PART I

Finding Noise

It is not acceptable for similar people, convicted of the same offense, to end up with dramatically different sentences—say, five years in jail for one and probation for another. And yet in many places, something like that happens. To be sure, the criminal justice system is pervaded by bias as well. But our focus in chapter 1 is on noise—and in particular, on what happened when a famous judge drew attention to it, found it scandalous, and launched a crusade that in a sense changed the world (but not enough). Our tale involves the United States, but we are confident that similar stories can be (and will be) told about many other nations. In some of those nations, the problem of noise is likely to be even worse than it is in the United States. We use the example of sentencing in part to show that noise can produce great unfairness.

Criminal sentencing has especially high drama, but we are also concerned with the private sector, where the stakes can be large, too. To illustrate the point, we turn in chapter 2 to a large insurance company. There, underwriters have the task of setting insurance premiums for potential clients, and claims adjusters must judge the value of claims. You might predict that these tasks would be simple and mechanical and that different professionals would come up with roughly the same amounts. We conducted a carefully designed experiment—a noise audit—to test that prediction. The results surprised us, but more importantly they astonished and dismayed the company's leadership. As we learned, the sheer volume of noise is costing the company a great deal of money. We use this example to show that noise can produce large economic losses.

Both of these examples involve studies of a large number of people

making a large number of judgments. But many important judgments are *singular* rather than repeated: how to handle an apparently unique business opportunity, whether to launch a whole new product, how to deal with a pandemic, whether to hire someone who just doesn't meet the standard profile. Can noise be found in decisions about unique situations like these? It is tempting to think that it is absent there. After all, noise is unwanted variability, and how can you have variability with singular decisions? In chapter 3 , we try to answer this question. The judgment that you make, even in a seemingly unique situation, is one in a cloud of possibilities. You will find a lot of noise there as well.

The theme that emerges from these three chapters can be summarized in one sentence, which will be a key theme of this book: *wherever there is judgment, there is noise—and more of it than you think.* Let's start to find out how much.

CHAPTER 1

Crime and Noisy Punishment

Suppose that someone has been convicted of a crime—shoplifting, possession of heroin, assault, or armed robbery. What is the sentence likely to be?

The answer should not depend on the particular judge to whom the case happens to be assigned, on whether it is hot or cold outside, or on whether a local sports team won the day before. It would be outrageous if three similar people, convicted of the same crime, received radically different penalties: probation for one, two years in jail for another, and ten years in jail for another. And yet that outrage can be found in many nations—not only in the distant past but also today.

All over the world, judges have long had a great deal of discretion in deciding on appropriate sentences. In many nations, experts have celebrated this discretion and have seen it as both just and humane. They have insisted that criminal sentences should be based on a host of factors involving not only the crime but also the defendant's character and circumstances. Individualized tailoring was the order of the day. If judges were constrained by rules, criminals would be treated in a dehumanized way; they would not be seen as unique individuals entitled to draw attention to the details of their situation. The very idea of due process of law seemed, to many, to call for open-ended judicial discretion.

In the 1970s, the universal enthusiasm for judicial discretion started to collapse for one simple reason: startling evidence of noise. In 1973, a famous judge, Marvin Frankel, drew public attention to the problem. Before he became a judge, Frankel was a defender of freedom of speech and a passionate human rights advocate who helped found the Lawyers' Committee

for Human Rights (an organization now known as Human Rights First).

Frankel could be fierce. And with respect to noise in the criminal justice system, he was outraged. Here is how he describes his motivation:

> *If a federal bank robbery defendant was convicted, he or she could receive a maximum of 25 years. That meant anything from 0 to 25 years. And where the number was set, I soon realized, depended less on the case or the individual defendant than on the individual judge, i.e., on the views, predilections, and biases of the judge. So the same defendant in the same case could get widely different sentences depending on which judge got the case.*

Frankel did not provide any kind of statistical analysis to support his argument. But he did offer a series of powerful anecdotes, showing unjustified disparities in the treatment of similar people. Two men, neither of whom had a criminal record, were convicted for cashing counterfeit checks in the amounts of $58.40 and $35.20, respectively. The first man was sentenced to fifteen *years*, the second to 30 *days*. For embezzlement actions that were similar to one another, one man was sentenced to 117 *days* in prison, while another was sentenced to 20 *years* . Pointing to numerous cases of this kind, Frankel deplored what he called the “almost wholly unchecked and sweeping powers” of federal judges, resulting in “arbitrary cruelties perpetrated daily,” which he deemed unacceptable in a “government of laws, not of men.”

Frankel called on Congress to end this “discrimination,” as he described those arbitrary cruelties. By that term, he mainly meant noise, in the form of inexplicable variations in sentencing. But he was also concerned about bias, in the form of racial and socioeconomic disparities. To combat both noise and bias, he urged that differences in treatment of criminal defendants should not be allowed unless the differences could be “justified by relevant tests capable of formulation and application with sufficient objectivity to ensure that the results will be more than the idiosyncratic ukases of particular officials, justices, or others.” (The term *idiosyncratic ukases* is a bit esoteric; by it, Frankel meant personal edicts.) Much more than that, Frankel argued for a reduction in noise through a “detailed profile or checklist of factors that would include, wherever possible, some form of numerical or other objective grading.”

Writing in the early 1970s, he did not go quite so far as to defend what he called “displacement of people by machines.” But startlingly, he came close. He believed that “the rule of law calls for a body of impersonal rules, applicable across the board, binding on judges as well as everyone else.” He explicitly argued for the use of “computers as an aid toward orderly thought in sentencing.” He also recommended the creation of a commission on sentencing.

Frankel’s book became one of the most influential in the entire history of criminal law—not only in the United States but also throughout the world. His work did suffer from a degree of informality. It was devastating but impressionistic. To test for the reality of noise, several people immediately followed up by exploring the level of noise in criminal sentencing.

An early large-scale study of this kind, chaired by Judge Frankel himself, took place in 1974. Fifty judges from various districts were asked to set sentences for defendants in hypothetical cases summarized in identical pre-sentence reports. The basic finding was that “absence of consensus was the norm” and that the variations across punishments were “astounding.” A heroin dealer could be incarcerated for one to ten years, depending on the judge. Punishments for a bank robber ranged from five to eighteen years in prison. The study found that in an extortion case, sentences varied from a whopping twenty years imprisonment and a $65,000 fine to a mere three years imprisonment and no fine. Most startling of all, in sixteen of twenty cases, there was no unanimity on whether any incarceration was appropriate.

This study was followed by a series of others, all of which found similarly shocking levels of noise. In 1977, for example, William Austin and Thomas Williams conducted a survey of forty-seven judges, asking them to respond to the same five cases, each involving low-level offenses. All the descriptions of the cases included summaries of the information used by judges in actual sentencing, such as the charge, the testimony, the previous criminal record (if any), social background, and evidence relating to character. The key finding was “substantial disparity.” In a case involving burglary, for example, the recommended sentences ranged from five years in prison to a mere thirty days (alongside a fine of $100). In a case involving possession of marijuana, some judges recommended prison terms; others recommended probation.

A much larger study, conducted in 1981, involved 208 federal judges who were exposed to the same sixteen hypothetical cases. Its central findings

were stunning:

> *In only 3 of the 16 cases was there a unanimous agreement to impose a prison term. Even where most judges agreed that a prison term was appropriate, there was a substantial variation in the lengths of prison terms recommended. In one fraud case in which the mean prison term was 8.5 years, the longest term was life in prison. In another case the mean prison term was 1.1 years, yet the longest prison term recommended was 15 years.*

As revealing as they are, these studies, which involve tightly controlled experiments, almost certainly understate the magnitude of noise in the real world of criminal justice. Real-life judges are exposed to far more information than what the study participants received in the carefully specified vignettes of these experiments. Some of this additional information is relevant, of course, but there is also ample evidence that irrelevant information, in the form of small and seemingly random factors, can produce major differences in outcomes. For example, judges have been found more likely to grant parole at the beginning of the day or after a food break than immediately before such a break. If judges are hungry, they are tougher.

A study of thousands of juvenile court decisions found that when the local football team loses a game on the weekend, the judges make harsher decisions on the Monday (and, to a lesser extent, for the rest of the week). Black defendants disproportionately bear the brunt of that increased harshness. A different study looked at 1.5 million judicial decisions over three decades and similarly found that judges are more severe on days that follow a loss by the local city's football team than they are on days that follow a win.

A study of six million decisions made by judges in France over twelve years found that defendants are given more leniency on their birthday. (The defendant's birthday, that is; we suspect that judges might be more lenient on their own birthdays as well, but as far as we know, that hypothesis has not been tested.) Even something as irrelevant as outside temperature can influence judges. A review of 207,000 immigration court decisions over four years found a significant effect of daily temperature variations: when it is hot outside, people are less likely to get asylum. If you are suffering political

persecution in your home country and want asylum elsewhere, you should hope and maybe even pray that your hearing falls on a cool day.

Reducing Noise in Sentencing

In the 1970s, Frankel's arguments, and the empirical findings supporting them, came to the attention of Edward M. Kennedy, brother of the slain president John F. Kennedy, and one of the most influential members of the US Senate. Kennedy was shocked and appalled. As early as 1975, he introduced sentencing reform legislation; it didn't go anywhere. But Kennedy was relentless. Pointing to the evidence, he continued to press for the enactment of that legislation, year after year. In 1984, he succeeded. Responding to the evidence of unjustified variability, Congress enacted the Sentencing Reform Act of 1984.

The new law was intended to reduce noise in the system by reducing "the unfettered discretion the law confers on those judges and parole authorities responsible for imposing and implementing the sentences." In particular, members of Congress referred to "unjustifiably wide" sentencing disparity, specifically citing findings that in the New York area, punishments for identical actual cases could range from three years to twenty years of imprisonment. Just as Judge Frankel had recommended, the law created the US Sentencing Commission, whose principal job was clear: to issue sentencing guidelines that were meant to be mandatory and that would establish a restricted range for criminal sentences.

In the following year, the commission established those guidelines, which were generally based on average sentences for similar crimes in an analysis of ten thousand actual cases. Supreme Court Justice Stephen Breyer, who was heavily involved in the process, defended the use of past practice by pointing to the intractable disagreement within the commission: "Why didn't the Commission sit down and really go and rationalize this thing and not just take history? The short answer to that is: we couldn't. We couldn't because there are such good arguments all over the place pointing in opposite directions.… Try listing all the crimes that there are in rank order of punishable merit.… Then collect results from your friends and see if they all match. I will tell you they won't."

Under the guidelines, judges have to consider two factors to establish

sentences: the crime and the defendant's criminal history. Crimes are assigned one of forty-three "offense levels," depending on their seriousness. The defendant's criminal history refers principally to the number and severity of a defendant's previous convictions. Once the crime and the criminal history are put together, the guidelines offer a relatively narrow range of sentencing, with the top of the range authorized to exceed the bottom by the greater of six months or 25%. Judges are permitted to depart from the range altogether by reference to what they see as aggravating or mitigating circumstances, but departures must be justified to an appellate court.

Even though the guidelines are mandatory, they are not entirely rigid. They do not go nearly as far as Judge Frankel wanted. They offer judges significant room to maneuver. Nonetheless, several studies, using a variety of methods and focused on a range of historical periods, reach the same conclusion: the guidelines cut the noise. More technically, they "reduced the net variation in sentence attributable to the happenstance of the identity of the sentencing judge."

The most elaborate study came from the commission itself. It compared sentences in bank robbery, cocaine distribution, heroin distribution, and bank embezzlement cases in 1985 (before the guidelines went into effect) with the sentences imposed between January 19, 1989, and September 30, 1990. Offenders were matched with respect to the factors deemed relevant to sentencing under the guidelines. For every offense, variations across judges were much smaller in the later period, after the Sentencing Reform Act had been implemented.

According to another study, the expected difference in sentence length between judges was 17%, or 4.9 months, in 1986 and 1987. That number fell to 11%, or 3.9 months, between 1988 and 1993. An independent study covering different periods found similar success in reducing interjudge disparities, which were defined as the differences in average sentences among judges with similar caseloads.

Despite these findings, the guidelines ran into a firestorm of criticism. Some people, including many judges, thought that some sentences were too severe—a point about bias, not noise. For our purposes, a much more interesting objection, which came from numerous judges, was that guidelines were deeply unfair because they prohibited judges from taking adequate account of the particulars of the case. The price of reducing noise was to make decisions unacceptably mechanical. Yale law professor Kate Stith and

federal judge José Cabranes wrote that “the need is not for blindness, but for insight, for equity,” which “can only occur in a judgment that takes account of the complexities of the individual case.”

This objection led to vigorous challenges to the guidelines, some of them based on law, others based on policy. Those challenges failed until, for technical reasons entirely unrelated to the debate summarized here, the Supreme Court struck the guidelines down in 2005. As a result of the court’s ruling, the guidelines became merely advisory. Notably, most federal judges were much happier after the Supreme Court decision. Seventy-five percent preferred the advisory regime, whereas just 3% thought the mandatory regime was better.

What have been the effects of changing the guidelines from mandatory to advisory? Harvard law professor Crystal Yang investigated this question, not with an experiment or a survey but with a massive data set of actual sentences, involving nearly four hundred thousand criminal defendants. Her central finding is that by multiple measures, interjudge disparities increased significantly after 2005. When the guidelines were mandatory, defendants who had been sentenced by a relatively harsh judge were sentenced to 2.8 months longer than if they had been sentenced by an average judge. When the guidelines became merely advisory, the disparity was doubled. Sounding much like Judge Frankel from forty years before, Yang writes that her “findings raise large equity concerns because the identity of the assigned sentencing judge contributes significantly to the disparate treatment of similar offenders convicted of similar crimes.”

After the guidelines became advisory, judges became more likely to base their sentencing decisions on their personal values. Mandatory guidelines reduce bias as well as noise. After the Supreme Court’s decision, there was a significant increase in the disparity between the sentences of African American defendants and white people convicted of the same crimes. At the same time, female judges became more likely than male judges were to exercise their increased discretion in favor of leniency. The same is true of judges appointed by Democratic presidents.

Three years after Frankel’s death in 2002, striking down the mandatory guidelines produced a return to something more like his nightmare: law without order.

The story of Judge Frankel's fight for sentencing guidelines offers a glimpse of several of the key points we will cover in this book. First, judgment is difficult because the world is a complicated, uncertain place. This complexity is obvious in the judiciary and holds in most other situations requiring professional judgment. Broadly, these situations include judgments made by doctors, nurses, lawyers, engineers, teachers, architects, Hollywood executives, members of hiring committees, book publishers, corporate executives of all kinds, and managers of sports teams. Disagreement is unavoidable wherever judgment is involved.

Second, the extent of these disagreements is much greater than we expect. While few people object to the principle of judicial discretion, almost everyone disapproves of the magnitude of the disparities it produces. *System noise*, that is, unwanted variability in judgments that should ideally be identical, can create rampant injustice, high economic costs, and errors of many kinds.

Third, noise can be reduced. The approach advocated by Frankel and implemented by the US Sentencing Commission—rules and guidelines—is one of several approaches that successfully reduce noise. Other approaches are better suited to other types of judgment. Some methods adopted to reduce noise can simultaneously reduce bias as well.

Fourth, efforts at noise reduction often raise objections and run into serious difficulties. These issues must be addressed, too, or the fight against noise will fail.

Speaking of Noise in Sentencing

"Experiments show large disparities among judges in the sentences they recommend for identical cases. This variability cannot be fair. A defendant's sentence should not depend on which judge the case happens to be assigned to."

"Criminal sentences should not depend on the judge's mood during the hearing, or on the outside temperature."

"Guidelines are one way to address this issue. But many people don't like them, because they limit judicial discretion, which might be necessary to ensure fairness and accuracy. After all, each case is unique, isn't it?"

CHAPTER 2

A Noisy System

Our initial encounter with noise, and what first triggered our interest in the topic, was not nearly so dramatic as a brush with the criminal justice system. Actually, the encounter was a kind of accident, involving an insurance company that had engaged the consulting firm with which two of us were affiliated.

Of course, the topic of insurance is not everyone's cup of tea. But our findings show the magnitude of the problem of noise in a for-profit organization that stands to lose a lot from noisy decisions. Our experience with the insurance company helps explain why the problem is so often unseen and what might be done about it.

The insurance company's executives were weighing the potential value of an effort to increase consistency—to reduce noise—in the judgments of people who made significant financial decisions on the firm's behalf. Everyone agreed that consistency is desirable. Everyone also agreed that these judgments could never be entirely consistent, because they are informal and partly subjective. Some noise is inevitable.

Disagreement emerged when it came to its magnitude. The executives doubted that noise could be a substantial problem for their company. Much to their credit, however, they agreed to settle the question by a kind of simple experiment that we will call a *noise audit* . The result surprised them. It also turned out to be a perfect illustration of the problem of noise.

A Lottery That Creates Noise

Many professionals in any large company are authorized to make judgments that bind the company. For example, this insurance company employs numerous underwriters who quote premiums for financial risks, such as insuring a bank against losses due to fraud or rogue trading. It also employs many claims adjusters who forecast the cost of future claims and also negotiate with claimants if disputes arise.

Every large branch of the company has several qualified underwriters. When a quote is requested, anyone who happens to be available may be assigned to prepare it. In effect, the particular underwriter who will determine a quote is selected by a lottery.

The exact value of the quote has significant consequences for the company. A high premium is advantageous if the quote is accepted, but such a premium risks losing the business to a competitor. A low premium is more likely to be accepted, but it is less advantageous to the company. For any risk, there is a Goldilocks price that is just right—neither too high nor too low—and there is a good chance that the average judgment of a large group of professionals is not too far from this Goldilocks number. Prices that are higher or lower than this number are costly—this is how the variability of noisy judgments hurts the bottom line.

The job of claims adjusters also affects the finances of the company. For example, suppose that a claim is submitted on behalf of a worker (the claimant) who permanently lost the use of his right hand in an industrial accident. An adjuster is assigned to the claim—just as the underwriter was assigned, because she happens to be available. The adjuster gathers the facts of the case and provides an estimate of its ultimate cost to the company. The same adjuster then takes charge of negotiating with the claimant's representative to ensure that the claimant receives the benefits promised in the policy while also protecting the company from making excessive payments.

The early estimate matters because it sets an implicit goal for the adjuster in future negotiations with the claimant. The insurance company is also legally obligated to reserve the predicted cost of each claim (i.e., to have enough cash to be able to pay it). Here again, there is a Goldilocks value from the perspective of the company. A settlement is not guaranteed, as there is an attorney for the claimant on the other side, who may choose to go to court if the offer is miserly. On the other hand, an overly generous reserve may allow the adjuster too much latitude to agree to frivolous demands. The adjuster's

judgment is consequential for the company—and even more consequential for the claimant.

We use the word *lottery* to emphasize the role of chance in the selection of one underwriter or adjuster. In the normal operation of the company, a single professional is assigned to a case, and no one can ever know what would have happened if another colleague had been selected instead.

Lotteries have their place, and they need not be unjust. Acceptable lotteries are used to allocate "goods," like courses in some universities, or "bads," like the draft in the military. They serve a purpose. But the judgment lotteries we talk about allocate nothing. They just produce uncertainty. Imagine an insurance company whose underwriters are noiseless and set the optimal premium, but a chance device then intervenes to modify the quote that the client actually sees. Evidently, there would be no justification for such a lottery. Neither is there any justification for a system in which the outcome depends on the identity of the person randomly chosen to make a professional judgment.

Noise Audits Reveal System Noise

The lottery that picks a particular judge to establish a criminal sentence or a single shooter to represent a team creates variability, but this variability remains unseen. A noise audit—like the one conducted on federal judges with respect to sentencing—is a way to reveal noise. In such an audit, the same case is evaluated by many individuals, and the variability of their responses is made visible.

The judgments of underwriters and claims adjusters lend themselves especially well to this exercise because their decisions are based on written information. To prepare for the noise audit, executives of the company constructed detailed descriptions of five representative cases for each group (underwriters and adjusters). Employees were asked to evaluate two or three cases each, working independently. They were not told that the purpose of the study was to examine the variability of their judgments.

Before reading on, you may want to think of your own answer to the following questions: In a well-run insurance company, if you randomly selected two qualified underwriters or claims adjusters, how different would you expect their estimates for the same case to be? Specifically, what would

be the difference between the two estimates, as a percentage of their average?

We asked numerous executives in the company for their answers, and in subsequent years, we have obtained estimates from a wide variety of people in different professions. Surprisingly, one answer is clearly more popular than all others. Most executives of the insurance company guessed 10% or less. When we asked 828 CEOs and senior executives from a variety of industries how much variation they expected to find in similar expert judgments, 10% was also the median answer and the most frequent one (the second most popular was 15%). A 10% difference would mean, for instance, that one of the two underwriters set a premium of $9,500 while the other quoted $10,500. Not a negligible difference, but one that an organization can be expected to tolerate.

Our noise audit found much greater differences. By our measure, the median difference in underwriting was 55%, about five times as large as was expected by most people, including the company's executives. This result means, for instance, that when one underwriter sets a premium at $9,500, the other does not set it at $10,500—but instead quotes $16,700. For claims adjusters, the median ratio was 43%. We stress that these results are medians: in half the pairs of cases, the difference between the two judgments was even larger.

The executives to whom we reported the results of the noise audit were quick to realize that the sheer volume of noise presented an expensive problem. One senior executive estimated that the company's annual cost of noise in underwriting—counting both the loss of business from excessive quotes and the losses incurred on underpriced contracts—was in the hundreds of millions of dollars.

No one could say precisely how much error (or how much bias) there was, because no one could know for sure the Goldilocks value for each case. But no one needed to see the bull's-eye to measure the scatter on the back of the target and to realize that the variability was a problem. The data showed that the price a customer is asked to pay depends to an uncomfortable extent on the lottery that picks the employee who will deal with that transaction. To say the least, customers would not be pleased to hear that they were signed up for such a lottery without their consent. More generally, people who deal with organizations expect a system that reliably delivers consistent judgments. They do not expect system noise.

Unwanted Variability Versus Wanted Diversity

A defining feature of system noise is that it is *unwanted,* and we should stress right here that variability in judgments is not always unwanted.

Consider matters of preference or taste. If ten film critics watch the same movie, if ten wine tasters rate the same wine, or if ten people read the same novel, we do not expect them to have the same opinion. Diversity of tastes is welcome and entirely expected. No one would want to live in a world in which everyone has exactly the same likes and dislikes. (Well, almost no one.) But diversity of tastes can help account for errors if a personal taste is mistaken for a professional judgment. If a film producer decides to go forward with an unusual project (about, say, the rise and fall of the rotary phone) because she personally likes the script, she might have made a major mistake if no one else likes it.

Variability in judgments is also expected and welcome in a competitive situation in which the best judgments will be rewarded. When several companies (or several teams in the same organization) compete to generate innovative solutions to the same customer problem, we don't want them to focus on the same approach. The same is true when multiple teams of researchers attack a scientific problem, such as the development of a vaccine: we very much want them to look at it from different angles. Even forecasters sometimes behave like competitive players. The analyst who correctly calls a recession that no one else has anticipated is sure to gain fame, whereas the one who never strays from the consensus remains obscure. In such settings, variability in ideas and judgments is again welcome, because variation is only the first step. In a second phase, the results of these judgments will be pitted against one another, and the best will triumph. In a market as in nature, selection cannot work without variation.

Matters of taste and competitive settings all pose interesting problems of judgment. But our focus is on judgments in which variability is undesirable. System noise is a problem of systems, which are organizations, not markets. When traders make different assessments of the value of a stock, some of them will make money, and others will not. Disagreements make markets. But if one of those traders is randomly chosen to make that assessment on behalf of her firm, and if we find out that her colleagues in the same firm would produce very different assessments, then the firm faces system noise, and that is a problem.

An elegant illustration of the issue arose when we presented our findings to the senior managers of an asset management firm, prompting them to run their own exploratory noise audit. They asked forty-two experienced investors in the firm to estimate the fair value of a stock (the price at which the investors would be indifferent to buying or selling). The investors based their analysis on a one-page description of the business; the data included simplified profit and loss, balance sheet, and cash flow statements for the past three years and projections for the next two. Median noise, measured in the same way as in the insurance company, was 41%. Such large differences among investors in the same firm, using the same valuation methods, cannot be good news.

Wherever the person making a judgment is randomly selected from a pool of equally qualified individuals, as is the case in this asset management firm, in the criminal justice system, and in the insurance company discussed earlier, noise is a problem. System noise plagues many organizations: an assignment process that is effectively random often decides which doctor sees you in a hospital, which judge hears your case in a courtroom, which patent examiner reviews your application, which customer service representative hears your complaint, and so on. Unwanted variability in these judgments can cause serious problems, including a loss of money and rampant unfairness.

A frequent misconception about unwanted variability in judgments is that it doesn’t matter, because random errors supposedly cancel one another out. Certainly, positive and negative errors in a judgment about the same case will tend to cancel one another out, and we will discuss in detail how this property can be used to reduce noise. But noisy systems do not make multiple judgments of the same case. They make noisy judgments of different cases. If one insurance policy is overpriced and another is underpriced, pricing may on average look right, but the insurance company has made two costly errors. If two felons who both should be sentenced to five years in prison receive sentences of three years and seven years, justice has not, on average, been done. In noisy systems, errors do not cancel out. They add up.

The Illusion of Agreement

A large literature going back several decades has documented noise in

professional judgment. Because we were aware of this literature, the results of the insurance company's noise audit did not surprise us. What did surprise us, however, was the reaction of the executives to whom we reported our findings: no one at the company had expected anything like the amount of noise we had observed. No one questioned the validity of the audit, and no one claimed that the observed amount of noise was acceptable. Yet the problem of noise—and its large cost—seemed like a new one for the organization. Noise was like a leak in the basement. It was tolerated not because it was thought acceptable but because it had remained unnoticed.

How could that be? How could professionals in the same role and in the same office differ so much from one another without becoming aware of it? How could executives fail to make this observation, which they understood to be a significant threat to the performance and reputation of their company? We came to see that the problem of system noise often goes unrecognized in organizations and that the common inattention to noise is as interesting as its prevalence. The noise audits suggested that respected professionals—and the organizations that employ them—maintained an *illusion of agreement* while in fact disagreeing in their daily professional judgments.

To begin to understand how the illusion of agreement arises, put yourself in the shoes of an underwriter on a normal working day. You have more than five years of experience, you know that you are well regarded among your colleagues, and you respect and like them. You know you are good at your job. After thoroughly analyzing the complex risks faced by a financial firm, you conclude that a premium of $200,000 is appropriate. The problem is complex but not much different from those you solve every day of the week.

Now imagine being told that your colleagues at the office have been given the same information and assessed the same risk. Could you believe that at least half of them have set a premium that is either higher than $255,000 or lower than $145,000? The thought is hard to accept. Indeed, we suspect that underwriters who heard about the noise audit and accepted its validity never truly believed that its conclusions applied to them personally.

Most of us, most of the time, live with the unquestioned belief that the world looks as it does because that's the way it is. There is one small step from this belief to another: "Other people view the world much the way I do." These beliefs, which have been called *naive realism*, are essential to the sense of a reality we share with other people. We rarely question these beliefs. We hold a single interpretation of the world around us at any one

time, and we normally invest little effort in generating plausible alternatives to it. One interpretation is enough, and we experience it as true. We do not go through life imagining alternative ways of seeing what we see.

In the case of professional judgments, the belief that others see the world much as we do is reinforced every day in multiple ways. First, we share with our colleagues a common language and set of rules about the considerations that should matter in our decisions. We also have the reassuring experience of agreeing with others on the absurdity of judgments that violate these rules. We view the occasional disagreements with colleagues as lapses of judgment on their part. We have little opportunity to notice that our agreed-on rules are vague, sufficient to eliminate some possibilities but not to specify a shared positive response to a particular case. We can live comfortably with colleagues without ever noticing that they actually do not see the world as we do.

One underwriter we interviewed described her experience in becoming a veteran in her department: “When I was new, I would discuss seventy-five percent of cases with my supervisor.… After a few years, I didn’t need to—I am now regarded as an expert.… Over time, I became more and more confident in my judgment.” Like many of us, this person had developed confidence in her judgment mainly by exercising it.

The psychology of this process is well understood. Confidence is nurtured by the subjective experience of judgments that are made with increasing fluency and ease, in part because they resemble judgments made in similar cases in the past. Over time, as this underwriter learned to agree with her past self, her confidence in her judgments increased. She gave no indication that—after the initial apprenticeship phase—she had learned to agree with others, had checked to what extent she did agree with them, or had even tried to prevent her practices from drifting away from those of her colleagues.

For the insurance company, the illusion of agreement was shattered only by the noise audit. How had the leaders of the company remained unaware of their noise problem? There are several possible answers here, but one that seems to play a large role in many settings is simply the discomfort of disagreement. Most organizations prefer consensus and harmony over dissent and conflict. The procedures in place often seem expressly designed to minimize the frequency of exposure to actual disagreements and, when such disagreements happen, to explain them away.

Nathan Kuncel, a professor of psychology at the University of Minnesota and a leading researcher on the prediction of performance, shared with us a story that illustrates this problem. Kuncel was helping a school's admissions office review its decision process. First a person read an application file, rated it, and then handed it off with ratings to a second reader, who then also rated it. Kuncel suggested—for reasons that will become obvious throughout this book—that it would be preferable to mask the first reader's ratings so as not to influence the second reader. The school's reply: "We used to do that, but it resulted in so many disagreements that we switched to the current system." This school is not the only organization that considers conflict avoidance at least as important as making the right decision.

Consider another mechanism that many companies resort to: postmortems of unfortunate judgments. As a learning mechanism, postmortems are useful. But if a mistake has truly been made—in the sense that a judgment strayed far from professional norms—discussing it will not be challenging. Experts will easily conclude that the judgment was way off the consensus. (They might also write it off as a rare exception.) Bad judgment is much easier to identify than good judgment. The calling out of egregious mistakes and the marginalization of bad colleagues will not help professionals become aware of how much they disagree when making broadly acceptable judgments. On the contrary, the easy consensus about bad judgments may even reinforce the illusion of agreement. The true lesson, about the ubiquity of system noise, will never be learned.

We hope you are starting to share our view that system noise is a serious problem. Its existence is not a surprise; noise is a consequence of the informal nature of judgment. However, as we will see throughout this book, the amount of noise observed when an organization takes a serious look almost always comes as a shock. Our conclusion is simple: wherever there is judgment, there is noise, and more of it than you think.

Speaking of System Noise in the Insurance Company

"We depend on the quality of professional judgments, by underwriters, claims adjusters, and others. We assign each case to one expert, but we operate under the wrong assumption that another expert would produce a similar judgment."

"System noise is five times larger than we thought—or than we can tolerate.

Without a noise audit, we would never have realized that. The noise audit shattered the illusion of agreement."

"System noise is a serious problem: it costs us hundreds of millions."

"Wherever there is judgment, there is noise—and more of it than we think."

CHAPTER 3

Singular Decisions

The case studies we have discussed thus far involve judgments that are made repeatedly. What is the right sentence for someone convicted of theft? What is the right premium for a particular risk? While each case is in some sense unique, judgments like these are *recurrent decisions* . Doctors diagnosing patients, judges hearing parole cases, admissions officers reviewing applications, accountants preparing tax forms—these are all examples of recurrent decisions.

Noise in recurrent decisions is demonstrated by a noise audit, such as those we introduced in the previous chapter. Unwanted variability is easy to define and measure when interchangeable professionals make decisions in similar cases. It seems much harder, or perhaps even impossible, to apply the idea of noise to a category of judgments that we call *singular decisions* .

Consider, for instance, the crisis the world faced in 2014. In West Africa, numerous people were dying from Ebola. Because the world is interconnected, projections suggested that infections would rapidly spread all over the world and hit Europe and North America particularly hard. In the United States, there were insistent calls to shut down air travel from affected regions and to take aggressive steps to close the borders. The political pressure to move in that direction was intense, and prominent and well-informed people favored those steps.

President Barack Obama was faced with one of the most difficult decisions of his presidency—one that he had not encountered before and never encountered again. He chose not to close any borders. Instead he sent three thousand people—health workers and soldiers—to West Africa. He led a diverse, international coalition of nations that did not always work well

together, using their resources and expertise to tackle the problem at its source.

Singular Versus Recurrent

Decisions that are made only once, like the president's Ebola response, are singular because they are not made recurrently by the same individual or team, they lack a prepackaged response, and they are marked by genuinely unique features. In dealing with Ebola, President Obama and his team had no real precedents on which to draw. Important political decisions are often good examples of singular decisions, as are the most fateful choices of military commanders.

In the private realm, decisions you make when choosing a job, buying a house, or proposing marriage have the same characteristics. Even if this is not your first job, house, or marriage, and despite the fact that countless people have faced these decisions before, the decision feels unique to you. In business, heads of companies are often called on to make what seem like unique decisions to them: whether to launch a potentially game-changing innovation, how much to close down during a pandemic, whether to open an office in a foreign country, or whether to capitulate to a government that seeks to regulate them.

Arguably, there is a continuum, not a category difference, between singular and recurrent decisions. Underwriters may deal with some cases that strike them as very much out of the ordinary. Conversely, if you are buying a house for the fourth time in your life, you have probably started to think of home buying as a recurrent decision. But extreme examples clearly suggest that the difference is meaningful. Going to war is one thing; going through annual budget reviews is another.

Noise in Singular Decisions

Singular decisions have traditionally been treated as quite separate from the recurrent judgments that interchangeable employees routinely make in large organizations. While social scientists have dealt with recurrent decisions, high-stakes singular decisions have been the province of historians and management gurus. The approaches to the two types of decisions have been

quite different. Analyses of recurrent decisions have often taken a statistical bent, with social scientists assessing many similar decisions to discern patterns, identify regularities, and measure accuracy. In contrast, discussions of singular decisions typically adopt a causal view; they are conducted in hindsight and are focused on identifying the causes of what happened. Historical analyses, like case studies of management successes and failures, aim to understand how an essentially unique judgment was made.

The nature of singular decisions raises an important question for the study of noise. We have defined noise as undesirable variability in judgments of the same problem. Since singular problems are never exactly repeated, this definition does not apply to them. After all, history is only run once. You will never be able to compare Obama's decision to send health workers and soldiers to West Africa in 2014 with the decisions other American presidents made about how to handle that particular problem at that particular time (though you can speculate). You may agree to compare your decision to marry that special someone with the decisions of other people like you, but that comparison will not be as relevant to you as the one we made between the quotes of underwriters on the same case. You and your spouse are unique. There is no direct way to observe the presence of noise in singular decisions.

Yet singular decisions are not free from the factors that produce noise in recurrent decisions. At the shooting range, the shooters on Team C (the noisy team) may be adjusting the gunsight on their rifle in different directions, or their hands may just be unsteady. If we observed only the first shooter on the team, we would have no idea how noisy the team is, but the sources of noise would still be there. Similarly, when you make a singular decision, you have to imagine that another decision maker, even one just as competent as you and sharing the same goals and values, would not reach the same conclusion from the same facts. And as the decision maker, you should recognize that you might have made a different decision if some irrelevant aspects of the situation or of the decision-making process had been different.

In other words, we cannot measure noise in a singular decision, but if we think counterfactually, we know for sure that noise is there. Just as the shooter's unsteady hand implies that a single shot *could* have landed somewhere else, noise in the decision makers and in the decision-making process implies that the singular decision *could* have been different.

Consider all the factors that affect a singular decision. If the experts in charge of analyzing the Ebola threat and preparing response plans had been

different people, with different backgrounds and life experiences, would their proposals to President Obama have been the same? If the same facts had been presented in a slightly different manner, would the conversation have unfolded the same way? If the key players had been in a different mood or had been meeting during a snowstorm, would the final decision have been identical? Seen in this light, the singular decision does not seem so determined. Depending on many factors that we are not even aware of, the decision could plausibly have been different.

For another exercise in counterfactual thinking, consider how different countries and regions responded to the COVID-19 crisis. Even when the virus hit them roughly at the same time and in a similar manner, there were wide differences in responses. This variation provides clear evidence of noise in different countries' decision making. But what if the epidemic had struck a single country? In that case, we wouldn't have observed any variability. But our inability to observe variability would not make the decision less noisy.

Controlling Noise in Singular Decisions

This theoretical discussion matters. If singular decisions are just as noisy as recurrent ones, then the strategies that reduce noise in recurrent decisions should also improve the quality of singular decisions.

This is a more counterintuitive prescription than it seems. When you have a one-of-a-kind decision to make, your instinct is probably to treat it as, well, one of a kind. Some even claim that the rules of probabilistic thinking are entirely irrelevant to singular decisions made under uncertainty and that such decisions call for a radically different approach.

Our observations here suggest the opposite advice. From the perspective of noise reduction, *a singular decision is a recurrent decision that happens only once.* Whether you make a decision only once or a hundred times, your goal should be to make it in a way that reduces both bias and noise. And practices that reduce error should be just as effective in your one-of-a-kind decisions as in your repeated ones.

Speaking of Singular Decisions

"The way you approach this unusual opportunity exposes you to noise."

“Remember: a singular decision is a recurrent decision that is made only once.”

“The personal experiences that made you who you are are not truly relevant to this decision.”