A Flaw in Human Judgment

DANIEL KAHNEMAN

AUTHOR OF *THINKING, FAST AND SLOW*

OLIVIER SIBONY

CASS R. SUNSTEIN

NOISE

A Flaw in Human Judgment

Daniel Kahneman
Olivier Sibony
Cass R. Sunstein

Copyright

William Collins
An imprint of HarperCollins*Publishers*
1 London Bridge Street
London SE1 9GF

WilliamCollinsBooks.com

HarperCollins*Publishers*
1st Floor, Watermarque Building, Ringsend Road
Dublin 4, Ireland

This eBook first published in Great Britain by William Collins in 2021

A catalogue record for this book is available from the British Library

Source ISBN: 9780008308995
Ebook Edition © May 2021 ISBN: 9780008309015

Version: 2021-04-30

Dedication

For Noga, Ori and Gili—DK
For Fantin and Lélia—OS
For Samantha—CRS

Contents

INTRODUCTION

Two Kinds of Error

Imagine that four teams of friends have gone to a shooting arcade. Each team consists of five people; they share one rifle, and each person fires one shot. Figure 1 shows their results.

In an ideal world, every shot would hit the bull's-eye.

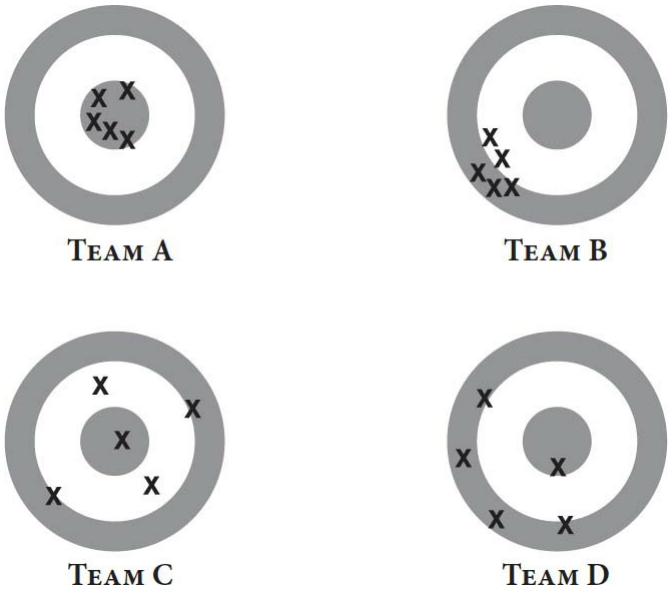

FIGURE 1: *Four teams*

That is nearly the case for Team A. The team's shots are tightly clustered around the bull's-eye, close to a perfect pattern.

We call Team B *biased* because its shots are systematically off target. As the figure illustrates, the consistency of the bias supports a prediction. If one of the team's members were to take another shot, we would bet on its landing in the same area as the first five. The consistency of the bias also invites a causal explanation: perhaps the gunsight on the team's rifle was bent.

We call Team C *noisy* because its shots are widely scattered. There is no obvious bias, because the impacts are roughly centered on the bull's-eye. If one of the team's members took another shot, we would know very little about where it is likely to hit. Furthermore, no interesting hypothesis comes to mind to explain the results of Team C. We know that its members are poor shots. We do not know why they are so noisy.

Team D is both biased and noisy. Like Team B, its shots are systematically off target; like Team C, its shots are widely scattered.

But this is not a book about target shooting. Our topic is human error. Bias and noise—systematic deviation and random scatter—are different components of error. The targets illustrate the difference.

The shooting range is a metaphor for what can go wrong in human judgment, especially in the diverse decisions that people make on behalf of organizations. In these situations, we will find the two types of error illustrated in figure 1. Some judgments are biased; they are systematically off target. Other judgments are noisy, as people who are expected to agree end up at very different points around the target. Many organizations, unfortunately, are afflicted by both bias and noise.

Figure 2 illustrates an important difference between bias and noise. It shows what you would see at the shooting range if you were shown only the backs of the targets at which the teams were shooting, without any indication of the bull's-eye they were aiming at.

From the back of the target, you cannot tell whether Team A or Team B is closer to the bull's-eye. But you can tell at a glance that Teams C and D are noisy and that Teams A and B are not. Indeed, you know just as much about scatter as you did in figure 1. A general property of noise is that you can recognize and measure it while knowing nothing about the target or bias.

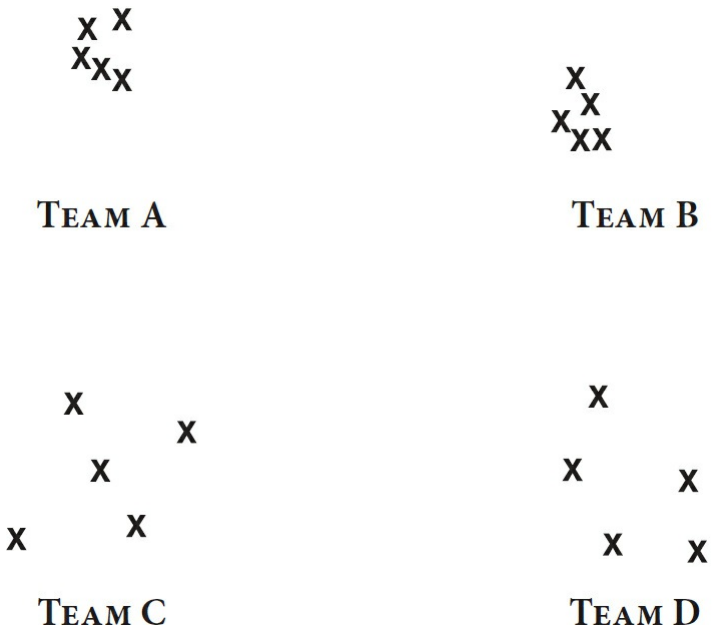

FIGURE 2: *Looking at the back of the target*

The general property of noise just mentioned is essential for our purposes in this book, because many of our conclusions are drawn from judgments whose true answer is unknown or even unknowable. When physicians offer different diagnoses for the same patient, we can study their disagreement without knowing what ails the patient. When film executives estimate the market for a movie, we can study the variability of their answers without knowing how much the film eventually made or even if it was produced at all. We don't need to know who is right to measure how much the judgments of the same case vary. All we have to do to measure noise is look at the back of the target.

To understand error in judgment, we must understand both bias and noise. Sometimes, as we will see, noise is the more important problem. But in public conversations about human error and in organizations all over the world, noise is rarely recognized. Bias is the star of the show. Noise is a bit player, usually offstage. The topic of bias has been discussed in thousands of scientific articles and dozens of popular books, few of which even mention the issue of noise. This book is our attempt to redress the balance.

In real-world decisions, the amount of noise is often scandalously high. Here are a few examples of the alarming amount of noise in situations in which accuracy matters:

- *Medicine is noisy.* Faced with the same patient, different doctors make different judgments about whether patients have skin cancer, breast cancer, heart disease, tuberculosis, pneumonia, depression, and a host of other conditions. Noise is especially high in psychiatry,

where subjective judgment is obviously important. However, considerable noise is also found in areas where it might not be expected, such as in the reading of X-rays.

- *Child custody decisions are noisy.* Case managers in child protection agencies must assess whether children are at risk of abuse and, if so, whether to place them in foster care. The system is noisy, given that some managers are much more likely than others to send a child to foster care. Years later, more of the unlucky children who have been assigned to foster care by these heavy-handed managers have poor life outcomes: higher delinquency rates, higher teen birth rates, and lower earnings.
- *Forecasts are noisy.* Professional forecasters offer highly variable predictions about likely sales of a new product, likely growth in the unemployment rate, the likelihood of bankruptcy for troubled companies, and just about everything else. Not only do they disagree with each other, but they also disagree with themselves. For example, when the same software developers were asked on two separate days to estimate the completion time for the same task, the hours they projected differed by 71%, on average.
- *Asylum decisions are noisy.* Whether an asylum seeker will be admitted into the United States depends on something like a lottery. A study of cases that were randomly allotted to different judges found that one judge admitted 5% of applicants, while another admitted 88%. The title of the study says it all: "Refugee Roulette." (We are going to see a lot of roulette.)
- *Personnel decisions are noisy.* Interviewers of job candidates make widely different assessments of the same people. Performance ratings of the same employees are also highly variable and depend more on the person doing the assessment than on the performance being assessed.
- *Bail decisions are noisy.* Whether an accused person will be granted bail or instead sent to jail pending trial depends partly on the identity of the judge who ends up hearing the case. Some judges are far more lenient than others. Judges also differ markedly in their assessment of which defendants present the highest risk of flight or reoffending.
- *Forensic science is noisy.* We have been trained to think of fingerprint identification as infallible. But fingerprint examiners sometimes differ in deciding whether a print found at a crime scene matches that of a suspect. Not only do experts disagree, but the same experts sometimes make inconsistent decisions when presented with the same print on different occasions. Similar variability has been documented in other forensic science disciplines, even DNA analysis.
- *Decisions to grant patents are noisy.* The authors of a leading study on patent applications emphasize the noise involved: "Whether the patent office grants or rejects a patent is significantly related to the happenstance of which examiner is assigned the application." This variability is obviously troublesome from the standpoint of equity.

All these noisy situations are the tip of a large iceberg. Wherever you look at human judgments, you are likely to find noise. To improve the quality of our judgments, we need to overcome noise as well as bias.

This book comes in six parts. In part 1, we explore the difference between noise and bias, and we show that both public and private organizations can be noisy, sometimes shockingly so. To appreciate the problem, we begin with judgments in two areas. The first involves criminal sentencing (and hence the public sector). The second involves

insurance (and hence the private sector). At first glance, the two areas could not be more different. But with respect to noise, they have much in common. To establish that point, we introduce the idea of a noise audit, designed to measure how much disagreement there is among professionals considering the same cases within an organization.

In part 2, we investigate the nature of human judgment and explore how to measure accuracy and error. Judgments are susceptible to both bias and noise. We describe a striking equivalence in the roles of the two types of error. Occasion noise is the variability in judgments of the same case by the same person or group on different occasions. A surprising amount of occasion noise arises in group discussion because of seemingly irrelevant factors, such as who speaks first.

Part 3 takes a deeper look at one type of judgment that has been researched extensively: predictive judgment. We explore the key advantage of rules, formulas, and algorithms over humans when it comes to making predictions: contrary to popular belief, it is not so much the superior insight of rules but their noiselessness. We discuss the ultimate limit on the quality of predictive judgment—objective ignorance of the future—and how it conspires with noise to limit the quality of prediction. Finally, we address a question that you will almost certainly have asked yourself by then: if noise is so ubiquitous, then why had you not noticed it before?

Part 4 turns to human psychology. We explain the central causes of noise. These include interpersonal differences arising from a variety of factors, including personality and cognitive style; idiosyncratic variations in the weighting of different considerations; and the different uses that people make of the very same scales. We explore why people are oblivious to noise and are frequently unsurprised by events and judgments they could not possibly have predicted.

Part 5 explores the practical question of how you can improve your judgments and prevent error. (Readers who are primarily interested in practical applications of noise reduction might skip the discussion of the challenges of prediction and of the psychology of judgment in parts 3 and 4 and move directly to this part.) We investigate efforts to tackle noise in medicine, business, education, government, and elsewhere. We introduce several noise-reduction techniques that we collect under the label of *decision hygiene.* We present five case studies of domains in which there is much documented noise and in which people have made sustained efforts to reduce it, with instructively varying degrees of success. The case studies include unreliable medical diagnoses, performance ratings, forensic

science, hiring decisions, and forecasting in general. We conclude by offering a system we call the *mediating assessments protocol:* a general-purpose approach to the evaluation of options that incorporates several key practices of decision hygiene and aims to produce less noisy and more reliable judgments.

What is the right level of noise? Part 6 turns to this question. Perhaps counterintuitively, the right level is not zero. In some areas, it just isn't feasible to eliminate noise. In other areas, it is too expensive to do so. In still other areas, efforts to reduce noise would compromise important competing values. For example, efforts to eliminate noise could undermine morale and give people a sense that they are being treated like cogs in a machine. When algorithms are part of the answer, they raise an assortment of objections; we address some of them here. Still, the current level of noise is unacceptable. We urge both private and public organizations to conduct noise audits and to undertake, with unprecedented seriousness, stronger efforts to reduce noise. Should they do so, organizations could reduce widespread unfairness—and reduce costs in many areas.

With that aspiration in mind, we end each chapter with a few brief propositions in the form of quotations. You can use these statements as they are or adapt them for any issues that matter to you, whether they involve health, safety, education, money, employment, entertainment, or something else. Understanding the problem of noise, and trying to solve it, is a work in progress and a collective endeavor. All of us have opportunities to contribute to this work. This book is written in the hope that we can seize those opportunities.

PART I

Finding Noise

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

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

Both of these examples involve studies of a large number of people making a large number of judgments. But many important judgments are *singular* rather than repeated: how to handle an apparently unique business opportunity, whether to launch a whole new product, how to deal with a

pandemic, whether to hire someone who just doesn't meet the standard profile. Can noise be found in decisions about unique situations like these? It is tempting to think that it is absent there. After all, noise is unwanted variability, and how can you have variability with singular decisions? In chapter 3, we try to answer this question. The judgment that you make, even in a seemingly unique situation, is one in a cloud of possibilities. You will find a lot of noise there as well.

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

CHAPTER 1

Crime and Noisy Punishment

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As revealing as they are, these studies, which involve tightly controlled experiments, almost certainly understate the magnitude of noise in the real

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

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

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

Reducing Noise in Sentencing

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

Responding to the evidence of unjustified variability, Congress enacted the Sentencing Reform Act of 1984.

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

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

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

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