in the quality of performance. Naturally, he praised only a cadet whose performance was far better than average. But the cadet was probably just lucky on that particular attempt and therefore likely to deteriorate regardless of whether or not he was praised. Similarly, the instructor would shout into a cadet's earphones only when the cadet's performance was unusually bad and therefore likely to improve regardless of what the instructor did. The instructor had attached a causal interpretation to the inevitable fluctuations of a random process.

The challenge called for a response, but a lesson in the algebra of prediction would not be enthusiastically received. Instead, I used chalk to mark a target on the floor. I asked every officer in the room to turn his back to the target and throw two coins at it in immediate succession, without looking. We measured the distances from the target and wrote the two results of each contestant on the blackboard. Then we rewrote the results in order, from the best to the worst performance on the first try. It was apparent that most (but not all) of those who had done best the first time deteriorated on their second try, and those who had done poorly on the first attempt generally improved. I pointed out to the instructors that what they saw on the board coincided with what we had heard about the performance of aerobatic maneuvers on successive attempts: poor performance was typically followed by improvement and good performance by deterioration, without any help from either praise or punishment.

The discovery I made on that day was that the flight instructors were trapped in an unfortunate contingency: because they punished cadets when performance was poor, they were mostly rewarded by a subsequent improvement, even if punishment was actually ineffective. Furthermore, the instructors were not alone in that predicament. I had stumbled onto a significant fact of the human condition: the feedback to which life exposes us is perverse. Because we tend to be nice to other people when they please us and nasty when they do not, we are statistically punished for being nice and rewarded for being nasty.

TALENT AND LUCK

A few years ago, John Brockman, who edits the online magazine *Edge*, asked a number of scientists to report their "favorite equation." These were my offerings:

success = talent + luck
great success = a little more talent + a lot of luck

The unsurprising idea that luck often contributes to success has surprising consequences when we apply it to the first two days of a high-level golf tournament. To keep things simple, assume that on both days the average score of the competitors was at par 72. We focus on a player who did very well on the first day, closing with a score of 66. What can we learn from that excellent score? An immediate inference is that the golfer is more talented than the average participant in the tournament. The formula for success suggests that another inference is equally justified: the golfer who did so well on day 1 probably enjoyed better-than-average luck on that day. If you accept that talent and luck both contribute to success, the conclusion that the successful golfer was lucky is as warranted as the conclusion that he is talented.

By the same token, if you focus on a player who scored 5 over par on that day, you have reason to infer both that he is rather weak *and* had a bad day. Of course, you know that neither of these inferences is certain. It is entirely possible that the player who scored 77 is actually very talented but had an exceptionally dreadful day. Uncertain though they are, the following inferences from the score on day 1 are plausible and will be correct more often than they are wrong.

above-average score on day 1 = above-average talent +
lucky on day 1

and

below-average score on day 1 = below-average talent +
unlucky on day 1

Now, suppose you know a golfer's score on day 1 and are asked to predict his score on day 2. You expect the golfer to retain the same level of talent on the second day, so your best guesses will be "above average" for the first player and "below average" for the second player. Luck, of course, is a different matter. Since you have no way of predicting the golfers' luck on the second (or any) day, your best guess must be that it will be average, neither good nor bad. This means that in the absence of any other information, your best guess about the players' score on day 2 should not be a repeat of their performance on day 1. This is the most you can say:

- The golfer who did well on day 1 is likely to be successful on day 2 as well, but less than on the first, because the unusual luck he probably enjoyed on day 1 is unlikely to hold.
- The golfer who did poorly on day 1 will probably be below average on day 2, but will improve, because his probable streak of bad luck is not likely to continue.

We also expect the difference between the two golfers to shrink on the second day, although our best guess is that the first player will still do better than the second.

My students were always surprised to hear that the best predicted performance on day 2 is more moderate, closer to the average than the evidence on which it is based (the score on day 1). This is why the pattern is called regression to the mean. The more extreme the original score, the more regression we expect, because an extremely good score suggests a very lucky day. The regressive prediction is reasonable, but its accuracy is not guaranteed. A few of the golfers who scored 66 on day 1 will do even better on the second day, if their luck improves. Most will do worse, because their luck will no longer be above average.

Now let us go against the time arrow. Arrange the players by their performance on day 2 and look at their performance on day 1. You will find precisely the same pattern of regression to the mean. The golfers who did best on day 2 were probably lucky on that day, and the best guess is that they had been less lucky and had done less well on day 1. The fact that you observe regression when you predict an early event from a later event should help convince you that regression does not have a causal explanation.

Regression effects are ubiquitous, and so are misguided causal stories to explain them. A well-known example is the "*Sports Illustrated* jinx," the claim that an athlete whose picture appears on the cover of the magazine is doomed to perform poorly the following season. Overconfidence and the pressure of meeting high expectations are often offered as explanations. But there is a simpler account of the jinx: an athlete who gets to be on the cover of *Sports Illustrated* must have performed exceptionally well in the preceding season, probably with the assistance of a nudge from luck—and luck is fickle.

I happened to watch the men's ski jump event in the Winter Olympics

while Amos and I were writing an article about intuitive prediction. Each athlete has two jumps in the event, and the results are combined for the final score. I was startled to hear the sportscaster's comments while athletes were preparing for their second jump: "Norway had a great first jump; he will be tense, hoping to protect his lead and will probably do worse" or "Sweden had a bad first jump and now he knows he has nothing to lose and will be relaxed, which should help him do better." The commentator had obviously detected regression to the mean and had invented a causal story for which there was no evidence. The story itself could even be true. Perhaps if we measured the athletes' pulse before each jump we might find that they are indeed more relaxed after a bad first jump. And perhaps not. The point to remember is that the change from the first to the second jump does not need a causal explanation. It is a mathematically inevitable consequence of the fact that luck played a role in the outcome of the first jump. Not a very satisfactory story—we would all prefer a causal account—but that is all there is.

UNDERSTANDING REGRESSION

Whether undetected or wrongly explained, the phenomenon of regression is strange to the human mind. So strange, indeed, that it was first identified and understood two hundred years after the theory of gravitation and differential calculus. Furthermore, it took one of the best minds of nineteenth-century Britain to make sense of it, and that with great difficulty.

Regression to the mean was discovered and named late in the nineteenth century by Sir Francis Galton, a half cousin of Charles Darwin and a renowned polymath. You can sense the thrill of discovery in an article he published in 1886 under the title "Regression towards Mediocrity in Hereditary Stature," which reports measurements of size in successive generations of seeds and in comparisons of the height of children to the height of their parents. He writes about his studies of seeds:

> They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small . . . The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.

Galton obviously expected his learned audience at the Royal Institution—the oldest independent research society in the world—to be as surprised by his "noteworthy observation" as he had been. What is truly noteworthy is that he was surprised by a statistical regularity that is as common as the air we breathe. Regression effects can be found wherever we look, but we do not recognize them for what they are. They hide in plain sight. It took Galton several years to work his way from his discovery of filial regression in size to the broader notion that regression inevitably occurs when the correlation between two measures is less than perfect, and he needed the help of the most brilliant statisticians of his time to reach that conclusion.

One of the hurdles Galton had to overcome was the problem of measuring regression between variables that are measured on different scales, such as weight and piano playing. This is done by using the population as a standard of reference. Imagine that weight and piano playing have been measured for 100 children in all grades of an elementary school, and that they have been ranked from high to low on each measure. If Jane ranks third in piano playing and twenty-seventh in weight, it is appropriate to say that she is a better pianist than she is tall. Let us make some assumptions that will simplify things:

At any age,

- Piano-playing success depends only on weekly hours of practice.
- Weight depends only on consumption of ice cream.
- Ice cream consumption and weekly hours of practice are unrelated.

Now, using ranks (or the *standard scores* that statisticians prefer), we can write some equations:

weight = age + ice cream consumption
piano playing = age + weekly hours of practice

You can see that there will be regression to the mean when we predict piano playing from weight, or vice versa. If all you know about Tom is that he ranks twelfth in weight (well above average), you can infer (statistically) that he is probably older than average and also that he probably consumes more ice cream than other children. If all you know about Barbara is that she is eighty-fifth in piano (far below the average of the group), you can infer that she is likely to be young and that she is likely to practice less than most other children.

The *correlation coefficient* between two measures, which varies between 0 and 1, is a measure of the relative weight of the factors they share. For example, we all share half our genes with each of our parents, and for traits in which environmental factors have relatively little influence, such as height, the correlation between parent and child is not far from .50. To appreciate the meaning of the correlation measure, the following are some examples of coefficients:

- The correlation between the size of objects measured with precision in English or in metric units is 1. Any factor that influences one measure also influences the other; 100% of determinants are shared.
- The correlation between self-reported height and weight among adult American males is .41. If you included women and children, the correlation would be much higher, because individuals' gender and age influence both their height and their weight, boosting the relative weight of shared factors.
- The correlation between SAT scores and college GPA is approximately .60. However, the correlation between aptitude tests and success in graduate school is much lower, largely because measured aptitude varies little in this selected group. If everyone has similar aptitude, differences in this measure are unlikely to play a large role in measures of success.
- The correlation between income and education level in the United States is approximately .40.
- The correlation between family income and the last four digits of their phone number is 0.

It took Francis Galton several years to figure out that correlation and regression are not two concepts—they are different perspectives on the same concept. The general rule is straightforward but has surprising consequences: whenever the correlation between two scores is imperfect, there will be regression to the mean. To illustrate Galton's insight, take a proposition that most people find quite interesting:

Highly intelligent women tend to marry men who are less intelligent than they are.

You can get a good conversation started at a party by asking for an explanation, and your friends will readily oblige. Even people who have had some exposure to statistics will spontaneously interpret the statement in causal terms. Some may think of highly intelligent women wanting to avoid the competition of equally intelligent men, or being forced to compromise in their choice of spouse because intelligent men do not want to compete with intelligent women. More far-fetched explanations will come up at a good party. Now consider this statement:

> The correlation between the intelligence scores of spouses is less than perfect.

This statement is obviously true and not interesting at all. Who would expect the correlation to be perfect? There is nothing to explain. But the statement you found interesting and the statement you found trivial are algebraically equivalent. If the correlation between the intelligence of spouses is less than perfect (and if men and women on average do not differ in intelligence), then it is a mathematical inevitability that highly intelligent women will be married to husbands who are on average less intelligent than they are (and vice versa, of course). The observed regression to the mean cannot be more interesting or more explainable than the imperfect correlation.

You probably sympathize with Galton's struggle with the concept of regression. Indeed, the statistician David Freedman used to say that if the topic of regression comes up in a criminal or civil trial, the side that must explain regression to the jury will lose the case. Why is it so hard? The main reason for the difficulty is a recurrent theme of this book: our mind is strongly biased toward causal explanations and does not deal well with "mere statistics." When our attention is called to an event, associative memory will look for its cause—more precisely, activation will automatically spread to any cause that is already stored in memory. Causal explanations will be evoked when regression is detected, but they will be wrong because the truth is that regression to the mean has an explanation but does not have a cause. The event that attracts our attention in the golfing tournament is the frequent deterioration of the performance of the golfers who were successful on day 1. The best explanation of it is that those golfers were unusually lucky that day, but this explanation lacks the causal force that our minds prefer. Indeed, we pay people quite well to provide interesting explanations of regression effects. A business commentator who correctly announces that "the business did better this year because it had done poorly last year" is likely to have a short tenure on the air.

Our difficulties with the concept of regression originate with both System 1 and System 2. Without special instruction, and in quite a few cases even after some statistical instruction, the relationship between correlation and regression remains obscure. System 2 finds it difficult to understand and learn. This is due in part to the insistent demand for causal interpretations, which is a feature of System 1.

> Depressed children treated with an energy drink improve significantly over a three-month period.

I made up this newspaper headline, but the fact it reports is true: if you treated a group of depressed children for some time with an energy drink, they would show a clinically significant improvement. It is also the case that depressed children who spend some time standing on their head or hug a cat for twenty minutes a day will also show improvement. Most readers of such headlines will automatically infer that the energy drink or the cat hugging caused an improvement, but this conclusion is completely unjustified. Depressed children are an extreme group, they are more depressed than most other children—and extreme groups regress to the mean over time. The correlation between depression scores on successive occasions of testing is less than perfect, so there will be regression to the mean: depressed children will get somewhat better over time even if they hug no cats and drink no Red Bull. In order to conclude that an energy drink—or any other treatment—is effective, you must compare a group of patients who receive this treatment to a "control group" that receives no treatment (or, better, receives a placebo). The control group is expected to improve by regression alone, and the aim of the experiment is to determine whether the treated patients improve more than regression can explain.

Incorrect causal interpretations of regression effects are not restricted to readers of the popular press. The statistician Howard Wainer has drawn up a long list of eminent researchers who have made the same mistake—confusing mere correlation with causation. Regression effects are a common source of trouble in research, and experienced scientists develop a healthy fear of the trap of unwarranted causal inference.

One of my favorite examples of the errors of intuitive prediction is adapted from Max Bazerman's excellent text *Judgment in Managerial Decision Making*:

You are the sales forecaster for a department store chain. All stores are similar in size and merchandise selection, but their sales differ because of location, competition, and random factors. You are given the results for 2011 and asked to forecast sales for 2012. You have been instructed to accept the overall forecast of economists that sales will increase overall by 10%. How would you complete the following table?

Store	2011	2012
1	$11,000,000	________
2	$23,000,000	________
3	$18,000,000	________
4	$29,000,000	________
Total	$81,000,000	$89,100,000

Having read this chapter, you know that the obvious solution of adding 10% to the sales of each store is wrong. You want your forecasts to be regressive, which requires adding more than 10% to the low-performing branches and adding less (or even subtracting) to others. But if you ask other people, you are likely to encounter puzzlement: Why do you bother them with an obvious question? As Galton painfully discovered, the concept of regression is far from obvious.

SPEAKING OF REGRESSION TO MEDIOCRITY

"She says experience has taught her that criticism is more effective than praise. What she doesn't understand is that it's all due to regression to the mean."

"Perhaps his second interview was less impressive than the first because he was afraid of disappointing us, but more likely it was his first that was unusually good."

"Our screening procedure is good but not perfect, so we should anticipate regression. We shouldn't be surprised that the very best candidates often fail to meet our expectations."