10 The Law of Small Numbers A study of the incidence of kidney cancer in the 3,141 counties of the United States reveals a remarkable pattern. The counties in which the incidence of kidney cancer is lowest are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. What do you make of this? Your mind has been very active in the last few seconds, and it was mainly a System 2 operation. You deliberately searched memory and formulated hypotheses. Some effort was involved; your pupils dilated, and your heart rate increased measurably. But System 1 was not idle: the operation of System 2 depended on the facts and suggestions retrieved from associative memory. You probably rejected the idea that Republican politics provide protection against kidney cancer. Very likely, you ended up focusing on the fact that the counties with low incidence of cancer are mostly rural. The witty statisticians Howard Wainer and Harris Zwerling, from whom I learned this example, commented, “It is both easy and tempting to infer that their low cancer rates are directly due to the clean living of the rural lifestyle—no air pollution, no water pollution, access to fresh food without additives.” This makes perfect sense. Now consider the counties in which the incidence of kidney cancer is highest. These ailing counties tend to be mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. Tongue-in-cheek, Wainer and Zwerling comment: “It is easy to infer that their high cancer rates might be directly due to the poverty of the rural lifestyle—no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco.” Something is wrong, of course. The rural lifestyle cannot explain both very high and very low incidence of kidney cancer. The key factor is not that the counties were rural or predominantly Republican. It is that rural counties have small populations. And the main lesson to be learned is not about epidemiology, it is about the difficult relationship between our mind and statistics. System 1 is highly adept in one form of thinking—it automatically and effortlessly identifies causal connections between events, sometimes even when the connection is spurious. When told about the high-incidence counties, you immediately assumed that these counties are different from other counties for a reason, that there must be a cause that explains this difference. As we shall see, however, System 1 is inept when faced with “merely statistical” facts, which change the probability of outcomes but do not cause them to happen. A random event, by definition, does not lend itself to explanation, but collections of random events do behave in a highly regular fashion. Imagine a large urn filled with marbles. Half the marbles are red, half are white. Next, imagine a very patient person (or a robot) who blindly draws 4 marbles from the urn, records the number of red balls in the sample, throws the balls back into the urn, and then does it all again, many times. If you summarize the results, you will find that the outcome “2 red, 2 white” occurs (almost exactly) 6 times as often as the outcome “4 red” or “4 white.” This relationship is a mathematical fact. You can predict the outcome of repeated sampling from an urn just as confidently as you can predict what will happen if you hit an egg with a hammer. You cannot predict every detail of how the shell will shatter, but you can be sure of the general idea. There is a difference: the satisfying sense of causation that you experience when thinking of a hammer hitting an egg is altogether absent when you think about sampling. A related statistical fact is relevant to the cancer example. From the same urn, two very patient marble counters take turns. Jack draws 4 marbles on each trial, Jill draws 7. They both record each time they observe a homogeneous sample—all white or all red. If they go on long enough, Jack will observe such extreme outcomes more often than Jill—by a factor of 8 (the expected percentages are 12.5% and 1.56%). Again, no hammer, no causation, but a mathematical fact: samples of 4 marbles yield extreme results more often than samples of 7 marbles do. Now imagine the population of the United States as marbles in a giant urn. Some marbles are marked KC, for kidney cancer. You draw samples of marbles and populate each county in turn. Rural samples are smaller than other samples. Just as in the game

of Jack and Jill, extreme outcomes (very high and/or very low cancer rates) are most likely to be found in sparsely populated counties. This is all there is to the story. We started from a fact that calls for a cause: the incidence of kidney cancer varies widely across counties and the differences are systematic. The explanation I offered is statistical: extreme outcomes (both high and low) are more likely to be found in small than in large samples. This explanation is not causal. The small population of a county neither causes nor prevents cancer; it merely allows the incidence of cancer to be much higher (or much lower) than it is in the larger population. The deeper truth is that there is nothing to explain. The incidence of cancer is not truly lower or higher than normal in a county with a small population, it just appears to be so in a particular year because of an accident of sampling. If we repeat the analysis next year, we will observe the same general pattern of extreme results in the small samples, but the counties where cancer was common last year will not necessarily have a high incidence this year. If this is the case, the differences between dense and rural counties do not really count as facts: they are what scientists call artifacts, observations that are produced entirely by some aspect of the method of research—in this case, by differences in sample size. The story I have told may have surprised you, but it was not a revelation. You have long known that the results of large samples deserve more trust than smaller samples, and even people who are innocent of statistical knowledge have heard about this law of large numbers. But "knowing" is not a yes-no affair and you may find that the following statements apply to you: The feature "sparsely populated" did not immediately stand out as relevant when you read the epidemiological story. You were at least mildly surprised by the size of the difference between samples of 4 and samples of 7. Even now, you must exert some mental effort to see that the following two statements mean exactly the same thing: Large samples are more precise than small samples. Small samples yield extreme results more often than large samples do. The first statement has a clear ring of truth, but until the second version makes intuitive sense, you have not truly understood the first. The bottom line: yes, you did know that the results of large samples are more precise, but you may now realize that you did not know it very well. You are not alone. The first study that Amos and I did together showed that even sophisticated researchers have poor intuitions and a wobbly understanding of sampling effects. THE LAW OF SMALL NUMBERS My collaboration with Amos in the early 1970s began with a discussion of the claim that people who have had no training in statistics are good "intuitive statisticians." He told my seminar and me of researchers at the University of Michigan who were generally optimistic about intuitive statistics. I had strong feelings about that claim, which I took personally: I had recently discovered that I was not a good intuitive statistician, and I did not believe that I was worse than others. For a research psychologist, sampling variation is not a curiosity; it is a nuisance and a costly obstacle, which turns the undertaking of every research project into a gamble. Suppose that you wish to confirm the hypothesis that the vocabulary of the average six-year-old girl is larger than the vocabulary of an average boy of the same age. The hypothesis is true in the population; the average vocabulary of girls is indeed larger. Girls and boys vary a great deal, however, and by the luck of the draw you could select a sample in which the difference is inconclusive, or even one in which boys actually score higher. If you are the researcher, this outcome is costly to you because you have wasted time and effort, and failed to confirm a hypothesis that was in fact true. Using a sufficiently large sample is the only way to reduce the risk. Researchers who pick too small a sample leave themselves at the mercy of sampling luck. The risk of error can be estimated for any given sample size by a fairly simple procedure. Traditionally, however, psychologists do not use calculations to decide on a sample size. They use their judgment, which is commonly flawed. An article I had read shortly before the debate with Amos demonstrated the mistake that researchers made (they still do) by a dramatic observation. The author pointed out that psychologists commonly chose samples so small that they exposed themselves to a 50% risk of failing to confirm their true hypotheses! No researcher in his right mind would accept such a risk. A plausible explanation was that psychologists' decisions about sample size reflected

Kahneman, Daniel. Thinking, Fast and Slow (pp. 111-prevalent intuitive misconceptions of the extent of sampling variation. The article shocked me, because it explained some troubles I had had in my own research. Like most research psychologists, I had routinely chosen samples that were too small and had often obtained results that made no sense. Now I knew why: the odd results were actually artifacts of my research method. My mistake was particularly embarrassing because I taught statistics and knew how to compute the sample size that would reduce the risk of failure to an acceptable level. But I had never chosen a sample size by computation. Like my colleagues, I had trusted tradition and my intuition in planning my experiments and had never thought seriously about the issue. When Amos visited the seminar, I had already reached the conclusion that my intuitions were deficient, and in the course of the seminar we quickly agreed that the Michigan optimists were wrong. Amos and I set out to examine whether I was the only fool or a member of a majority of fools, by testing whether researchers selected for mathematical expertise would make similar mistakes. We developed a questionnaire that described realistic research situations, including replications of successful experiments. It asked the researchers to choose sample sizes, to assess the risks of failure to which their decisions exposed them, and to provide advice to hypothetical graduate students planning their research. Amos collected the responses of a group of sophisticated participants (including authors of two statistical textbooks) at a meeting of the Society of Mathematical Psychology. The results were straightforward: I was not the only fool. Every one of the mistakes I had made was shared by a large majority of our respondents. It was evident that even the experts paid insufficient attention to sample size. Amos and I called our first joint article "Belief in the Law of Small Numbers." We explained, tongue-in-cheek, that "intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well." We also included a strongly worded recommendation that researchers regard their "statistical intuitions with proper suspicion and replace impression formation by computation whenever possible." A BIAS OF CONFIDENCE OVER DOUBT In a telephone poll of 300 seniors, 60% support the president. If you had to summarize the message of this sentence in exactly three words, what would they be? Almost certainly you would choose "elderly support president." These words provide the gist of the story. The omitted details of the poll, that it was done on the phone with a sample of 300, are of no interest in themselves; they provide background information that attracts little attention. Your summary would be the same if the sample size had been different. Of course, a completely absurd number would draw your attention ("a telephone poll of 6 [or 60 million] elderly voters …"). Unless you are a professional, however, you may not react very differently to a sample of 150 and to a sample of 3,000. That is the meaning of the statement that "people are not adequately sensitive to sample size." The message about the poll contains information of two kinds: the story and the source of the story. Naturally, you focus on the story rather than on the reliability of the results. When the reliability is obviously low, however, the message will be discredited. If you are told that "a partisan group has conducted a flawed and biased poll to show that the elderly support the president …" you will of course reject the findings of the poll, and they will not become part of what you believe. Instead, the partisan poll and its false results will become a new story about political lies. You can choose to disbelieve a message in such clear-cut cases. But do you discriminate sufficiently between "I read in The New York Times …" and "I heard at the watercooler …"? Can your System 1 distinguish degrees of belief? The principle of WY SIATI suggests that it cannot. As I described earlier, System 1 is not prone to doubt. It suppresses ambiguity and spontaneously constructs stories that are as coherent as possible. Unless the message is immediately negated, the associations that it evokes will spread as if the message were true.

System 2 is capable of doubt, because it can maintain incompatible possibilities at the same time. However, sustaining doubt is harder work than sliding into certainty. The law of small numbers is a manifestation of a general bias that favors certainty over doubt, which will turn up in many guises in following chapters. The strong bias toward believing that small samples closely resemble the population from which they are drawn is also part of a larger story: we are prone to exaggerate the consistency and coherence of what we see. The exaggerated faith of researchers in what can be learned from a few observations is closely related to the halo effect, the sense we often get that we know and understand

a person about whom we actually know very little. System 1 runs ahead of the facts in constructing a rich image on the basis of scraps of evidence. A machine for jumping to conclusions will act as if it believed in the law of small numbers. More generally, it will produce a representation of reality that makes too much sense. CAUSE AND CHANCE The associative machinery seeks causes. The difficulty we have with statistical regularities is that they call for a different approach. Instead of focusing on how the event at hand came to be, the statistical view relates it to what could have happened instead. Nothing in particular caused it to be what it is—chance selected it from among its alternatives. Our predilection for causal thinking exposes us to serious mistakes in evaluating the randomness of truly random events. For an example, take the sex of six babies born in sequence at a hospital. The sequence of boys and girls is obviously random; the events are independent of each other, and the number of boys and girls who were born in the hospital in the last few hours has no effect whatsoever on the sex of the next baby. Now consider three possible sequences: BBBGGG GGGGGG BGBBGB Are the sequences equally likely? The intuitive answer—"of course not!"—is false. Because the events are independent and because the outcomes B and G are (approximately) equally likely, then any possible sequence of six births is as likely as any other. Even now that you know this conclusion is true, it remains counterintuitive, because only the third sequence appears random. As expected, B GBBGB is judged much more likely than the other two sequences. We are pattern seekers, believers in a coherent world, in which regularities (such as a sequence of six girls) appear not by accident but as a result of mechanical causality or of someone's intention. We do not expect to see regularity produced by a random process, and when we detect what appears to be a rule, we quickly reject the idea that the process is truly random. Random processes produce many sequences that convince people that the process is not random after all. You can see why assuming causality could have had evolutionary advantages. It is part of the general vigilance that we have inherited from ancestors. We are automatically on the lookout for the possibility that the environment has changed. Lions may appear on the plain at random times, but it would be safer to notice and respond to an apparent increase in the rate of appearance of prides of lions, even if it is actually due to the fluctuations of a random process. The widespread misunderstanding of randomness sometimes has significant consequences. In our article on representativeness, Amos and I cited the statistician William Feller, who illustrated the ease with which people see patterns where none exists. During the intensive rocket bombing of London in World War II, it was generally believed that the bombing could not be random because a map of the hits revealed conspicuous gaps. Some suspected that German spies were located in the unharmed areas. A careful statistical analysis revealed that the distribution of hits was typical of a random process—and typical as well in evoking a strong impression that it was not random. "To the untrained eye," Feller remarks, "randomness appears as regularity or tendency to cluster." I soon had an occasion to apply what I had learned from Feller. The Yom Kippur War broke out in 1973, and my only significant contribution to the war effort was to advise high officers in the Israeli Air Force to stop an investigation. The air war initially went quite badly for Israel, because of the unexpectedly good performance of Egyptian ground-to-air missiles. Losses were high, and they appeared to be unevenly distributed. I was told of two squadrons flying from the same base, one of which had lost four planes while the other had lost none. An inquiry was initiated in the hope of learning what it was that the unfortunate squadron was doing wrong. There was no prior reason to believe that one of the squadrons was more effective than the other, and no operational differences were found, but of course the lives of the pilots differed in many random ways, including, as I recall, how often they went home between missions and something about the conduct of debriefings. My advice was that the command should accept that the different outcomes were due to blind luck, and that the interviewing of the pilots should stop. I reasoned that luck was the most likely answer, that a random search for a nonobvious cause was hopeless, and that in the meantime the pilots in the squadron that had sustained losses did not need the extra burden of being made to feel that they and their dead friends were at fault. Some years later, Amos and his students Tom Gilovich and Robert Vallone caused a stir with their study of misperceptions of randomness in basketball. The "fact" that players occasionally acquire a hot hand is generally accepted by players,

coaches, and fans. The inference is irresistible: a player sinks three or four baskets in a row and you cannot help forming the causal judgment that this player is now hot, with a temporarily increased propensity to score. Players on both teams adapt to this judgment—teammates are more likely to pass to the hot scorer and the defense is more likely to double-team. Analysis of thousands of sequences of shots led to a disappointing conclusion: there is no such thing as a hot hand in professional basketball, either in shooting from the field or scoring from the foul line. Of course, some players are more accurate than others, but the sequence of successes and missed shots satisfies all tests of randomness. The hot hand is entirely in the eye of the beholders, who are consistently too quick to perceive order and causality in randomness. The hot hand is a massive and widespread cognitive illusion. The public reaction to this research is part of the story. The finding was picked up by the press because of its surprising conclusion, and the general response was disbelief. When the celebrated coach of the Boston Celtics, Red Auerbach, heard of Gilovich and his study, he responded, "Who is this guy? So he makes a study. I couldn't care less." The tendency to see patterns in randomness is overwhelming—certainly more impressive than a guy making a study. The illusion of pattern affects our lives in many ways off the basketball court. How many good years should you wait before concluding that an investment adviser is unusually skilled? How many successful acquisitions should be needed for a board of directors to believe that the CEO has extraordinary flair for such deals? The simple answer to these questions is that if you follow your intuition, you will more often than not err by misclassifying a random event as systematic. We are far too willing to reject the belief that much of what we see in life is random. I began this chapter with the example of cancer incidence across the United States. The example appears in a book intended for statistics teachers, but I learned about it from an amusing article by the two statisticians I quoted earlier, Howard Wainer and Harris Zwerling. Their essay focused on a large investment, some $1.7 billion, which the Gates Foundation made to follow up intriguing findings on the characteristics of the most successful schools. Many researchers have sought the secret of successful education by identifying the most successful schools in the hope of discovering what distinguishes them from others. One of the conclusions of this research is that the most successful schools, on average, are small. In a survey of 1,662 schools in Pennsylvania, for instance, 6 of the top 50 were small, which is an overrepresentation by a factor of 4. These data encouraged the Gates Foundation to make a substantial investment in the creation of small schools, sometimes by splitting large schools into smaller units. At least half a dozen other prominent institutions, such as the Annenberg Foundation and the Pew Charitable Trust, joined the effort, as did the U.S. Department of Education's Smaller Learning Communities Program. This probably makes intuitive sense to you. It is easy to construct a causal story that explains how small schools are able to provide superior education and thus produce high-achieving scholars by giving them more personal attention and encouragement than they could get in larger schools. Unfortunately, the causal analysis is pointless because the facts are wrong. If the statisticians who reported to the Gates Foundation had asked about the characteristics of the worst schools, they would have found that bad schools also tend to be smaller than average. The truth is that small schools are not better on average; they are simply more variable. If anything, say Wainer and Zwerling, large schools tend to produce better results, especially in higher grades where a variety of curricular options is valuable. Thanks to recent advances in cognitive psychology, we can now see clearly what Amos and I could only glimpse: the law of small numbers is part of two larger stories about the workings of the mind. The exaggerated faith in small samples is only one example of a more general illusion—we pay more attention to the content of messages than to information about their reliability, and as a result end up with a view of the world around us that is simpler and more coherent than the data justify. Jumping to conclusions is a safer sport in the world of our imagination than it is in reality. Statistics produce many observations that appear to beg for causal explanations but do not lend themselves to such explanations. Many facts of the world are due to chance, including accidents of sampling. Causal explanations of chance events are inevitably wrong.

SPEAKING OF THE LAW OF SMALL NUMBERS “Yes, the studio has had three successful films since the new CEO took over. But it is too early to declare he has a hot hand.” “I won’t believe that the new trader is a genius before consulting a statistician who could estimate the likelihood of his streak being a chance event.” “The sample of observations is too small to make any inferences. Let’s not follow the law of small numbers.” “I plan to keep the results of the experiment secret until we have a sufficiently large sample. Otherwise we will face pressure to reach a conclusion prematurely.”

11 Anchors Amos and I once rigged a wheel of fortune. It was marked from 0 to 100, but we had it built so that it would stop only at 10 or 65. We recruited students of the University of Oregon as participants in our experiment. One of us would stand in front of a small group, spin the wheel, and ask them to write down the number on which the wheel stopped, which of course was either 10 or 65. We then asked them two questions: Is the percentage of African nations among UN members larger or smaller than the number you just wrote? What is your best guess of the percentage of African nations in the UN? The spin of a wheel of fortune—even one that is not rigged—cannot possibly yield useful information about anything, and the participants in our experiment should simply have ignored it. But they did not ignore it. The average estimates of those who saw 10 and 65 were 25% and 45%, respectively. The phenomenon we were studying is so common and so important in the everyday world that you should know its name: it is an anchoring effect. It occurs when people consider a particular value for an unknown quantity before estimating that quantity. What happens is one of the most reliable and robust results of experimental psychology: the estimates stay close to the number that people considered—hence the image of an anchor. If you are asked whether Gandhi was more than 114 years old when he died you will end up with a much higher estimate of his age at death than you would if the anchoring question referred to death at 35. If you consider how much you should pay for a house, you will be influenced by the asking price. The same house will appear more valuable if its listing price is high than if it is low, even if you are determined to resist the influence of this number; and so on—the list of anchoring effects is endless. Any number that you are asked to consider as a possible solution to an estimation problem will induce an anchoring effect. We were not the first to observe the effects of anchors, but our experiment was the first demonstration of its absurdity: people’s judgments were influenced by an obviously uninformative number. There was no way to describe the anchoring effect of a wheel of fortune as reasonable. Amos and I published the experiment in our

Science paper, and it is one of the best known of the findings we reported there. There was only one trouble: Amos and I did not fully agree on the psychology of the anchoring effect. He supported one interpretation, I liked another, and we never found a way to settle the argument. The problem was finally solved decades later by the efforts of numerous investigators. It is now clear that Amos and I were both right. Two different mechanisms produce anchoring effects—one for each system. There is a form of anchoring that occurs in a deliberate process of adjustment, an operation of System 2. And there is anchoring that occurs by a priming effect, an automatic manifestation of System 1. ANCHORING AS ADJUSTMENT Amos liked the idea of an adjust-and-anchor heuristic as a strategy for estimating uncertain quantities: start from an anchoring number, assess whether it is too high or too low, and gradually adjust your estimate by mentally "moving" from the anchor. The adjustment typically ends prematurely, because people stop when they are no longer certain that they should move farther. Decades after our disagreement, and years after Amos's death, convincing evidence of such a process was offered independently by two psychologists who had worked closely with Amos early in their careers: Eldar Shafir and Tom Gilovich together with their own students—Amos's intellectual grandchildren! To get the idea, take a sheet of paper and draw a 2½-inch line going up, starting at the bottom of the page—without a ruler. Now take another sheet, and start at the top and draw a line going down until it is 2½ inches from the bottom. Compare the lines. There is a good chance that your first estimate of 2½ inches was shorter than the second. The reason is that you do not know exactly what such a line looks like; there is a range of uncertainty. You stop near the bottom of the region of uncertainty when you start from the bottom of the page and near the top of the region when you start from the top. Robyn Le Boeuf and Shafir found many examples of that mechanism in daily experience. Insufficient adjustment neatly explains why you are likely to drive too fast when you come off the highway onto city streets—especially if you are talking with someone as you drive. Insufficient adjustment is also a source of tension between exasperated parents and teenagers who enjoy loud music in their room. Le Boeuf and Shafir note that a "well-intentioned child who turns down exceptionally loud music to meet a parent's demand that it be played at a 'reasonable' volume may fail to adjust sufficiently from a high anchor, and may feel that genuine attempts at compromise are being overlooked." The driver and the child both deliberately adjust down, and both fail to adjust enough. Now consider these questions: When did George Washington become president? What is the boiling temperature of water at the top of Mount Everest? The first thing that happens when you consider each of these questions is that an anchor comes to your mind, and you know both that it is wrong and the direction of the correct answer. You know immediately that George Washington became president after 1776, and you also know that the boiling temperature of water at the top of Mount Everest is lower than 100°C. You have to adjust in the appropriate direction by finding arguments to move away from the anchor. As in the case of the lines, you are likely to stop when you are no longer sure you should go farther—at the near edge of the region of uncertainty. Nick Epley and Tom Gilovich found evidence that adjustment is a deliberate attempt to find reasons to move away from the anchor: people who are instructed to shake their head when they hear the anchor, as if they rejected it, move farther from the anchor, and people who nod their head show enhanced anchoring. Epley and Gilovich also confirmed that adjustment is an effortful operation. People adjust less (stay closer to the anchor) when their mental resources are depleted, either because their memory is loaded with digits or because they are slightly drunk. Insufficient adjustment is a failure of a weak or lazy System 2. So we now know that Amos was right for at least some cases of anchoring, which involve a deliberate System 2 adjustment in a specified direction from an anchor. ANCHORING AS PRIMING EFFECT When Amos and I debated anchoring, I agreed that adjustment sometimes occurs, but I was uneasy. Adjustment is a deliberate and conscious activity, but in most cases of anchoring there is no corresponding subjective experience. Consider these two questions: Was Gandhi more or less than 144 years old when he died? How old was Gandhi when he died? Did you produce your estimate by adjusting down from 144? Probably not, but the absurdly high number still affected your estimate. My hunch was that anchoring is a case of suggestion. This is the word we use when someone causes us to see, hear, or feel something by merely bringing it to mind. For example, the question “Do you now feel a slight numbness in your left leg?” always prompts quite a few people to report that their left leg does indeed feel a little strange. Amos was more conservative than I was about hunches, and he correctly pointed out that appealing to suggestion did not help us understand anchoring, because we did not know how to explain suggestion. I had to agree that he was right, but I never became enthusiastic about the idea of insufficient adjustment as the sole cause of anchoring effects. We conducted many inconclusive experiments in an effort to understand anchoring, but we failed and eventually gave up the idea of writing more about it. The puzzle that defeated us is now solved, because the concept of suggestion is no longer obscure: suggestion is a priming effect, which selectively evokes compatible evidence. You did not believe for a moment that Gandhi lived for 144 years, but your associative machinery surely generated an impression of a very ancient person. System 1 understands sentences by trying to make them true, and the selective activation of compatible thoughts produces a family of systematic errors that make us gullible and prone to believe too strongly whatever we believe. We can now see why Amos and I did not realize that there were two types of anchoring: the research techniques and theoretical ideas we needed did not yet exist. They were developed, much later, by other people. A process that resembles suggestion is indeed at work in many situations: System 1 tries its best to construct a world in which the anchor is the true number. This is one of the manifestations of

associative coherence that I described in the first part of the book. The German psychologists Thomas Mussweiler and Fritz Strack offered the most compelling demonstrations of the role of associative coherence in anchoring. In one experiment, they asked an anchoring question about temperature: “Is the annual mean temperature in Germany higher or lower than 20°C (68°F)?” or “Is the annual mean temperature in Germany higher or lower than 5°C (41°F)?” All participants were then briefly shown words that they were asked to identify. The researchers found that 68°F made it easier to recognize summer words (like sun and beach), and 40°F facilitated winter words (like frost and ski). The selective activation of compatible memories explains anchoring: the high and the low numbers activate different sets of ideas in memory. The estimates of annual temperature draw on these biased samples of ideas and are therefore biased as well. In another elegant study in the same vein, participants were asked about the average price of German cars. A high anchor selectively primed the names of luxury brands (Mercedes, Audi), whereas the low anchor primed brands associated with mass-market cars (Volkswagen). We saw earlier that any prime will tend to evoke information that is compatible with it. Suggestion and anchoring are both explained by the same automatic operation of System 1. Although I did not know how to prove it at the time, my hunch about the link between anchoring and suggestion turned out to be correct. THE ANCHORING INDEX Many psychological phenomena can be demonstrated experimentally, but few can actually be measured. The effect of anchors is an exception. Anchoring can be measured, and it is an impressively large effect. Some visitors at the San Francisco Exploratorium were asked the following two questions: Is the height of the tallest redwood more or less than 1,200 feet? What is your best guess about the height of the tallest redwood? The “high anchor” in this experiment was 1,200 feet. For other participants, the first question referred to a “low anchor” of 180 feet. The difference between the two anchors was 1,020 feet. As expected, the two groups produced very different mean estimates: 844 and 282 feet. The difference between them was 562 feet. The anchoring index is simply the ratio of the two differences (562/1,020) expressed as a percentage: 55%. The anchoring measure would be 100% for people who slavishly adopt the anchor as an estimate, and zero for people who are able to ignore the anchor altogether. The value of 55% that was observed in this example is typical. Similar values have been observed in numerous other problems. The anchoring effect is not a laboratory curiosity; it can be just as strong in the real world. In an experiment conducted some years ago, real-estate agents were given an opportunity to assess the value of a house that was actually on the market. They visited the house and studied a comprehensive booklet of information that included an asking price. Half the agents saw an asking price that was substantially higher than the listed price of the house; the other half saw an asking price that was substantially lower. Each agent gave her opinion about a reasonable buying price for the house and the lowest price at which she would agree to sell the house if she owned it. The agents were then asked about the factors that had affected their judgment. Remarkably, the asking price was not one of these factors; the agents took pride in their ability to ignore it. They insisted that the listing price had no effect on their responses, but they were wrong: the anchoring effect was 41%. Indeed, the professionals were almost as susceptible to anchoring effects as business school students with no real-estate experience, whose anchoring index was 48%. The only difference between the two groups was that the students conceded that they were influenced by the anchor, while the professionals denied that influence. Powerful anchoring effects are found in decisions that people make about money, such as when they choose how much to contribute to a cause. To demonstrate this effect, we told participants in the Exploratorium study about the environmental damage caused by oil tankers in the Pacific Ocean and asked about their willingness to make an annual contribution "to save 50,000 offshore Pacific Coast seabirds from small offshore oil spills, until ways are found to prevent spills or require tanker owners to pay for the operation." This question requires intensity matching: the respondents are asked, in effect, to find the dollar amount of a contribution that matches the intensity of their feelings about the plight of the seabirds. Some of the visitors were first asked an anchoring question,

such as, "Would you be willing to pay $5 …," before the point-blank question of how much they would contribute. When no anchor was mentioned, the visitors at the Exploratorium—generally an environmentally sensitive crowd—said they were willing to pay $64, on average. When the anchoring amount was only $5, contributions averaged $20. When the anchor was a rather extravagant $400, the willingness to pay rose to an average of $143. The difference between the high-anchor and low-anchor groups was $123. The anchoring effect was above 30%, indicating that increasing the initial request by $100 brought a return of $30 in average willingness to pay. Similar or even larger anchoring effects have been obtained in numerous studies of estimates and of willingness to pay. For example, French residents of the heavily polluted Marseilles region were asked what increase in living costs they would accept if they could live in a less polluted region. The anchoring effect was over 50% in that study. Anchoring effects are easily observed in online trading, where the same item is often offered at different "buy now" prices. The "estimate" in fine-art auctions is also an anchor that influences the first bid. There are situations in which anchoring appears reasonable. After all, it is not surprising that people who are asked difficult questions clutch at straws, and the anchor is a plausible straw. If you know next to nothing about the trees of California and are asked whether a redwood can be taller than 1,200 feet, you might infer that this number is not too far from the truth. Somebody who knows the true height thought up that question, so the anchor may be a valuable hint. However, a key finding of anchoring research is that anchors that are obviously random can be just as effective as potentially informative anchors. When we used a wheel of fortune to anchor estimates of the proportion of African nations in the UN, the anchoring index was 44%, well within the range of effects observed with anchors that could plausibly be taken as hints. Anchoring effects of similar size have been observed in experiments in which the last few digits of the respondent's Social Security number was used as the anchor (e.g., for estimating the number of physicians in their city). The conclusion is clear: anchors do not have their effects because people believe they are informative. The power of random anchors has been demonstrated in some unsettling ways. German judges with an average of more than fifteen years of experience on the bench first read a description of a woman who had been caught shoplifting, then rolled a pair of dice that were loaded so every roll resulted in either a 3 or a 9. As soon as the dice came to a stop, the judges were asked whether they would sentence the woman to a term in prison greater or lesser, in months, than the number showing on the dice. Finally, the judges were instructed to specify the exact prison sentence they would give to the shoplifter. On average, those who had rolled a 9 said they would sentence her to 8 months; those who rolled a 3 said they would sentence her to 5 months; the anchoring effect was 50%. USES AND ABUSES OF ANCHORS By now you should be convinced that anchoring effects—sometimes due to priming, sometimes to insufficient adjustment—are everywhere. The psychological mechanisms that produce anchoring make us far more suggestible than most of us would want to be. And of course there are quite a few people who are willing and able to exploit our gullibility. Anchoring effects explain why, for example, arbitrary rationing is an effective marketing ploy. A few years ago, supermarket shoppers in Sioux City, Iowa, encountered a sales promotion for Campbell's soup at about 10% off the regular price. On some days, a sign on the shelf said LIMIT OF 12 PER PERSON. On other days, the sign said NO LIMIT PER PERSON. Shoppers purchased an average of 7 cans when the limit was in force, twice as many as they bought when the limit was removed. Anchoring is not the sole explanation. Rationing also implies that the goods are flying off the shelves, and shoppers should feel some urgency about stocking up. But we also know that the mention of 12 cans as a possible purchase would produce anchoring even if the number were produced by a roulette wheel. We see the same strategy at work in the negotiation over the price of a home, when the seller makes the first move by setting the list price. As in many other games, moving first is an advantage in single-issue negotiations—for example, when price is the only issue to be settled between a buyer and a seller. As you may have experienced when negotiating for the first time in a bazaar, the initial anchor has a powerful effect. My advice to students when I taught negotiations was that if you think

the other side has made an outrageous proposal, you should not come back with an equally outrageous counteroffer, creating a gap that will be difficult to bridge in further negotiations. Instead you should make a scene, storm out or threaten to do so, and make it clear—to yourself as well as to the other side—that you will not continue the negotiation with that number on the table. The psychologists Adam Galinsky and Thomas Mussweiler proposed more subtle ways to resist the anchoring effect in negotiations. They instructed negotiators to focus their attention and search their memory for arguments against the anchor. The instruction to activate System 2 was successful. For example, the anchoring effect is reduced or eliminated when the second mover focuses his attention on the minimal offer that the opponent would accept, or on the costs to the opponent of failing to reach an agreement. In general, a strategy of deliberately "thinking the opposite" may be a good defense against anchoring effects, because it negates the biased recruitment of thoughts that produces these effects. Finally, try your hand at working out the effect of anchoring on a problem of public policy: the size of damages in personal injury cases. These awards are sometimes very large. Businesses that are frequent targets of such lawsuits, such as hospitals and chemical companies, have lobbied to set a cap on the awards. Before you read this chapter you might have thought that capping awards is certainly good for potential defendants, but now you should not be so sure. Consider the effect of capping awards at $1 million. This rule would eliminate all larger awards, but the anchor would also pull up the size of many awards that would otherwise be much smaller. It would almost certainly benefit serious offenders and large firms much more than small ones. ANCHORING AND THE TWO SYSTEMS The effects of random anchors have much to tell us about the relationship between System 1 and System 2. Anchoring effects have always been studied in tasks of judgment and choice that are ultimately completed by System 2. However, System 2 works on data that is retrieved from memory, in an automatic and involuntary operation of System 1. System 2 is therefore susceptible to the biasing influence of anchors that make some information easier to retrieve. Furthermore, System 2 has no control over the effect and no knowledge of it. The participants who have been exposed to random or absurd anchors (such as Gandhi's death at age 144) confidently deny that this obviously useless information could have influenced their estimate, and they are wrong. We saw in the discussion of the law of small numbers that a message, unless it is immediately rejected as a lie, will have the same effect on the associative system regardless of its reliability. The gist of the message is the story, which is based on whatever information is available, even if the quantity of the information is slight and its quality is poor: WYSIATI. When you read a story about the heroic rescue of a wounded mountain climber, its effect on your associative memory is much the same if it is a news report or the synopsis of a film. Anchoring results from this associative activation. Whether the story is true, or believable, matters little, if at all. The powerful effect of random anchors is an extreme case of this phenomenon, because a random anchor obviously provides no information at all. Earlier I discussed the bewildering variety of priming effects, in which your thoughts and behavior may be influenced by stimuli to which you pay no attention at all, and even by stimuli of which you are completely unaware. The main moral of priming research is that our thoughts and our behavior are influenced, much more than we know or want, by the environment of the moment. Many people find the priming results unbelievable, because they do not correspond to subjective experience. Many others find the results upsetting, because they threaten the subjective sense of agency and autonomy. If the content of a screen saver on an irrelevant computer can affect your willingness to help strangers without your being aware of it, how free are you? Anchoring effects are threatening in a similar way. You are always aware of the anchor and even pay attention to it, but you do not know how it guides and constrains your thinking, because you cannot imagine how you would have thought if the anchor had been different (or absent). However, you should assume that any number that is on the table has had an anchoring effect on you, and if the stakes are high you should mobilize yourself (your System 2) to combat the effect. SPEAKING OF ANCHORS "The firm we want to acquire sent us their business plan, with the revenue they expect. We shouldn't let that number

influence our thinking. Set it aside." "Plans are best-case scenarios. Let's avoid anchoring on plans when we forecast actual outcomes. Thinking about ways the plan could go wrong is one way to do it." "Our aim in the negotiation is to get them anchored on this number." "Let's make it clear that if that is their proposal, the negotiations are over. We do not want to start there." "The defendant's lawyers put in a frivolous reference in which they mentioned a ridiculously low amount of damages, and they got the judge anchored on it!"

12 The Science of Availability Amos and I had our most productive year in 1971–72, which we spent in Eugene, Oregon. We were the guests of the Oregon Research Institute, which housed several future stars of all the fields in which we worked—judgment, decision making, and intuitive prediction. Our main host was Paul Slovic, who had been Amos's classmate at Ann Arbor and remained a lifelong friend. Paul was on his way to becoming the leading psychologist among scholars of risk, a position he has held for decades, collecting many honors along the way. Paul and his wife, Roz, introduced us to life in Eugene, and soon we were doing what people in Eugene do—jogging, barbecuing, and taking children to basketball games. We also worked very hard, running dozens of experiments and writing our articles on judgment heuristics. At night I wrote Attention and Effort. It was a busy year. One of our projects was the study of what we called

the availability heuristic. We thought of that heuristic when we asked ourselves what people actually do when they wish to estimate the frequency of a category, such as “people who divorce after the age of 60” or “dangerous plants.” The answer was straightforward: instances of the class will be retrieved from memory, and if retrieval is easy and fluent, the category will be judged to be large. We defined the availability heuristic as the process of judging frequency by “the ease with which instances come to mind.” The statement seemed clear when we formulated it, but the concept of availability has been refined since then. The two-system approach had not yet been developed when we studied availability, and we did not attempt to determine whether this heuristic is a deliberate problem-solving strategy or an automatic operation. We now know that both systems are involved. A question we considered early was how many instances must be retrieved to get an impression of the ease with which they come to mind. We now know the answer: none. For an example, think of the number of words that can be constructed from the two sets of letters below. XUZONLCJM TAPCERHOB You knew almost immediately, without generating any instances, that one set offers far more possibilities than the other, probably by a factor of 10 or more. Similarly, you do not need to retrieve specific news stories to have a good idea of the relative frequency with which different countries have appeared in the news during the past year (Belgium, China, France, Congo, Nicaragua, Romania …). The availability heuristic, like other heuristics of judgment, substitutes one question for another: you wish to estimate the size of a category or the frequency of an event, but you report an impression of the ease with which instances come to mind. Substitution of questions inevitably produces systematic errors. You can discover how the heuristic leads to biases by following a simple procedure: list factors other than frequency that make it easy to come up with instances. Each factor in your list will be a potential source of bias. Here are some examples: A salient event that attracts your attention will be easily retrieved from memory. Divorces among Hollywood celebrities and sex scandals among politicians attract much attention, and instances will come easily to mind. You are therefore likely to exaggerate the frequency of both Hollywood divorces and political sex scandals. A dramatic event temporarily increases the availability of its category. A plane crash that attracts media coverage will temporarily alter your feelings about the safety of flying. Accidents are on your mind, for a while, after you see a car burning at the side of the road, and the world is for a while a more dangerous place. Personal experiences, pictures, and vivid examples are more available than incidents that happened to others, or mere words, or statistics. A judicial error that affects you will undermine your faith in the justice system more than a similar incident you read about in a newspaper.

Resisting this large collection of potential availability biases is possible, but tiresome. You must make the effort to reconsider your impressions and intuitions by asking such questions as, “Is our belief that thefts by teenagers are a major problem due to a few recent instances in our neighborhood?” or “Could it be that I feel no need to get a flu shot because none of my acquaintances got the flu last year?” Maintaining one’s vigilance against biases is a chore—but the chance to avoid a costly mistake is sometimes worth the effort. One of the best-known studies of availability suggests that awareness of your own biases can contribute to peace in marriages, and probably in other joint projects. In a famous study, spouses were asked, “How large was your personal contribution to keeping the place tidy, in percentages?” They also answered similar questions about “taking out the garbage,” “initiating social engagements,” etc. Would the self-estimated contributions add up to 100%, or more, or less? As expected, the self-assessed contributions added up to more than 100%. The explanation is a simple availability bias: both spouses remember their own individual efforts and contributions much more clearly than those of the other, and the difference in availability leads to a difference in judged frequency. The bias is not necessarily self-serving: spouses also overestimated their contribution to causing quarrels, although to a smaller extent than their contributions to more desirable outcomes. The same bias contributes to the common observation that many members of a collaborative team feel they have done more than their share and also feel that the others are not adequately grateful for their individual contributions. I am generally not optimistic about the potential for personal control of biases, but this is an exception. The opportunity for successful debiasing exists because the circumstances in which issues of credit allocation come up are easy to identify, the more so because tensions often arise when several people at once feel that their efforts are not adequately recognized. The mere observation that there is usually more than 100% credit to go around is sometimes sufficient to defuse the situation. In any event, it is a good thing for every individual to remember. You will occasionally do more than your share, but it is useful to know that you are likely to have that feeling even when each member of the team feels the same way. THE PSYCHOLOGY OF AVAILABILITY A major advance in the understanding of the availability heuristic occurred in the early 1990s, when a group of German psychologists led by Norbert Schwarz raised an intriguing question: How will people's impressions of the frequency of a category be affected by a requirement to list a specified number of instances? Imagine yourself a subject in that experiment: First, list six instances in which you behaved assertively. Next, evaluate how assertive you are. Imagine that you had been asked for twelve instances of assertive behavior (a number most people find difficult). Would your view of your own assertiveness be different? Schwarz and his colleagues observed that the task of listing instances may enhance the judgments of the trait by two different routes: the number of instances retrieved the ease with which they come to mind The request to list twelve instances pits the two determinants against each other. On the one hand, you have just retrieved an impressive number of cases in which you were assertive. On the other hand, while the first three or four instances of your own assertiveness probably came easily to you, you almost certainly struggled to come up with the last few to complete a set of twelve; fluency was low. Which will count more—the amount retrieved or the ease and fluency of the retrieval? The contest yielded a clear-cut winner: people who had just listed twelve instances rated themselves as less assertive than people who had listed only six. Furthermore, participants who had been asked to list twelve cases in which they had not behaved assertively ended up thinking of themselves as quite assertive! If you cannot easily come up with instances of meek behavior, you are likely to conclude that you are not meek at all. Self-ratings were dominated by the ease with which examples had come to mind. The experience of fluent retrieval of instances trumped the number retrieved. An even more direct demonstration of the role of fluency was offered by other psychologists in the same group. All the participants in their experiment listed six instances of assertive (or nonassertive) behavior, while maintaining a specified facial expression. "Smilers" were instructed to contract the zygomaticus muscle, which produces

a light smile; "frowners" were required to furrow their brow. As you already know, frowning normally accompanies cognitive strain and the effect is symmetric: when people are instructed to frown while doing a task, they actually try harder and experience greater cognitive strain. The researchers anticipated that the frowners would have more difficulty retrieving examples of assertive behavior and would therefore rate themselves as relatively lacking in assertiveness. And so it was. Psychologists enjoy experiments that yield paradoxical results, and they have applied Schwarz's discovery with gusto. For example, people: believe that they use their bicycles less often after recalling many rather than few instances are less confident in a choice when they are asked to produce more arguments to support it are less confident that an event was avoidable after listing more ways it could have been avoided are less impressed by a car after listing many of its advantages A professor at UCLA found an ingenious way to exploit the availability bias. He asked different groups of students to list ways to improve the course, and he varied the required number of improvements. As expected, the students who listed more ways to improve the class rated it higher! Perhaps the most interesting finding of this paradoxical research is that the paradox is not always found: people sometimes go by content rather than by ease of retrieval. The proof that you truly understand a pattern of behavior is that you know how to reverse it. Schwarz and his colleagues took on this challenge of discovering the conditions under which this reversal would take place. The ease with which instances of assertiveness come to the subject's mind changes during the task. The first few instances are easy, but retrieval soon becomes much harder. Of course, the subject also expects fluency to drop gradually, but the drop of fluency between six and twelve instances appears to be steeper than the participant expected. The results suggest that the participants make an inference: if I am having so much more trouble than expected coming up with instances of my assertiveness, then I can't be very assertive. Note that this inference rests on a surprise—fluency being worse than expected. The availability heuristic that the subjects apply is better described as an "unexplained unavailability" heuristic. Schwarz and his colleagues reasoned that they could disrupt the heuristic by providing the subjects with an explanation for the fluency of retrieval that they experienced. They told the participants they would hear background music while recalling instances and that the music would affect performance in the memory task. Some subjects were told that the music would help, others were told to expect diminished fluency. As predicted, participants whose experience of fluency was "explained" did not use it as a heuristic; the subjects who were told that music would make retrieval more difficult rated themselves as equally assertive when they retrieved twelve instances as when they retrieved six. Other cover stories have been used with the same result: judgments are no longer influenced by ease of retrieval when the experience of fluency is given a spurious explanation by the presence of curved or straight text boxes, by the background color of the screen, or by other irrelevant factors that the experimenters dreamed up. As I have described it, the process that leads to judgment by availability appears to involve a complex chain of reasoning. The subjects have an experience of diminishing fluency as they produce instances. They evidently have expectations about the rate at which fluency decreases, and those expectations are wrong: the difficulty of coming up with new instances increases more rapidly than they expect. It is the unexpectedly low fluency that causes people who were asked for twelve instances to describe themselves as unassertive. When the surprise is eliminated, low fluency no longer influences the judgment. The process appears to consist of a sophisticated set of inferences. Is the automatic System 1 capable of it? The answer is that in fact no complex reasoning is needed. Among the basic features of System 1 is its ability to set expectations and to be surprised when these expectations are violated. The system also retrieves possible causes of a surprise, usually by finding a possible cause among recent surprises. Furthermore, System 2 can reset the expectations of System 1 on the fly, so that an event that would normally be surprising is now almost normal. Suppose you are told that the three-year-old boy who lives next door frequently wears a top hat in his stroller. You will be far less surprised when you actually see him with his top hat than you would

have been without the warning. In Schwarz's experiment, the background music has been mentioned as a possible cause of retrieval problems. The difficulty of retrieving twelve instances is no longer a surprise and therefore is less likely to be evoked by the task of judging assertiveness. Schwarz and his colleagues discovered that people who are personally involved in the judgment are more likely to consider the number of instances they retrieve from memory and less likely to go by fluency. They recruited two groups of students for a study of risks to cardiac health. Half the students had a family history of cardiac disease and were expected to take the task more seriously than the others, who had no such history. All were asked to recall either three or eight behaviors in their routine that could affect their cardiac health (some were asked for risky behaviors, others for protective behaviors). Students with no family history of heart disease were casual about the task and followed the availability heuristic. Students who found it difficult to find eight instances of risky behavior felt themselves relatively safe, and those who struggled to retrieve examples of safe behaviors felt themselves at risk. The students with a family history of heart disease showed the opposite pattern—they felt safer when they retrieved many instances of safe behavior and felt greater danger when they retrieved many instances of risky behavior. They were also more likely to feel that their future behavior would be affected by the experience of evaluating their risk. The conclusion is that the ease with which instances come to mind is a System 1 heuristic, which is replaced by a focus on content when System 2 is more engaged. Multiple lines of evidence converge on the conclusion that people who let themselves be guided by System 1 are more strongly susceptible to availability biases than others who are in a state of higher vigilance. The following are some conditions in which people "go with the flow" and are affected more strongly by ease of retrieval than by the content they retrieved: when they are engaged in another effortful task at the same time when they are in a good mood because they just thought of a happy episode in their life if they score low on a depression scale if they are knowledgeable novices on the topic of the task, in contrast to true experts when they score high on a scale of faith in intuition if they are (or are made to feel) powerful I find the last finding particularly intriguing. The authors introduce their article with a famous quote: "I don't spend a lot of time taking polls around the world to tell me what I think is the right way to act. I've just got to know how I feel" (George W. Bush, November 2002). They go on to show that reliance on intuition is only in part a personality trait. Merely reminding people of a time when they had power increases their apparent trust in their own intuition. SPEAKING OF AVAILABILITY "Because of the coincidence of two planes crashing last month, she now prefers to take the train. That's silly. The risk hasn't really changed; it is an availability bias." "He underestimates the risks of indoor pollution because there are few media stories on them. That's an availability effect. He should look at the statistics." "She has been watching too many spy movies recently, so she's seeing conspiracies everywhere." "The CEO has had several successes in a row, so failure doesn't come easily to her mind. The availability bias is making her overconfident."

[reaching fair use limit so skip 2.14]

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.

[skip]

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.

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

[skip]

18 Taming Intuitive Predictions Life presents us with many occasions to forecast. Economists forecast inflation and unemployment, financial analysts forecast earnings, military experts predict casualties, venture capitalists assess profitability, publishers and producers predict audiences, contractors estimate the time required to complete projects, chefs anticipate the demand for the dishes on their menu, engineers estimate the amount of concrete needed for a building, fireground commanders assess the number of trucks that will be needed to put out a fire. In our private lives, we forecast our spouse’s reaction to a proposed move or our own future adjustment to a new job. Some predictive judgments, such as those made by engineers, rely largely on look-up tables, precise calculations, and explicit analyses of outcomes observed on similar occasions. Others involve intuition and System 1, in two main varieties. Some intuitions draw primarily on skill and expertise acquired by repeated experience. The rapid and automatic judgments and choices of chess masters, fireground commanders, and physicians that Gary Klein has described in Sources of Power and elsewhere illustrate these skilled intuitions, in which a solution to the current problem comes to mind quickly because familiar cues are recognized. Other intuitions, which are sometimes subjectively indistinguishable from the first, arise from the operation of heuristics that often substitute an easy question for the harder one that was asked. Intuitive judgments can be made with high confidence even when they are based on nonregressive assessments of weak evidence. Of course, many judgments, especially in the professional domain, are influenced by

Kahneman, Daniel. Thinking, Fast and Slow (pp. 184-186). Penguin Books Ltd. Kindle Edition.