# 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.”

PART II

# Your Mind Is a Measuring Instrument

**M**easurement, in everyday life as in science, is the act of using an instrument to assign a value on a scale to an object or event. You measure the length of a carpet in inches, using a tape measure. You measure the temperature in degrees Fahrenheit or Celsius by consulting a thermometer.

The act of making a judgment is similar. When judges determine the appropriate prison term for a crime, they assign a value on a scale. So do underwriters when they set a dollar value to insure a risk, or doctors when they make a diagnosis. (The scale need not be numerical: "guilty beyond a reasonable doubt," "advanced melanoma," and "surgery is recommended" are judgments, too.)

Judgment can therefore be described as *measurement in which the instrument is a human mind*. Implicit in the notion of measurement is the goal of accuracy—to approach truth and minimize error. The goal of judgment is not to impress, not to take a stand, not to persuade. It is important to note that the concept of judgment as we use it here is borrowed from the technical psychological literature, and that it is a much narrower concept than the same word has in everyday language. *Judgment* is not a synonym for *thinking,* and *making accurate judgments* is not a synonym for *having good judgment.*

As we define it, a judgment is a conclusion that can be summarized in a word or phrase. If an intelligence analyst writes a long report leading to the conclusion that a regime is unstable, only the conclusion is a judgment. *Judgment,* like *measurement,* refers both to the mental

activity of making a judgment and to its product. And we will sometimes use *judge* as a technical term to describe people who make judgments, even when they have nothing to do with the judiciary.

Although accuracy is the goal, perfection in achieving this goal is never achieved even in scientific measurement, much less in judgment. There is always some error, some of which is bias and some of which is noise.

To experience how noise and bias contribute to error, we invite you to play a game that will take you less than one minute. If you have a smartphone with a stopwatch, it probably has a lap function, which enables you to measure consecutive time intervals without stopping the stopwatch or even looking at the display. Your goal is to produce five consecutive laps of exactly ten seconds without looking at the phone. You may want to observe a ten-second interval a few times before you begin. Go.

Now look at the lap durations recorded on your phone. (The phone itself was not free from noise, but there was very little of it.) You will see that the laps are not all exactly ten seconds and that they vary over a substantial range. You tried to reproduce the same timing exactly, but you were unable to do so. The variability you could not control is an instance of noise.

This finding is hardly surprising, because noise is universal in physiology and psychology. Variability across individuals is a biological given; no two peas in a pod are truly identical. Within the same person, there is variability, too. Your heartbeat is not exactly regular. You cannot repeat the same gesture with perfect precision. And when you have your hearing examined by an audiologist, there will be some sounds so soft you never hear them, and others so loud you always do. But there will also be some sounds that you will sometimes hear and sometimes miss.

Now look at the five numbers on your phone. Do you see a pattern? For instance, are all five laps shorter than ten seconds, a pattern suggesting that your internal clock is running fast? In this simple task, the bias is the difference, positive or negative, between the mean of your laps and ten seconds. Noise constitutes the variability of your results, analogous to the scatter of shots we saw earlier. In statistics, the most common measure of variability is *standard deviation,* and we will use it to measure noise in judgments.

We can think of most judgments, specifically *predictive* judgments, as similar to the measurements you just made. When we make a prediction, we attempt to come close to a true value. An economic forecaster aims to be as close as possible to the true value of the growth in next year's gross domestic product; a doctor aims to make the correct diagnosis. (Note that the term *prediction,* in the technical sense used in this book, does not imply predicting the future: for our purposes, the diagnosis of an existing medical condition is a prediction.)

We will rely extensively on the analogy between judgment and measurement because it helps explain the role of noise in error. People who make predictive judgments are just like the shooter who aims at the bull's-eye or the physicist who strives to measure the true weight of a particle. Noise in their judgments implies error. Simply put, when a judgment aims at a true value, two different judgments cannot both be right. Like measuring instruments, some people generally show more error than others in a particular task—perhaps because of deficiencies in skill or training. But, like measuring instruments, the people who make judgments are never perfect. We need to understand and measure their errors.

Of course, most professional judgments are far more complex than the measurement of a time interval. In chapter 4, we define different types of professional judgments and explore what they aim at. In chapter 5, we discuss how to measure error and how to quantify the contribution of system noise to it. Chapter 6 dives deeper into system noise and identifies its components, which are different types of noise. In chapter 7, we explore one of these components: occasion noise. Finally, in chapter 8, we show how groups often amplify noise in judgments.

A simple conclusion emerges from these chapters: like a measuring instrument, the human mind is imperfect—it is both biased and noisy. Why, and by how much? Let's find out.

# CHAPTER 9

# Judgments and Models

Many people are interested in forecasting people's future performance on the job—their own and that of others. The forecasting of performance is therefore a useful example of predictive professional judgment. Consider, for instance, two executives in a large company. Monica and Nathalie were assessed by a specialized consulting firm when they were hired and received ratings on a 1-to-10 scale for leadership, communication, interpersonal skills, job-related technical skills, and motivation for the next position (table 2). Your task is to predict their performance evaluations two years after they were hired, also using a 1-to-10 scale.

**Table 2:** *Two candidates for an executive position*

| | Leadership | Communication | Interpersonal skills | Technical skills | Motivation | Your prediction |
|---|---|---|---|---|---|---|
| **Monica** | 4 | 6 | 4 | 8 | 8 | |
| **Nathalie** | 8 | 10 | 6 | 7 | 6 | |

Most people, when faced with this type of problem, simply eyeball each line and produce a quick judgment, sometimes after mentally computing the average of the scores. If you just did that, you probably concluded that Nathalie was the stronger candidate and that the difference between her and Monica was 1 or 2 points.

## Judgment or Formula?

The informal approach you took to this problem is known as *clinical judgment*. You consider the information, perhaps engage in a quick computation, consult your intuition, and come up with a judgment. In fact, clinical judgment is the process that we have described simply as judgment in this book.

Now suppose that you performed the prediction task as a participant in an experiment. Monica and Nathalie were drawn from a database of several hundred managers who were hired some years ago, and who received ratings on five separate dimensions. You used these ratings to predict the managers' success on the job. Evaluations of their performance in their new roles are now available. How closely would these evaluations align with your clinical judgments of their potential?

This example is loosely based on an actual study of performance prediction. If you had been a participant in that study, you would probably not be pleased with its results. Doctoral-level psychologists, employed by an international consulting firm to make such predictions, achieved a correlation of .15 with performance evaluations (PC = 55%). In other words, when they rated one candidate as stronger than another —as you did with Monica and Nathalie—the probability that their favored candidate would end up with a higher performance rating was 55%, barely better than chance. To say the least, that is not an impressive result.

Perhaps you think that accuracy was poor because the ratings you were shown were useless for prediction. So we must ask, how much useful predictive information do the candidates' ratings actually contain? How can they be combined into a predictive score that will have the highest possible correlation with performance?

A standard statistical method answers these questions. In the present study, it yields an optimal correlation of .32 (PC = 60%), far from impressive but substantially higher than what clinical predictions achieved.

This technique, called *multiple regression,* produces a predictive score that is a weighted average of the predictors. It finds the optimal set of weights, chosen to maximize the correlation between the composite prediction and the target variable. The optimal weights minimize the MSE (mean squared error) of the predictions—a prime example of the dominant role of the least squares principle in statistics.

As you might expect, the predictor that is most closely correlated with the target variable gets a large weight, and useless predictors get a weight of zero. Weights could also be negative: the candidate's number of unpaid traffic tickets would probably get a negative weight as a predictor of managerial success.

The use of multiple regression is an example of *mechanical prediction.* There are many kinds of mechanical prediction, ranging from simple rules ("hire anyone who completed high school") to sophisticated artificial intelligence models. But linear regression models are the most common (they have been called "the workhorse of judgment and decision-making research"). To minimize jargon, we will refer to linear models as *simple models.*

The study that we illustrated with Monica and Nathalie was one of many comparisons of clinical and mechanical predictions, which all share a simple structure:

- A set of *predictor variables* (in our example, the ratings of candidates) are used to predict a *target outcome* (the job evaluations of the same people);
- Human judges make *clinical predictions;*
- A rule (such as multiple regression) uses the same predictors to produce *mechanical predictions* of the same outcomes;
- The overall accuracy of clinical and mechanical predictions is compared.

## Meehl: The Optimal Model Beats You

When people are introduced to clinical and mechanical prediction, they want to know how the two compare. How good is human judgment, relative to a formula?

The question had been asked before, but it attracted much attention only in 1954, when Paul Meehl, a professor of psychology at the University of Minnesota, published a book titled *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Meehl reviewed twenty studies in which a clinical judgment was pitted against a mechanical prediction for such outcomes as academic success and psychiatric prognosis. He reached the strong conclusion that simple mechanical rules were generally superior to human judgment. Meehl discovered that clinicians and other professionals are distressingly weak in what they often see as their unique strength: the ability to integrate information.

To appreciate how surprising this finding is, and how it relates to noise, you have to understand how a simple mechanical prediction model works. Its defining characteristic is that the same rule is applied to all the cases. Each predictor has a weight, and that weight does not vary from one case to the next. You might think that this severe constraint puts models at a great disadvantage relative to human judges. In our example, perhaps you thought that Monica's combination of motivation and technical skills would be an important asset and would offset her limitations in other areas. And perhaps you also thought that Nathalie's weaknesses in these two areas would not be a serious issue, given her other strengths. Implicitly, you imagined different routes to success for the two women. These plausible clinical speculations effectively assign different weights to the same predictors in the two cases—a subtlety that is out of the reach of a simple model.

Another constraint of the simple model is that an increase of 1 unit in a predictor always produces the same effect (and half the effect of an increase of 2 units). Clinical intuition often violates this rule. If, for instance, you were impressed by Nathalie's perfect 10 on communication skills and decided this score was worth a boost in your prediction, you did something that a simple model will not do. In a weighted-average formula, the difference between a score of 10 and a score of 9 must be the same as the difference between a 7 and a 6. Clinical judgment does not obey that rule. Instead, it reflects the common intuition that the same difference can be inconsequential in one context and critical in another. You may want to check, but we suspect that no simple model could account exactly for your judgments about Monica and Nathalie.

The study we used for these cases was a clear example of Meehl's pattern. As we noted, clinical predictions achieved a .15 correlation (PC = 55%) with job performance, but mechanical prediction achieved a correlation of .32 (PC = 60%). Think about the confidence that you experienced in the relative merits of the cases of Monica and Nathalie. Meehl's results strongly suggest that any satisfaction you felt with the quality of your judgment was an illusion: the *illusion of validity.*

The illusion of validity is found wherever predictive judgments are made, because of a common failure to distinguish between two stages of the prediction task: evaluating cases on the evidence available and predicting actual outcomes. You can often be quite confident in your assessment of which of two candidates *looks* better, but guessing

which of them will actually *be* better is an altogether different kettle of fish. It is safe to assert, for instance, that Nathalie looks like a stronger candidate than Monica, but it is not at all safe to assert that Nathalie will be a more successful executive than Monica. The reason is straightforward: you know most of what you need to know to assess the two cases, but gazing into the future is deeply uncertain.

Unfortunately, the difference gets blurred in our thinking. If you find yourself confused by the distinction between cases and predictions, you are in excellent company: Everybody finds that distinction confusing. If you are as confident in your predictions as you are in your evaluation of cases, however, you are a victim of the illusion of validity.

Clinicians are not immune to the illusion of validity. You can surely imagine the response of clinical psychologists to Meehl's finding that trivial formulas, consistently applied, outdo clinical judgment. The reaction combined shock, disbelief, and contempt for the shallow research that pretended to study the marvels of clinical intuition. The reaction is easy to understand: Meehl's pattern contradicts the subjective experience of judgment, and most of us will trust our experience over a scholar's claim.

Meehl himself was ambivalent about his findings. Because his name is associated with the superiority of statistics over clinical judgment, we might imagine him as a relentless critic of human insight, or as the godfather of quants, as we would say today. But that would be a caricature. Meehl, in addition to his academic career, was a practicing psychoanalyst. A picture of Freud hung in his office. He was a polymath who taught classes not just in psychology but also in philosophy and law and who wrote about metaphysics, religion, political science, and even parapsychology. (He insisted that "there is something to telepathy.") None of these characteristics fits the stereotype of a hard-nosed numbers guy. Meehl had no ill will toward clinicians—far from it. But as he put it, the evidence for the advantage of the mechanical approach to combining inputs was "massive and consistent."

"Massive and consistent" is a fair description. A 2000 review of 136 studies confirmed unambiguously that mechanical aggregation outperforms clinical judgment. The research surveyed in the article covered a wide variety of topics, including diagnosis of jaundice, fitness for military service, and marital satisfaction. Mechanical prediction was more accurate in 63 of the studies, a statistical tie was declared for another 65, and clinical prediction won the contest in 8 cases. These

results understate the advantages of mechanical prediction, which is also faster and cheaper than clinical judgment. Moreover, human judges actually had an unfair advantage in many of these studies, because they had access to "private" information that was not supplied to the computer model. The findings support a blunt conclusion: *simple models beat humans.*

## Goldberg: The Model of You Beats You

Meehl's finding raises important questions. Why, exactly, is the formula superior? What does the formula do better? In fact, a better question would be to ask what humans do worse. The answer is that people are inferior to statistical models in many ways. One of their critical weaknesses is that they are noisy.

To support this conclusion, we turn to a different stream of research on simple models, which began in the small city of Eugene, Oregon. Paul Hoffman was a wealthy and visionary psychologist who was impatient with academia. He founded a research institute where he collected under one roof a few extraordinarily effective researchers, who turned Eugene into a world-famous center for the study of human judgment.

One of these researchers was Lewis Goldberg, who is best known for his leading role in the development of the Big Five model of personality. In the late 1960s, following earlier work by Hoffman, Goldberg studied statistical models that describe the judgments of an individual.

It is just as easy to build such a model of a judge as it is to build a model of reality. The same predictors are used. In our initial example, the predictors are the five ratings of a manager's performance. And the same tool, multiple regression, is used. The only difference is the target variable. Instead of predicting a set of real outcomes, the formula is applied to predict a set of judgments—for instance, *your* judgments of Monica, Nathalie, and other managers.

The idea of modeling your judgments as a weighted average may seem altogether bizarre, because this is not how you form your opinions. When you thought clinically about Monica and Nathalie, you didn't apply the same rule to both cases. Indeed, you did not apply any

rule at all. The model of the judge is not a realistic description of how a judge actually judges.

However, even if you do not actually compute a linear formula, you might still make your judgments *as if* you did. Expert billiard players act as if they have solved the complex equations that describe the mechanics of a particular shot, even if they are doing nothing of the kind. Similarly, you could be generating predictions as if you used a simple formula—even if what you actually do is vastly more complex. An as-if model that predicts what people will do with reasonable accuracy is useful, even when it is obviously wrong as a description of the process. This is the case for simple models of judgment. A comprehensive review of studies of judgment found that, in 237 studies, the average correlation between the model of the judge and the judge's clinical judgments was .80 (PC = 79%). While far from perfect, this correlation is high enough to support an as-if theory.

The question that drove Goldberg's research was how well a simple model of the judge would predict real outcomes. Since the model is a crude approximation of the judge, we could sensibly assume that it cannot perform as well. How much accuracy is lost when the model replaces the judge?

The answer may surprise you. Predictions did not lose accuracy when the model generated predictions. They improved. In most cases, the model out-predicted the professional on which it was based. The ersatz was better than the original product.

This conclusion has been confirmed by studies in many fields. An early replication of Goldberg's work involved the forecasting of graduate school success. The researchers asked ninety-eight participants to predict the GPAs of ninety students from ten cues. On the basis of these predictions, the researchers built a linear model of each participant's judgments and compared how accurately the participants and the models of the participants predicted GPA. For every one of the ninety-eight participants, the model did better than the participant did! Decades later, a review of fifty years of research concluded that models of judges consistently outperformed the judges they modeled.

We do not know if the participants in these studies received personal feedback on their performance. But you can surely imagine your own dismay if someone told you that a crude model of your judgments—almost a caricature—was actually more accurate than you were. For

most of us, the activity of judgment is complex, rich, and interesting precisely because it does not fit simple rules. We feel best about ourselves and about our ability to make judgments when we invent and apply complex rules or have an insight that makes an individual case different from others—in short, when we make judgments that are not reducible to a plain operation of weighted averaging. The model-of-the-judge studies reinforce Meehl's conclusion that the subtlety is largely wasted. Complexity and richness do not generally lead to more accurate predictions.

Why is that so? To understand Goldberg's finding, we need to understand what accounts for the differences between you and the model of you. What causes the discrepancies between your actual judgments and the output of a simple model that predicts them?

A statistical model of your judgments cannot possibly add anything to the information they contain. All the model can do is subtract and simplify. In particular, the simple model of your judgments will not represent any complex rules that you consistently follow. If you think that the difference between 10 and 9 on a rating of communications skill is more significant than the difference between 7 and 6, or that a well-rounded candidate who scores a solid 7 on all dimensions is preferable to one who achieves the same average with clear strengths and marked weaknesses, the model of you will not reproduce your complex rules—even if you apply them with flawless consistency.

Failing to reproduce your subtle rules will result in a loss of accuracy when your subtlety is valid. Suppose, for instance, that you must predict success at a difficult task from two inputs, skill and motivation. A weighted average is not a good formula, because no amount of motivation is sufficient to overcome a severe skill deficit, and vice versa. If you use a more complex combination of the two inputs, your predictive accuracy will be enhanced and will be higher than that achieved by a model that fails to capture this subtlety. On the other hand, complex rules will often give you only the illusion of validity and in fact harm the quality of your judgments. Some subtleties are valid, but many are not.

In addition, a simple model of you will not represent the pattern noise in your judgments. It cannot replicate the positive and negative errors that arise from arbitrary reactions you may have to a particular case. Neither will the model capture the influences of the momentary context and of your mental state when you make a particular judgment. Most

likely, these noisy errors of judgment are not systematically correlated with anything, which means that for most purposes, they can be considered random.

The effect of removing noise from your judgments will always be an improvement of your predictive accuracy. For example, suppose that the correlation between your forecasts and an outcome is .50 (PC = 67%), but 50% of the variance of your judgments consists of noise. If your judgments could be made noise-free—as a model of you would be—their correlation with the same outcome would jump to .71 (PC = 75%). Reducing noise mechanically increases the validity of predictive judgment.

In short, replacing you with a model of you does two things: it eliminates your subtlety, and it eliminates your pattern noise. The robust finding that the model of the judge is more valid than the judge conveys an important message: the gains from subtle rules in human judgment—when they exist—are generally not sufficient to compensate for the detrimental effects of noise. You may believe that you are subtler, more insightful, and more nuanced than the linear caricature of your thinking. But in fact, you are mostly noisier.

Why do complex rules of prediction harm accuracy, despite the strong feeling we have that they draw on valid insights? For one thing, many of the complex rules that people invent are not likely to be generally true. But there is another problem: even when the complex rules are valid in principle, they inevitably apply under conditions that are rarely observed. For example, suppose you have concluded that exceptionally original candidates are worth hiring, even when their scores on other dimensions are mediocre. The problem is that exceptionally original candidates are, by definition, exceptionally rare. Since an evaluation of originality is likely to be unreliable, many high scores on that metric are flukes, and truly original talent often remains undetected. The performance evaluations that could confirm that "originals" end up as superstars are also imperfect. Errors of measurement at both ends inevitably attenuate the validity of predictions—and rare events are particularly likely to be missed. The advantages of true subtlety are quickly drowned in measurement error.

A study by Martin Yu and Nathan Kuncel reported a more radical version of Goldberg's demonstration. This study (which was the basis for the example of Monica and Nathalie) used data from an international consulting firm that employed experts to assess 847

candidates for executive positions, in three separate samples. The experts scored the results on seven distinct assessment dimensions and used their clinical judgment to assign an overall predictive score to each, with rather unimpressive results.

Yu and Kuncel decided to compare judges not to the best simple model of themselves but to a *random* linear model. They generated ten thousand sets of random weights for the seven predictors, and applied the ten thousand random formulas to predict job performance.

Their striking finding was that *any* linear model, when applied consistently to all cases, was likely to outdo human judges in predicting an outcome from the same information. In one of the three samples, 77% of the ten thousand randomly weighted linear models did better than the human experts. In the other two samples, 100% of the random models outperformed the humans. Or, to put it bluntly, it proved almost impossible in that study to generate a simple model that did worse than the experts did.

The conclusion from this research is stronger than the one we took away from Goldberg's work on the model of the judge—and indeed it is an extreme example. In this setting, human judges performed very poorly in absolute terms, which helps explain why even unimpressive linear models outdid them. Of course, we should not conclude that any model beats any human. Still, the fact that mechanical adherence to a simple rule (Yu and Kuncel call it "mindless consistency") could significantly improve judgment in a difficult problem illustrates the massive effect of noise on the validity of clinical predictions.

This quick tour has shown how noise impairs clinical judgment. In predictive judgments, human experts are easily outperformed by simple formulas—models of reality, models of a judge, or even randomly generated models. This finding argues in favor of using noisefree methods: rules and algorithms, which are the topic of the next chapter.

## Speaking of Judgments and Models

"People believe they capture complexity and add subtlety when they make judgments. But the complexity and the subtlety are mostly wasted—usually they do not add to the accuracy of simple models."

"More than sixty years after the publication of Paul Meehl's book, the idea that mechanical prediction is superior to people is still

shocking."

"There is so much noise in judgment that a noise-free model of a judge achieves more accurate predictions than the actual judge does."

# CHAPTER 21

# Selection and Aggregation in Forecasting

Many judgments involve forecasting. What is the unemployment rate likely to be in the next quarter? How many electric cars will be sold next year? What will be the effects of climate change in 2050? How long will it take to complete a new building? What will be the annual earnings of a particular company? How will a new employee perform? What will be the cost of a new air pollution regulation? Who will win an election? The answers to such questions have major consequences. Fundamental choices of private and public institutions often depend on them.

Analysts of forecasting—of when it goes wrong and why—make a sharp distinction between bias and noise (also called inconsistency or unreliability). Everyone agrees that in some contexts, forecasters are biased. For example, official agencies show unrealistic optimism in their budget forecasts. On average, they project unrealistically high economic growth and unrealistically low deficits. For practical purposes, it matters little whether their unrealistic optimism is a product of a cognitive bias or political considerations.

In addition, forecasters tend to be overconfident: if asked to formulate their forecasts as confidence intervals rather than as point estimates, they tend to pick narrower intervals than they should. For instance, an ongoing quarterly survey asks the chief financial officers of US companies to estimate the annual return of the S&P 500 index for the next year. The CFOs provide two numbers: a minimum, below which they think there is a one-in-ten chance the actual return will be, and a maximum, which they believe the actual return has a one-in-ten chance of exceeding. Thus the two numbers are the bounds of an 80% confidence interval. Yet the realized returns fall in that interval only 36%

of the time. The CFOs are far too confident in the precision of their forecasts.

Forecasters are also noisy. A reference text, J. Scott Armstrong's *Principles of Forecasting,* points out that even among experts, "unreliability is a source of error in judgmental forecasting." In fact noise is a major source of error. Occasion noise is common; forecasters do not always agree with themselves. Between-person noise is also pervasive; forecasters disagree with one another, even if they are specialists. If you ask law professors to predict Supreme Court rulings, you will find a great deal of noise. If you ask specialists to project the annual benefits of air pollution regulation, you will find massive variability, with ranges of, for example, $3 billion to $9 billion. If you ask a group of economists to make forecasts about unemployment and growth, you will also find great variability. We have already seen many examples of noisy forecasts, and research on forecasting uncovers many more.

## Improving Forecasts

The research also offers suggestions for reducing noise and bias. We will not review them exhaustively here, but we will focus on two noise-reduction strategies that have broad applicability. One is an application of the principle we mentioned in chapter 18: selecting better judges produces better judgments. The other is one of the most universally applicable decision hygiene strategies: aggregating multiple independent estimates.

The easiest way to aggregate several forecasts is to average them. Averaging is mathematically guaranteed to reduce noise: specifically, it divides it by the square root of the number of judgments averaged. This means that if you average one hundred judgments, you will reduce noise by 90%, and if you average four hundred judgments, you will reduce it by 95%—essentially eliminating it. This statistical law is the engine of the wisdom-of-crowds approach, discussed in chapter 7.

Because averaging does nothing to reduce bias, its effect on total error (MSE) depends on the proportions of bias and noise in it. That is why the wisdom of crowds works best when judgments are independent, and therefore less likely to contain shared biases. Empirically, ample evidence suggests that averaging multiple forecasts

greatly increases accuracy, for instance in the “consensus” forecast of economic forecasters of stock analysts. With respect to sales forecasting, weather forecasting, and economic forecasting, the unweighted average of a group of forecasters outperforms most and sometimes all individual forecasts. Averaging forecasts obtained by different methods has the same effect: in an analysis of thirty empirical comparisons in diverse domains, combined forecasts reduced errors by an average of 12.5%.

Straight averaging is not the only way to aggregate forecasts. A *select-crowd* strategy, which selects the best judges according to the accuracy of their recent judgments and averages the judgments of a small number of judges (e.g., five), can be as effective as straight averaging. It is also easier for decision makers who respect expertise to understand and adopt a strategy that relies not only on aggregation but also on selection.

One method to produce aggregate forecasts is to use *prediction markets,* in which individuals bet on likely outcomes and are thus incentivized to make the right forecasts. Much of the time, prediction markets have been found to do very well, in the sense that if the prediction market price suggests that events are, say, 70% likely to happen, they happen about 70% of the time. Many companies in various industries have used prediction markets to aggregate diverse views.

Another formal process for aggregating diverse views is known as the Delphi method. In its classic form, this method involves multiple rounds during which the participants submit estimates (or votes) to a moderator and remain anonymous to one another. At each new round, the participants provide reasons for their estimates and respond to the reasons given by others, still anonymously. The process encourages estimates to converge (and sometimes forces them to do so by requiring new judgments to fall within a specific range of the distribution of previous-round judgments). The method benefits both from aggregation and social learning.

The Delphi method has worked well in many situations, but it can be challenging to implement. A simpler version, *mini-Delphi,* can be deployed within a single meeting. Also called *estimate-talk-estimate,* it requires participants first to produce separate (and silent) estimates, then to explain and justify them, and finally to make a new estimate in response to the estimates and explanations of others. The consensus

judgment is the average of the individual estimates obtained in that second round.

## The Good Judgment Project

Some of the most innovative work on the quality of forecasting, going well beyond what we have explored thus far, started in 2011, when three prominent behavioral scientists founded the Good Judgment Project. Philip Tetlock (whom we encountered in chapter 11 when we discussed his assessment of long-term forecasts of political events); his spouse, Barbara Mellers; and Don Moore teamed up to improve our understanding of forecasting and, in particular, why some people are good at it.

The Good Judgment Project started with the recruitment of tens of thousands of volunteers—not specialists or experts but ordinary people from many walks of life. They were asked to answer hundreds of questions, such as these:

- *Will North Korea detonate a nuclear device before the end of the year?*
- *Will Russia officially annex additional Ukrainian territory in the next three months?*
- *Will India or Brazil become a permanent member of the UN Security Council in the next two years?*
- *In the next year, will any country withdraw from the eurozone?*

As these examples show, the project has focused on large questions about world events. Importantly, efforts to answer such questions raise many of the same problems that more mundane forecasts do. If a lawyer is asking whether a client will win in court, or if a television studio is asked whether a proposed show will be a big hit, forecasting skills are involved. Tetlock and his colleagues wanted to learn whether some people are especially good forecasters. They also wanted to learn whether the ability to forecast could be taught or at least improved.

To understand the central findings, we need to explain some key aspects of the method adopted by Tetlock and his team to evaluate forecasters. First, they used a large number of forecasts, not just one or a few, where luck might be responsible for success or failure. If you predict that your favorite sports team will win its next game, and it does, you are not necessarily a good forecaster. Maybe you *always* predict

that your favorite team will win: if that's your strategy, and if they win only half the time, your forecasting ability is not especially impressive. To reduce the role of luck, the researchers examined how participants did, on average, across numerous forecasts.

Second, the researchers asked participants to make their forecasts in terms of probabilities that an event would happen, rather than a binary "it will happen" or "it will not happen." To many people, forecasting means the latter—taking a stand one way or the other. However, given our objective ignorance of future events, it is much better to formulate probabilistic forecasts. If someone said in 2016, "Hillary Clinton is 70% likely to be elected president," he is not necessarily a bad forecaster. Things that are correctly said to be 70% likely will not happen 30% of the time. To know whether forecasters are good, we should ask whether their probability estimates map onto reality. Suppose that a particular forecaster named Margaret says that 500 different events are 60% likely. If 300 of them actually happen, then we can conclude that Margaret's confidence is well *calibrated.* Good calibration is one requirement for good forecasting.

Third, as an added refinement, Tetlock and colleagues did not just ask their forecasters to make *one* probability estimate about whether an event would happen in, say, twelve months. They gave the participants the opportunity to revise their forecasts continuously in light of new information. Suppose that you had estimated, back in 2016, that the United Kingdom had only a 30% chance of leaving the European Union before the end of 2019. As new polls came out, suggesting that the "Leave" vote was gaining ground, you probably would have revised your forecast upward. When the result of the referendum was known, it was still uncertain whether the United Kingdom would leave the union within that time frame, but it certainly looked a lot more probable. (Brexit technically happened in 2020.)

With each new piece of information, Tetlock and his colleagues allowed the forecasters to update their forecasts. For scoring purposes, each one of these updates is treated as a new forecast. That way, participants in the Good Judgment Project are incentivized to monitor the news and update their forecasts continuously. This approach mirrors what is expected of forecasters in business and government, who should also be updating their forecasts frequently on the basis of new information, despite the risk of being criticized for changing their minds. (A well-known response to this criticism, sometimes attributed to

John Maynard Keynes, is, “When the facts change, I change my mind. What do *you* do?”)

Fourth, to score the performance of the forecasters, the Good Judgment Project used a system developed by Glenn W. Brier in 1950. *Brier scores,* as they are known, measure the distance between what people forecast and what actually happens.

Brier scores are a clever way to get around a pervasive problem associated with probabilistic forecasts: the incentive for forecasters to hedge their bets by never taking a bold stance. Think again of Margaret, whom we described as a well-calibrated forecaster because she rated 500 events as 60% likely, and 300 of those events did happen. This result may not be as impressive as it seems. If Margaret is a weather forecaster who *always* predicts a 60% chance of rain and there are 300 rainy days out of 500, Margaret’s forecasts are well calibrated but also practically useless. Margaret, in essence, is telling you that, just in case, you might want to carry an umbrella every day. Compare her with Nicholas, who predicts a 100% chance of rain on the 300 days when it will rain, and a 0% chance of rain on the 200 dry days. Nicholas has the same perfect calibration as Margaret: when either forecaster predicts that X% of the days will be rainy, rain falls precisely X% of the time. But Nicholas’s forecasts are much more valuable: instead of hedging his bets, he is willing to tell you whether you should take an umbrella. Technically, Nicholas is said to have a high *resolution* in addition to good calibration.

Brier scores reward both good calibration and good resolution. To produce a good score, you have not only to be right on average (i.e., well calibrated) but also to be willing to take a stand and differentiate among forecasts (i.e., have high resolution). Brier scores are based on the logic of mean squared errors, and lower scores are better: a score of 0 would be perfect.

So, now that we know how they were scored, how well did the Good Judgment Project volunteers do? One of the major findings was that the overwhelming majority of the volunteers did poorly, but about 2% stood out. As mentioned earlier, Tetlock calls these well-performing people superforecasters. They were hardly unerring, but their predictions were much better than chance. Remarkably, one government official said that the group did significantly “better than the average for intelligence community analysts who could read intercepts and other secret data.” This comparison is worth pausing over.

Intelligence community analysts are trained to make accurate forecasts; they are not amateurs. In addition, they have access to classified information. And yet they do not do as well as the superforecasters do.

## Perpetual Beta

What makes superforecasters so good? Consistent with our argument in chapter 18, we could reasonably speculate that they are unusually intelligent. That speculation is not wrong. On GMA tests, the superforecasters do better than the average volunteer in the Good Judgment Project (and the average volunteer is significantly above the national average). But the difference isn't all that large, and many volunteers who do extremely well on intelligence tests do not qualify as superforecasters. Apart from general intelligence, we could reasonably expect that superforecasters are unusually good with numbers. And they are. But their real advantage is not their talent at math; it is their ease in thinking analytically and probabilistically.

Consider superforecasters' willingness and ability to structure and disaggregate problems. Rather than form a holistic judgment about a big geopolitical question (whether a nation will leave the European Union, whether a war will break out in a particular place, whether a public official will be assassinated), they break it up into its component parts. They ask, "What would it take for the answer to be yes? What would it take for the answer to be no?" Instead of offering a gut feeling or some kind of global hunch, they ask and try to answer an assortment of subsidiary questions.

Superforecasters also excel at taking the outside view, and they care a lot about base rates. As explained for the Gambardi problem in chapter 13, before you focus on the specifics of Gambardi's profile, it helps to know the probability that the average CEO will be fired or quit in the next two years. Superforecasters systematically look for base rates. Asked whether the next year will bring an armed clash between China and Vietnam over a border dispute, superforecasters do not focus only or immediately on whether China and Vietnam are getting along right now. They might have an intuition about this, in light of the news and analysis they have read. But they know that their intuition about one event is generally not a good guide. Instead they start by looking for a base rate: they ask how often past border disputes have

escalated into armed clashes. If such clashes are rare, superforecasters will begin by incorporating that fact and only then turn to the details of the China–Vietnam situation.

In short, what distinguishes the superforecasters isn't their sheer intelligence; it's *how* they apply it. The skills they bring to bear reflect the sort of cognitive style we described in chapter 18 as likely to result in better judgments, particularly a high level of "active open-mindedness." Recall the test for actively open-minded thinking: it includes such statements as "People should take into consideration evidence that goes against their beliefs" and "It is more useful to pay attention to people who disagree with you than to pay attention to those who agree." Clearly, people who score high on this test are not shy about updating their judgments (without overreacting) when new information becomes available.

To characterize the thinking style of superforecasters, Tetlock uses the phrase "perpetual beta," a term used by computer programmers for a program that is not meant to be released in a final version but that is endlessly used, analyzed, and improved. Tetlock finds that "the strongest predictor of rising into the ranks of superforecasters is perpetual beta, the degree to which one is committed to belief updating and self-improvement." As he puts it, "What makes them so good is less what they are than what they do—the hard work of research, the careful thought and self-criticism, the gathering and synthesizing of other perspectives, the granular judgments and relentless updating." They like a particular cycle of thinking: "try, fail, analyze, adjust, try again."

## Noise and Bias in Forecasting

At this point, you might be tempted to think that people can be trained to be superforecasters or at least to perform more like them. And indeed, Tetlock and his collaborators have worked to do exactly that. Their efforts should be considered the second stage of understanding why superforecasters perform so well and how to make them perform better.

In an important study, Tetlock and his team randomly assigned regular (nonsuper) forecasters to three groups, in which they tested the effect of different interventions on the quality of subsequent judgments.

These interventions exemplify three of the strategies we have described to improve judgments:

1. *Training:* Several forecasters completed a tutorial designed to improve their abilities by teaching them probabilistic reasoning. In the tutorial, the forecasters learned about various biases (including base-rate neglect, overconfidence, and confirmation bias); the importance of averaging multiple predictions from diverse sources; and considering reference classes.

2. *Teaming (a form of aggregation):* Some forecasters were asked to work in teams in which they could see and debate one another's predictions. Teaming could increase accuracy by encouraging forecasters to deal with opposing arguments and to be actively open-minded.

3. *Selection:* All forecasters were scored for accuracy, and at the end of a full year, the top 2% were designated as superforecasters and given the opportunity to work together in elite teams the following year.

As it turns out, all three interventions worked, in the sense that they improved people's Brier scores. Training made a difference, teaming made a larger one, and selection had an even larger effect.

This important finding confirms the value of aggregating judgments and selecting good judges. But it is not the full story. Armed with data about the effects of each intervention, Ville Satopää, who collaborated with Tetlock and Mellers, developed a sophisticated statistical technique to tease out how, exactly, each intervention improved forecasts. In principle, he reasoned, there are three major reasons why some forecasters can perform better or worse than others:

1. They can be more skilled at finding and analyzing data in the environment that are relevant to the prediction they have to make. This explanation points to the importance of information.

2. Some forecasters may have a general tendency to err on a particular side of the true value of a forecast. If, out of hundreds of forecasts, you systematically overestimate or underestimate the probability that certain changes from the status quo will occur, you can be said to suffer from a form of bias, in favor of either change or stability.

3. Some forecasters may be less susceptible to noise (or random errors). In forecasting, as in any judgment, noise can have many triggers. Forecasters may overreact to a particular piece of news (this is an example of what we have called pattern noise), they may be subject to occasion noise, or they may be noisy in their use of the probability scale. All these errors (and many more) are unpredictable in their size and direction.

Satopää, Tetlock, Mellers, and their colleague Marat Salikhov called their model the BIN (bias, information, and noise) model for forecasting. They set out to measure how much each of the three components was responsible for the performance improvement in each of the three interventions.

Their answer was simple: all three interventions worked primarily by reducing noise. As the researchers put it, "Whenever an intervention boosted accuracy, it worked mainly by suppressing random errors in judgment. Curiously, the original intent of the training intervention was to reduce bias."

Since the training was designed to reduce biases, a less-than-super forecaster would have predicted that bias reduction would be the major effect of the training. Yet the training worked by reducing noise. The surprise is easily explained. Tetlock's training is designed to fight *psychological* biases. As you now know, the effect of psychological biases is not always a statistical bias. When they affect different individuals on different judgments in different ways, psychological biases produce noise. This is clearly the case here, as the events being forecast are quite varied. The same biases can lead a forecaster to overreact or underreact, depending on the topic. We should not expect them to produce a *statistical* bias, defined as the general tendency of a forecaster to believe that events will happen or not happen. As a result, training forecasters to fight their psychological biases works—by reducing noise.

Teaming had a comparably large effect on noise reduction, but it also significantly improved the ability of the teams to extract information. This result is consistent with the logic of aggregation: several brains that work together are better at finding information than one is. If Alice and Brian are working together, and Alice has spotted signals that Brian has missed, their joint forecast will be better. When working in groups, the superforecasters seem capable of avoiding the dangers of

group polarization and information cascades. Instead, they pool their data and insights and, in their actively open-minded way, make the most of the combined information. Satopää and his colleagues explain this advantage: "Teaming—unlike training … allows forecasters to harness the information."

Selection had the largest total effect. Some of the improvement comes from a better use of information. Superforecasters are better than others at finding relevant information—possibly because they are smarter, more motivated, and more experienced at making these kinds of forecasts than is the average participant. But the main effect of selection is, again, to reduce noise. Superforecasters are less noisy than regular players or even trained teams. This finding, too, was a surprise to Satopää and the other researchers: "'Superforecasters' may owe their success more to superior discipline in tamping down measurement error, than to incisive readings of the news" that others cannot replicate.

## Where Selection and Aggregation Work

The success of the superforecasting project highlights the value of two decision hygiene strategies: *selection* (the superforecasters are, well, super) and *aggregation* (when they work in teams, forecasters perform better). The two strategies are broadly applicable in many judgments. Whenever possible, you should aim to combine the strategies, by constructing teams of judges (e.g., forecasters, investment professionals, recruiting officers) who are selected for being both good at what they do *and* complementary to one another.

So far, we have considered the improved precision that is achieved by averaging multiple independent judgments, as in the wisdom-of-crowds experiments. Aggregating the estimates of higher-validity judges will further improve accuracy. Yet another gain in accuracy can be obtained by combining judgments that are both independent and complementary. Imagine that four people are witnesses to a crime: it is essential, of course, to make sure that they do not influence one another. If, in addition, they have seen the crime from four different angles, the quality of the information they provide will be much better.

The task of assembling a team of professionals to make judgments together resembles the task of assembling a battery of tests to predict

the future performance of candidates at school or on the job. The standard tool for that task is multiple regression (introduced in chapter 9). It works by selecting variables in succession. The test that best predicts the outcome is selected first. However, the next test to be included is not necessarily the second most valid. Instead, it is the one that *adds* the most predictive power to the first test, by providing predictions that are both valid and not redundant with the first. For example, suppose you have two tests of mental aptitude, which correlate .50 and .45 with future performance, and a test of personality that correlates only .30 with performance but is uncorrelated with aptitude tests. The optimal solution is to pick the more valid aptitude test first, then the personality test, which brings more new information.

Similarly, if you are assembling a team of judges, you should of course pick the best judge first. But your next choice may be a moderately valid individual who brings some new skill to the table rather than a more valid judge who is highly similar to the first one. A team selected in this manner will be superior because the validity of pooled judgments increases faster when the judgments are uncorrelated with one another than when they are redundant. Pattern noise will be relatively high in such a team because individual judgments of each case will differ. Paradoxically, the average of that noisy group will be more accurate than the average of a unanimous one.

An important caveat is in order. Regardless of diversity, aggregation can only reduce noise if judgments are truly independent. As our discussion of noise in groups has highlighted, group deliberation often adds more error in bias than it removes in noise. Organizations that want to harness the power of diversity must welcome the disagreements that will arise when team members reach their judgments independently. Eliciting and aggregating judgments that are both independent and diverse will often be the easiest, cheapest, and most broadly applicable decision hygiene strategy.

## Speaking of Selection and Aggregation

“Let’s take the average of four independent judgments—this is guaranteed to reduce noise by half.”
“We should strive to be in perpetual beta, like the superforecasters.”
“Before we discuss this situation, what is the relevant base rate?”

“We have a good team, but how can we ensure more diversity of opinions?”