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 forprofit 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 5

Measuring Error

It is obvious that a consistent bias can produce costly errors. If a scale adds a constant amount to your weight, if an enthusiastic manager routinely predicts that projects will take half the time they end up taking, or if a timid executive is unduly pessimistic about future sales year after year, the result will be numerous serious mistakes.

We have now seen that noise can produce costly errors as well. If a manager most often predicts that projects will take half the time they ultimately take, and occasionally predicts they will take twice their actual time, it is unhelpful to say that the manager is "on average" right. The different errors add up; they do not cancel out.

An important question, therefore, is how, and how much, bias and noise contribute to error. This chapter aims to answer that question. Its basic message is straightforward: in professional judgments of all kinds, whenever accuracy is the goal, *bias and noise play the same role in the calculation of overall error*. In some cases, the larger contributor will be bias; in other cases it will be noise (and these cases are more common than one might expect). But in every case, a reduction of noise has the same impact on overall error as does a reduction of bias by the same amount. For that reason, the measurement and reduction of noise should have the same high priority as the measurement and reduction of bias.

This conclusion rests on a particular approach to the measurement of error, which has a long history and is generally accepted in science and in statistics. In this chapter, we provide an introductory overview of that history and a sketch of the underlying reasoning.

Should GoodSell Reduce Noise?

Begin by imagining a large retail company named GoodSell, which employs many sales forecasters. Their job is to predict GoodSell's market share in various regions. Perhaps after reading a book on the topic of noise, Amy Simkin, head of the forecasting department at GoodSell, has conducted a noise audit. All forecasters produced independent estimates of the market share in the same region.

Figure 3 shows the (implausibly smooth) results of the noise audit. Amy can see that the forecasts were distributed in the familiar bellshaped curve, also known as the normal or Gaussian distribution. The most frequent forecast, represented by the peak of the bell curve, is 44%. Amy can also see that the forecasting system of the company is quite noisy: the forecasts, which would be identical if all were accurate, vary over a considerable range.

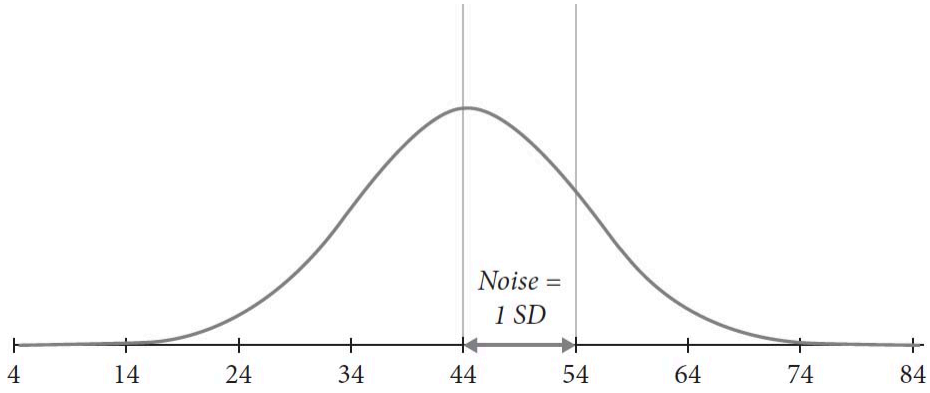

FIGURE 3: *Distribution of GoodSell's market share forecasts for one region*

We can attach a number to the amount of noise in GoodSell's forecasting system. Just as we did when you used your stopwatch to measure laps, we can compute the *standard deviation* of the forecasts. As its name indicates, the standard deviation represents a typical distance from the mean. In this example, it is 10 percentage points. As is true for every normal distribution, about two-thirds of the forecasts are contained within one standard deviation on either side of the mean —in this example, between a 34% and a 54% market share. Amy now has an estimate of the amount of system noise in the forecasts of market share. (A better noise audit would use several forecasting

problems for a more robust estimate, but one is enough for our purpose here.)

As was the case with the executives of the real insurance company of chapter 2, Amy is shocked by the results and wants to take action. The unacceptable amount of noise indicates that the forecasters are not disciplined in implementing the procedures they are expected to follow. Amy asks for authority to hire a noise consultant to achieve more uniformity and discipline in her forecasters' work. Unfortunately, she does not get approval. Her boss's reply seems sensible enough: how, he asks, could we reduce errors when we don't know if our forecasts are right or wrong? Surely, he says, if there is a large average error in the forecasts (i.e., a large bias), addressing it should be the priority. Before undertaking anything to improve its forecasts, he concludes, GoodSell must wait and find out if they are correct.

One year after the original noise audit, the outcome that the forecasters were trying to predict is known. Market share in the target region turned out to be 34%. Now we also know each forecaster's error, which is simply the difference between the forecast and the outcome. The error is 0 for a forecast of 34%, it is 10% for the mean forecast of 44%, and it is −10% for a lowball forecast of 24%.

Figure 4 shows the distribution of errors. It is the same as the distribution of forecasts in figure 3, but the true value (34%) has been subtracted from each forecast. The shape of the distribution has not changed, and the standard deviation (our measure of noise) is still 10%.

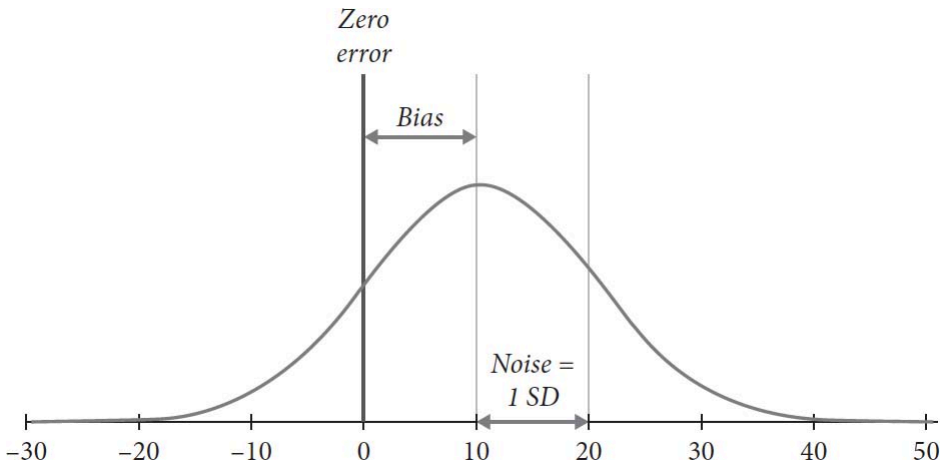

FIGURE 4: *Distribution of errors in GoodSell's forecasts for one region*

The difference between figures 3 and 4 is analogous to the difference between a pattern of shots seen from the back and the front of the target in figures 1 and 2 (see the introduction). Knowing the position of the target was not necessary to observe noise in shooting; similarly, knowing the true outcome adds nothing at all to what was already known about noise in forecasting.

Amy Simkin and her boss now know something they did not know earlier: the amount of bias in the forecasts. Bias is simply the average of errors, which in this case is also 10%. Bias and noise, therefore, happen to be numerically identical in this set of data. (To be clear, this equality of noise and bias is by no means a general rule, but a case in which bias and noise are equal makes it easier to understand their roles.) We can see that most forecasters had made an optimistic error—that is, they overestimated the market share that would be achieved: most of them erred on the right-hand side of the zero-error vertical bar. (In fact, using the properties of the normal distribution, we know that is the case for 84% of the forecasts.)

As Amy's boss notes with barely concealed satisfaction, he was right. There was a lot of bias in the forecasts! And indeed, it is now evident that reducing bias would be a good thing. But, Amy still wonders, would it have been a good idea a year ago—and would it be a good idea now—to reduce noise, too? How would the value of such an improvement compare with the value of reducing bias?

Mean Squares

To answer Amy's question, we need a "scoring rule" for errors, a way to weight and combine individual errors into a single measure of overall error. Fortunately, such a tool exists. It is the *method of least squares,* invented in 1795 by Carl Friedrich Gauss, a famous mathematical prodigy born in 1777, who began a career of major discoveries in his teens.

Gauss proposed a rule for scoring the contribution of individual errors to overall error. His measure of overall error—called *mean squared error* (MSE)—is the average of the squares of the individual errors of measurement.

Gauss's detailed arguments for his approach to the measurement of overall error are far beyond the scope of this book, and his solution is not immediately obvious. Why use the squares of errors? The idea seems arbitrary, even bizarre. Yet, as you will see, it builds on an intuition that you almost certainly share.

To see why, let us turn to what appears to be a completely different problem but turns out to be the same one. Imagine that you are given a ruler and asked to measure the length of a line to the nearest millimeter. You are allowed to make five measurements. They are represented by the downward-pointing triangles on the line in figure 5.

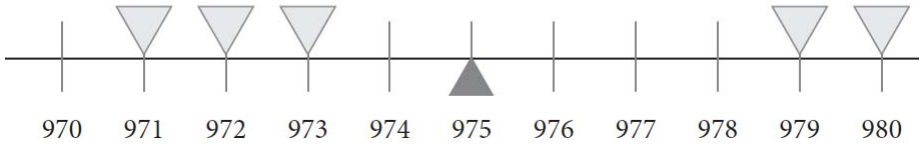

FIGURE 5: *Five measurements of the same length*

As you can see, the five measurements are all between 971 and 980 millimeters. What is your best estimate of the true length of the line? There are two obvious contenders. One possibility is the median number, the measurement that sits between the two shorter measurements and the two longer ones. It is 973 millimeters. The other possibility is the arithmetic mean, known in common parlance as the average, which in this example is 975 millimeters and shown here as an upwardpointing arrow. Your intuition probably favors the mean, and your intuition is correct. The mean contains more information; it is affected by the size of the numbers, while the median is affected only by their order.

There is a tight link between this problem of estimation, about which you have a clear intuition, and the problem of overall error measurement that concerns us here. They are, in fact, two sides of the same coin. That is because the best estimate is one that minimizes the overall error of the available measurements. Accordingly, if your intuition about the mean being the best estimate is correct, the formula you use to measure overall error should be one that yields the arithmetic mean as the value for which error is minimized.

MSE has that property—and it is the only definition of overall error that has it. In figure 6, we have computed the value of MSE in the set of five measurements for ten possible integer values of the line's true length. For instance, if the true value was 971, the errors in the five measurements would be 0, 1, 2, 8, and 9. The squares of these errors add up to 150, and their mean is 30. This is a large number, reflecting the fact that some measurements are far from the true value. You can see that MSE decreases as we get closer to 975—the mean—and increases again beyond that point. The mean is our best estimate because it is the value that minimizes overall error.

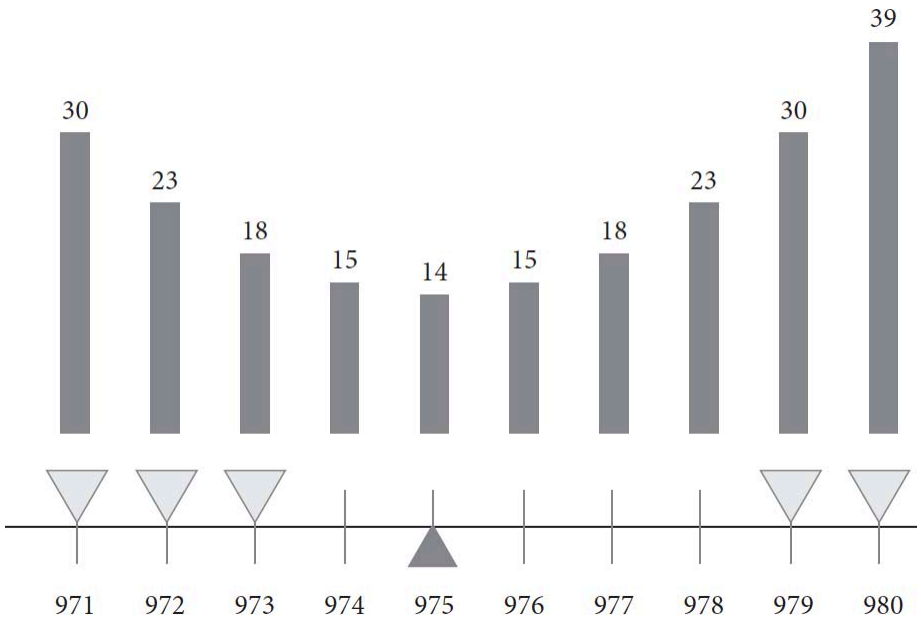

FIGURE 6: *Mean squared error (MSE) for ten possible values of the true length*

You can also see that the overall error increases rapidly when your estimate diverges from the mean. When your estimate increases by just 3 millimeters, from 976 to 979, for instance, MSE doubles. This is a key feature of MSE: squaring gives large errors a far greater weight than it gives small ones.

You now see why Gauss's formula to measure overall error is called mean squared error and why his approach to estimation is called the least squares method. The squaring of errors is its central idea, and no other formula would be compatible with your intuition that the mean is the best estimate.

The advantages of Gauss's approach were quickly recognized by other mathematicians. Among his many feats, Gauss used MSE (and other mathematical innovations) to solve a puzzle that had defeated the best astronomers of Europe: the rediscovery of Ceres, an asteroid that had been traced only briefly before it disappeared into the glare of the sun in 1801. The astronomers had been trying to estimate Ceres's trajectory, but the way they accounted for the measurement error of their telescopes was wrong, and the planet did not reappear anywhere near the location their results suggested. Gauss redid their calculations, using the least squares method. When the astronomers trained their telescopes to the spot that he had indicated, they found Ceres!

Scientists in diverse disciplines were quick to adopt the least squares method. Over two centuries later, it remains the standard way to evaluate errors wherever achieving accuracy is the goal. The weighting of errors by their square is central to statistics. In the vast majority of applications across all scientific disciplines, MSE rules. As we are about to see, the approach has surprising implications.

The Error Equations

The role of bias and noise in error is easily summarized in two expressions that we will call *the error equations.* The first of these equations decomposes the error in a single measurement into the two components with which you are now familiar: bias—the average error—and a residual "noisy error." The noisy error is positive when the error is larger than the bias, negative when it is smaller. The average of noisy errors is zero. Nothing new in the first error equation.

Error in a single measurement = Bias + Noisy Error

The second error equation is a decomposition of MSE, the measure of overall error we have now introduced. Using some simple algebra, MSE can be shown to be equal to the sum of the squares of bias and noise. (Recall that noise is the standard deviation of measurements, which is identical to the standard deviation of noisy errors.) Therefore:

Overall Error (MSE) = Bias^2 + Noise^2

The form of this equation—a sum of two squares—may remind you of a high-school favorite, the Pythagorean theorem. As you might remember, in a right triangle, the sum of the squares of the two shorter sides equals the square of the longest one. This suggests a simple visualization of the error equation, in which MSE, Bias^2, and Noise^2 are the areas of three squares on the sides of a right triangle. Figure 7 shows how MSE (the area of the darker square) equals the sum of the areas of the other two squares. In the left panel, there is more noise than bias; in the right panel, more bias than noise. But MSE is the same, and the error equation holds in both cases.

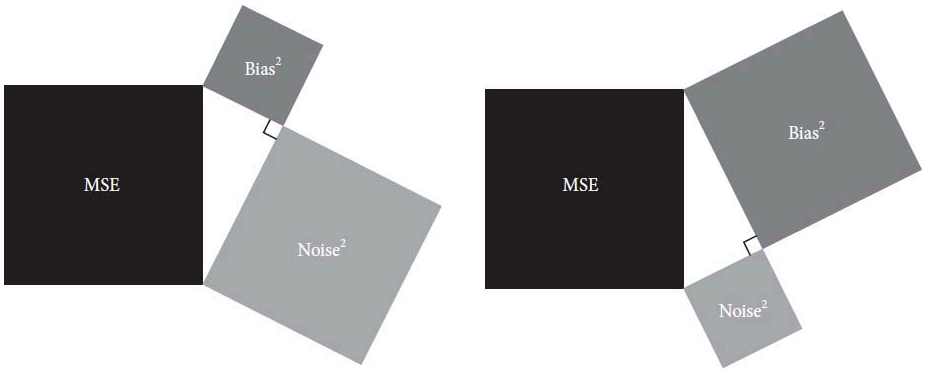

FIGURE 7: *Two decompositions of MSE*

As the mathematical expression and its visual representation both suggest, bias and noise play identical roles in the error equation. They are independent of each other and equally weighted in the determination of overall error. (Note that we will use a similar decomposition into a sum of squares when we analyze the components of noise in later chapters.)

The error equation provides an answer to the practical question that Amy raised: how will reductions in either noise or bias by the same amount affect overall error? The answer is straightforward: bias and noise are interchangeable in the error equation, and the decrease in overall error will be the same, regardless of which of the two is reduced. In figure 4, in which bias and noise happen to be equal (both are 10%), their contributions to overall error are equal.

The error equation also provides unequivocal support for Amy Simkin's initial impulse to try to reduce noise. Whenever you observe

noise, you should work to reduce it! The equation shows that Amy's boss was wrong when he suggested that GoodSell wait to measure the bias in its forecasts and only then decide what to do. In terms of overall error, noise and bias are independent: the benefit of reducing noise is the same, regardless of the amount of bias.

This notion is highly counterintuitive but crucial. To illustrate it, figure 8 shows the effect of reducing bias and noise by the same amount. To help you appreciate what has been accomplished in both panels, the original distribution of errors (from figure 4) is represented by a broken line.

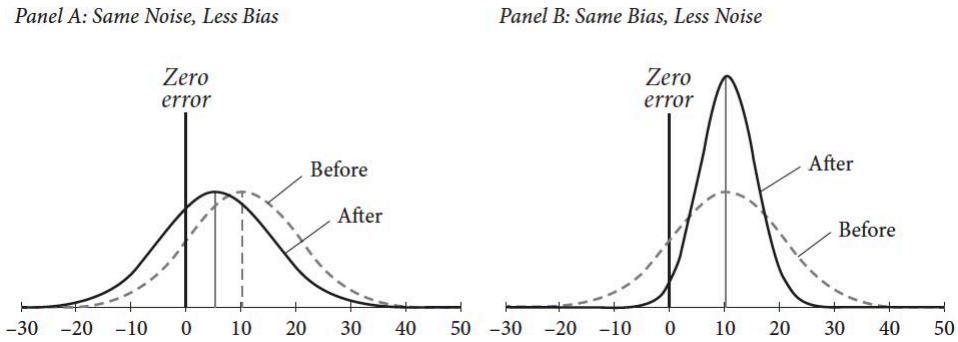

FIGURE 8: *Distribution of errors with bias reduced by half vs. noise reduced by half*

In panel A, we assume that Amy's boss decided to do things his way: he found out what the bias was, then somehow managed to reduce it by half (perhaps by providing feedback to the overoptimistic forecasters). Nothing was done about noise. The improvement is visible: the whole distribution of forecasts has shifted closer to the true value.

In panel B, we show what would have happened if Amy had won the argument. Bias is unchanged, but noise is reduced by half. The paradox here is that noise reduction seems to have made things worse. The forecasts are now more concentrated (less noisy) but not more accurate (not less biased). Whereas 84% of forecasts were on one side of the true value, almost all (98%) now err in the direction of overshooting the true value. Noise reduction seems to have made the forecasts more precisely wrong—hardly the sort of improvement for which Amy hoped!

Despite appearances, however, overall error has been reduced just as much in panel B as in panel A. The illusion of deterioration in panel B arises from an erroneous intuition about bias. The relevant measure

of bias is not the imbalance of positive and negative errors. It is average error, which is the distance between the peak of the bell curve and the true value. In panel B, this average error has not changed from the original situation—it is still high, at 10%, but not worse. True, the presence of bias is now more striking, because it accounts for a larger proportion of overall error (80% rather than 50%). But that is because noise has been reduced. Conversely, in panel A, bias has been reduced, but noise has not. The net result is that MSE is the same in both panels: reducing noise or reducing bias by the same amount has the same effect on MSE.

As this example illustrates, MSE conflicts with common intuitions about the scoring of predictive judgments. To minimize MSE, you must concentrate on avoiding large errors. If you measure length, for example, the effect of reducing an error from 11cm to 10cm is 21 times as large as the effect of going from an error of 1cm to a perfect hit. Unfortunately, people's intuitions in this regard are almost the mirror image of what they should be: people are very keen to get perfect hits and highly sensitive to small errors, but they hardly care at all about the difference between two large errors. Even if you sincerely believe that your goal is to make accurate judgments, your emotional reaction to results may be incompatible with the achievement of accuracy as science defines it.

Of course, the best solution here would be to reduce both noise and bias. Since bias and noise are independent, there is no reason to choose between Amy Simkin and her boss. In that regard, if GoodSell decides to reduce noise, the fact that noise reduction makes bias more visible—indeed, impossible to miss—may turn out to be a blessing. Achieving noise reduction will ensure that bias reduction is next on the company's agenda.

Admittedly, reducing noise would be less of a priority if bias were much larger than noise. But the GoodSell example offers another lesson worth highlighting. In this simplified model, we have assumed that noise and bias are equal. Given the form of the error equation, their contributions to total error are equal, too: bias accounts for 50% of overall error, and so does noise. Yet, as we have noted, 84% of the forecasters err in the same direction. It takes a bias this large (six out of seven people making mistakes in the same direction!) to have as much effect as noise has. We should not be surprised, therefore, to find situations in which there is more noise than bias.

We illustrated the application of the error equation to a single case, one particular region of GoodSell's territory. Of course, it is always desirable to carry out a noise audit on multiple cases at once. Nothing changes. The error equation is applied to the separate cases; and an overall equation is obtained by taking the averages of MSE, bias squared and noise squared over the cases. It would have been better for Amy Simkin to obtain multiple forecasts for several regions, either from the same or from different forecasters. Averaging results would give her a more accurate picture of bias and noise in the forecasting system of GoodSell.

The Cost of Noise

The error equation is the intellectual foundation of this book. It provides the rationale for the goal of reducing system noise in predictive judgments, a goal that is in principle as important as the reduction of statistical bias. (We should emphasize that statistical bias is not a synonym for social discrimination; it is simply the average error in a set of judgments.)

The error equation and the conclusions we have drawn from it depend on the use of MSE as the measure of overall error. The rule is appropriate for purely predictive judgments, including forecasts and estimates, all of which aim to approach a true value with maximum accuracy (the least bias) and precision (the least noise).

The error equation does not apply to evaluative judgments, however, because the concept of error, which depends on the existence of a true value, is far more difficult to apply. Furthermore, even if errors could be specified, their costs would rarely be symmetrical and would be unlikely to be precisely proportional to their square.

For a company that makes elevators, for example, the consequences of errors in estimating the maximum load of an elevator are obviously asymmetrical: underestimation is costly, but overestimation could be catastrophic. Squared error is similarly irrelevant to the decision of when to leave home to catch a train. For that decision, the consequences of being either one minute late or five minutes late are the same. And when the insurance company of chapter 2 prices policies or estimates the value of claims, errors in both directions are costly, but there is no reason to assume that their costs are equivalent.

These examples highlight the need to specify the roles of predictive and evaluative judgments in decisions. A widely accepted maxim of good decision making is that you should not mix your values and your facts. Good decision making must be based on objective and accurate predictive judgments that are completely unaffected by hopes and fears, or by preferences and values. For the elevator company, the first step would be a neutral calculation of the maximum technical load of the elevator under different engineering solutions. Safety becomes a dominant consideration only in the second step, when an evaluative judgment determines the choice of an acceptable safety margin to set the maximum capacity. (To be sure, that choice will also greatly depend on factual judgments involving, for example, the costs and benefits of that safety margin.) Similarly, the first step in deciding when to leave for the station should be an objective determination of the probabilities of different travel times. The respective costs of missing your train and of wasting time at the station become relevant only in your choice of the risk you are willing to accept.

The same logic applies to much more consequential decisions. A military commander must weigh many considerations when deciding whether to launch an offensive, but much of the intelligence on which the leader relies is a matter of predictive judgment. A government responding to a health crisis, such as a pandemic, must weigh the pros and cons of various options, but no evaluation is possible without accurate predictions about the likely consequences of each option (including the decision to do nothing).

In all these examples, the final decisions require evaluative judgments. The decision makers must consider multiple options and apply their values to make the optimal choice. But the decisions depend on underlying predictions, which should be value-neutral. Their goal is accuracy—hitting as close as possible to the bull's-eye—and MSE is the appropriate measure of error. Predictive judgments will be improved by procedures that reduce noise, as long as they do not increase bias to a larger extent.

Speaking of the Error Equation

"Oddly, reducing bias and noise by the same amount has the same effect on accuracy."

“Reducing noise in predictive judgment is always useful, regardless of what you know about bias.”

“When judgments are split 84 to 16 between those that are above and below the true value, there is a large bias—that’s when bias and noise are equal.”

“Predictive judgments are involved in every decision, and accuracy should be their only goal. Keep your values and your facts separate.”

CHAPTER 19

Debiasing and Decision Hygiene

Many researchers and organizations have pursued the goal of debiasing judgments. This chapter examines their central findings. We will distinguish between different types of debiasing interventions and discuss one such intervention that deserves further investigation. We will then turn to the reduction of noise and introduce the idea of decision hygiene.

Ex Post and Ex Ante Debiasing

A good way to characterize the two main approaches to debiasing is to return to the measurement analogy. Suppose that you know that your bathroom scale adds, on average, half a pound to your weight. Your scale is biased. But this does not make it useless. You can address its bias in one of two possible ways. You can correct every reading from your unkindly scale by subtracting half a pound. To be sure, that might get a bit tiresome (and you might forget to do it). An alternative might be to adjust the dial and improve the instrument's accuracy, once and for all.

These two approaches to debiasing measurements have direct analogues in interventions to debias judgments. They work either ex post, by correcting judgments after they have been made, or ex ante, by intervening before a judgment or decision.

Ex post, or corrective, debiasing is often carried out intuitively. Suppose that you are supervising a team in charge of a project and that the team estimates that it can complete its project in three months. You

might want to add a buffer to the members' judgment and plan for four months, or more, thus correcting a bias (the planning fallacy) you assume is present.

This kind of bias correction is sometimes undertaken more systematically. In the United Kingdom, HM Treasury has published *The Green Book*, a guide on how to evaluate programs and projects. The book urges planners to address optimistic biases by applying percentage adjustments to estimates of the cost and duration of a project. These adjustments should ideally be based on an organization's historic levels of optimism bias. If no such historical data is available, *The Green Book* recommends applying generic adjustment percentages for each type of project.

Ex ante or preventive debiasing interventions fall in turn into two broad categories. Some of the most promising are designed to modify the environment in which the judgment or decision takes place. Such modifications, or *nudges*, as they are known, aim to reduce the effect of biases or even to enlist biases to produce a better decision. A simple example is automatic enrollment in pension plans. Designed to overcome inertia, procrastination, and optimistic bias, automatic enrollment ensures that employees will be saving for retirement unless they deliberately opt out. Automatic enrollment has proved to be extremely effective in increasing participation rates. The program is sometimes accompanied by Save More Tomorrow plans, by which employees can agree to earmark a certain percentage of their future wage increases for savings. Automatic enrollment can be used in many places—for example, automatic enrollment in green energy, in free school meal plans for poor children, or in various other benefits programs.

Other nudges work on different aspects of choice architecture. They might make the right decision the easy decision—for example, by reducing administrative burdens for getting access to care for mental health problems. Or they might make certain characteristics of a product or an activity salient—for example, by making once-hidden fees explicit and clear. Grocery stores and websites can easily be designed to nudge people in a way that overcomes their biases. If healthy foods are put in prominent places, more people are likely to buy them.

A different type of ex ante debiasing involves training decision makers to recognize their biases and to overcome them. Some of these

interventions have been called *boosting;* they aim to improve people's capacities—for example, by teaching them statistical literacy.

Educating people to overcome their biases is an honorable enterprise, but it is more challenging than it seems. Of course, education is useful. For instance, people who have taken years of advanced statistics classes are less likely to make errors in statistical reasoning. But teaching people to avoid biases is hard. Decades of research have shown that professionals who have learned to avoid biases in their area of expertise often struggle to apply what they have learned to different fields. Weather forecasters, for instance, have learned to avoid overconfidence in their forecasts. When they announce a 70% chance of rain, it rains, by and large, 70% of the time. Yet they can be just as overconfident as other people when asked general-knowledge questions. The challenge of learning to overcome a bias is to recognize that a new problem is similar to one we have seen elsewhere and that a bias that we have seen in one place is likely to materialize in other places.

Researchers and educators have had some success using nontraditional teaching methods to facilitate this recognition. In one study, Carey Morewedge of Boston University and his colleagues used instructional videos and "serious games." Participants learned to recognize errors caused by confirmation bias, anchoring, and other psychological biases. After each game, they received feedback on the errors they had made and learned how to avoid making them again. The games (and, to a lesser extent, the videos) reduced the number of errors that participants made on a test immediately afterward and again eight weeks later, when they were asked similar questions. In a separate study, Anne-Laure Sellier and her colleagues found that MBA students who had played an instructional video game in which they learned to overcome confirmation bias applied this learning when solving a business case in another class. They did so even though they were not told that there was any connection between the two exercises.

A Limitation of Debiasing

Whether they correct biases ex post or prevent their effects through nudging or boosting, most debiasing approaches have one thing in

common: they target a specific bias, which they assume is present. This often-reasonable assumption is sometimes wrong.

Consider again the example of project planning. You can reasonably assume that overconfidence affects project teams in general, but you cannot be sure that it is the only bias (or even the main one) affecting a particular project team. Maybe the team leader has had a bad experience with a similar project and so has learned to be especially conservative when making estimates. The team thus exhibits the opposite error from the one you thought you should correct. Or perhaps the team developed its forecast by analogy with another similar project and was anchored on the time it took to complete that project. Or maybe the project team, anticipating that you would add a buffer to its estimate, has preempted your adjustment by making its recommendation even more bullish than its true belief.

Or consider an investment decision. Overconfidence about the investment's prospects may certainly be at work, but another powerful bias, loss aversion, has the opposite effect, making decision makers loath to risk losing their initial outlay. Or consider a company allocating resources across multiple projects. Decision makers may be both bullish about the effect of new initiatives (overconfidence again) and too timid in diverting resources from existing units (a problem caused by *status quo bias,* which, as the name indicates, is our preference for leaving things as they are).

As these examples illustrate, it is difficult to know exactly which psychological biases are affecting a judgment. In any situation of some complexity, multiple psychological biases may be at work, conspiring to add error in the same direction or offsetting one another, with unpredictable consequences.

The upshot is that ex post or ex ante debiasing—which, respectively, correct or prevent specific psychological biases—are useful in some situations. These approaches work where the general direction of error is known and manifests itself as a clear statistical bias. Types of decisions that are expected to be strongly biased are likely to benefit from debiasing interventions. For instance, the planning fallacy is a sufficiently robust finding to warrant debiasing interventions against overconfident planning.

The problem is that in many situations, the likely direction of error is not known in advance. Such situations include all those in which the effect of psychological biases is variable among judges and essentially

unpredictable—resulting in system noise. To reduce error under circumstances like these, we need to cast a broader net to try to detect more than one psychological bias at a time.

The Decision Observer

We suggest undertaking this search for biases neither before nor after the decision is made, but in real time. Of course, people are rarely aware of their own biases when they are being misled by them. This lack of awareness is itself a known bias, the *bias blind spot.* People often recognize biases more easily in others than they do in themselves. We suggest that observers can be trained to spot, in real time, the diagnostic signs that one or several familiar biases are affecting someone else's decisions or recommendations.

To illustrate how the process might work, imagine a group that attempts to make a complex and consequential judgment. The judgment could be of any type: a government deciding on possible responses to a pandemic or other crisis, a case conference in which physicians are exploring the best treatment for a patient with complex symptoms, a corporate board deciding on a major strategic move. Now imagine a *decision observer,* someone who watches this group and uses a checklist to diagnose whether any biases may be pushing the group away from the best possible judgment.

A decision observer is not an easy role to play, and no doubt, in some organizations it is not realistic. Detecting biases is useless if the ultimate decision makers are not committed to fighting them. Indeed, the decision makers must be the ones who initiate the process of decision observation and who support the role of the decision observer. We certainly do not recommend that you make yourself a self-appointed decision observer. You will neither win friends nor influence people.

Informal experiments suggest, however, that real progress can be made with this approach. At least, the approach is helpful under the right conditions, especially when the leaders of an organization or team are truly committed to the effort, and when the decision observers are well chosen—and not susceptible to serious biases of their own.

Decision observers in these cases fall in three categories. In some organizations, the role can be played by a supervisor. Instead of

monitoring only the substance of the proposals that are submitted by a project team, the supervisor might also pay close attention to the *process* by which they are developed and to the team's dynamics. This makes the observer alert to biases that may have affected the proposal's development. Other organizations might assign a member of each working team to be the team's "bias buster"; this guardian of the decision process reminds teammates in real time of the biases that may mislead them. The downside of this approach is that the decision observer is placed in the position of a devil's advocate inside the team and may quickly run out of political capital. Finally, other organizations might rely on an outside facilitator, who has the advantage of a neutral perspective (and the attending disadvantages in terms of inside knowledge and costs).

To be effective, decision observers need some training and tools. One such tool is a checklist of the biases they are attempting to detect. The case for relying on a checklist is clear: checklists have a long history of improving decisions in high-stakes contexts and are particularly well suited to preventing the repetition of past errors.

Here is an example. In the United States, federal agencies must compile a formal regulatory impact analysis before they issue expensive regulations designed to clean the air or water, reduce deaths in the workplace, increase food safety, respond to public health crises, reduce greenhouse gas emissions, or increase homeland security. A dense, technical document with an unlovely name (OMB Circular A-4) and spanning nearly fifty pages sets out the requirements of the analysis. The requirements are clearly designed to counteract bias. Agencies must explain why the regulation is needed, consider both more and less stringent alternatives, consider both costs and benefits, present the information in an unbiased manner, and discount the future appropriately. But in many agencies, government officials have not complied with the requirements of that dense, technical document. (They might not even have read it.) In response, federal officials produced a simple checklist, consisting of just one and one-half pages, to reduce the risk that agencies will ignore, or fail to attend to, any of the major requirements.

To illustrate what a bias checklist might look like, we have included one as appendix B. This generic checklist is merely an example; any decision observer will certainly want to develop one that is customized to the needs of the organization, both to enhance its relevance and

facilitate its adoption. Importantly, a checklist is not an exhaustive list of all the biases that can affect a decision; it aims to focus on the most frequent and most consequential ones.

Decision observation with appropriate bias checklists can help limit the effect of biases. Although we have seen some encouraging results in informal, small-scale efforts, we are not aware of any systematic exploration of the effects of this approach or of the pros and cons of the various possible ways to deploy it. We hope to inspire more experimentation, both by practitioners and by researchers, of the practice of real-time debiasing by decision observers.

Noise Reduction: Decision Hygiene

Bias is error we can often see and even explain. It is directional: that is why a nudge can limit the detrimental effects of a bias, or why an effort to boost judgment can combat specific biases. It is also often visible: that is why an observer can hope to diagnose biases in real time as a decision is being made.

Noise, on the other hand, is unpredictable error that we cannot easily see or explain. That is why we so often neglect it—even when it causes grave damage. For this reason, strategies for noise reduction are to debiasing what preventive hygiene measures are to medical treatment: the goal is to prevent an unspecified range of potential errors before they occur.

We call this approach to noise reduction *decision hygiene.* When you wash your hands, you may not know precisely which germ you are avoiding—you just know that handwashing is good prevention for a variety of germs (especially but not only during a pandemic). Similarly, following the principles of decision hygiene means that you adopt techniques that reduce noise without ever knowing which underlying errors you are helping to avoid.

The analogy with handwashing is intentional. Hygiene measures can be tedious. Their benefits are not directly visible; you might never know what problem they prevented from occurring. Conversely, when problems do arise, they may not be traceable to a specific breakdown in hygiene observance. For these reasons, handwashing compliance is difficult to enforce, even among health-care professionals, who are well aware of its importance.

Just like handwashing and other forms of prevention, decision hygiene is invaluable but thankless. Correcting a well-identified bias may at least give you a tangible sense of achieving something. But the procedures that reduce noise will not. They will, statistically, prevent many errors. Yet you will never know *which* errors. Noise is an invisible enemy, and preventing the assault of an invisible enemy can yield only an invisible victory.

Given how much damage noise can cause, that invisible victory is nonetheless worth the battle. The following chapters introduce several decision hygiene strategies used in multiple domains, including forensic science, forecasting, medicine, and human resources. In chapter 25, we will review these strategies and show how they can be combined in an integrated approach to noise reduction.

Speaking of Debiasing and Decision Hygiene

"Do you know what specific bias you're fighting and in what direction it affects the outcome? If not, there are probably several biases at work, and it is hard to predict which one will dominate."

"Before we start discussing this decision, let's designate a decision observer."

"We have kept good decision hygiene in this decision process; chances are the decision is as good as it can be."