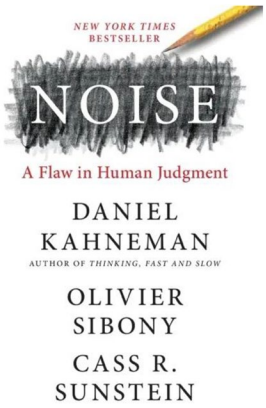
Noise: A Flaw in Human Judgment

by Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein

*Reviewed by*David Aldous



Little, Brown Spark, 2021,
464 pp., US$ 32.00

A statistics professor who had proposed to write a 400-page book on *variance* and promised a best-seller would undoubtedly have been laughed out of the publisher's office. These three eminent authors, from what one might call applied psychology, were evidently more convincing. Their stated theme, "wherever there is judgment, there is *noise*—and more of it than we think" (p. 12)—summarizes the book well. Judgment refers to human judgment and noise to the fact that different humans often make different judgments in the same case. This is hardly news, but the book provides an interesting range of examples plus suggestions for improvements.

To quantitatively minded folk, judgment involves deciding on a number. Three initial examples in the book are (i) a judge deciding on the length of a prison sentence for a convicted felon; (ii) a jury deciding on the amount of damages in a successful lawsuit; (iii) an insurance underwriter deciding the premium to charge on an unusual policy. We have data both from natural experiments (for example, different judges giving sentences of different lengths in different but similar cases) and from artificial experiments (for example, different individuals each stating what they would decide in the same case). And indeed, data show that in such contexts, the variability is much larger than one might expect or wish. For instance, in considering the difference between two underwriters' decisions as a percentage of the average, one study showed that the median difference was 55% (p. 26).

The relevant freshman probability and statistics course starts by considering a group of individuals predicting the value of (say) some economic indicator at a given future time. The average accuracy of those predictions can be measured at that future time by the mean squared error (MSE). A line or two of algebra gives the key identity

$$\text{MSE} = \text{bias}^2 + \text{noise}^2, \qquad (*)$$

where *bias* is the difference between the average of the group predictions and the true value, and *noise* is the standard deviation of the predictions. In that "ultimately known true value" setting, we will know the values of both terms on the right of (∗). The key conceptual point is that in most examples like (i)–(iii) above, there is no "true value," but (∗) is still correct in that both bias and noise contribute (and, in some sense, contribute equally) to the prediction error. This book comments that we tend to pay attention to possible bias—one judge may be lenient and another harsh—but do not pay enough attention to noise. That is, for jail sentences, we might look at the average sentence for a particular crime and ponder whether that is too long or short in our opinion, but we are inclined not to look at the variability.

The authors develop and illustrate this "noise issue" in many directions. There is a general discussion of the spectrum between predictive judgment (of a number) and evaluative judgment (of, for example, exam grades). They suggest that noise be decomposed into level noise, pattern noise, and occasion noise. They briefly mention cognitive biases of

https://doi.org/10.1007/s00283-022-10207-9

individuals—anchoring and matching, etc.—as described in Daniel Kahneman's bestseller *Thinking, Fast and Slow* [1]. They give cases in which the use of simple rules beats human judgment and (in the context of little data rather than big data) also beats classical statistical regression techniques. One nice example involves a recent study of "predicting life trajectories" from childhood [4]. It turned out that no such method could predict well—this is intrinsic unpredictability. In the spirit of the narrative fallacy emphasized in *The Black Swan* [5], they remark that we tend to think of events in terms of causality rather than randomness in the sense of intrinsic unpredictability. A curious omission is scant mention of the use of analytics to complement judgment in professional sports, in the spirit of *Moneyball* [2].

The material above, from Parts I–IV of the book, is all rather uncontroversial, and perhaps familiar from earlier accounts of decisions under uncertainty, but (as the authors say) it is often not emphasized enough, so this book provides a useful explicit account. The remaining third of the book, comprising Parts V–VI (Chapters 18–28), is more novel, describing and advocating ways to reduce noise as a feature of organizational and public policy (in contrast to the focus on individual psychology in [1]). Forming small groups and using the averaged judgments is one natural method. This has been extensively studied, in contexts such as geopolitical forecasting, and in Chapter 21 the book describes work of Philip E. Tetlock [6], which emphasizes recording an individual's predictions to quantify accuracy. This allows one to identify *superforecasters*, individuals whose predictions are substantially more accurate than those of others. In Chapter 22, striking examples are given of divergent medical assessments of the same patients (angiograms, laparoscopies, chest X-rays for tuberculosis, biopsies for melanoma, mammograms), and the authors recommend more use of explicit guidelines in diagnostics rather than pure individual judgment. In Chapter 24, it is noted that academic studies have long highlighted the ineffectiveness of the traditional job interview as a key ingredient in hiring. A more elaborate procedure used by the Google company is described (p. 308, edited) as follows:

> The interviewers' task in structured behavioral interviews is not to decide whether they like a candidate overall; it is to collect data about each assessment in the evaluation structure and to assign scores. Interviewers are required to ask predetermined questions and score them against a predetermined rating scale, whose rubric gives examples of average, good, or great answers.

My own institution has recently started requiring such "predetermined questions" as part of the faculty hiring process, and my impression is that existing faculty members dislike it as objectionably formulaic. In Chapter 25, the authors suggest their own elaborate six-step structure for decision-making in organizations.

The final part, Part VI, which might prompt general intellectual discussion, concerns possible negative effects of replacing judgment by rules in order to reduce noise. There is a spectrum from explicit rules to standards, such as "unreasonable behavior" in the physical world or "violent or dehumanizing speech" on Facebook, and of course, such standards are inevitably open to very differing interpretations by humans. So the authors generally prefer rules to standards. In Chapter 28, it is noted that in the context of algorithms, there has been extensive worry in the *Weapons of Math Destruction* genre [3] about bias from training samples of past outcomes that might themselves be biased. And there are psychological issues on both sides. On one side, the authors comment that people in positions of authority do not like to have their discretion taken away. Not surprising! On the other side, a person applying for a loan or a job wants to be treated as an individual with an opportunity to make a case, not to receive a mechanical yes/no response. This is part of a "dignity" issue, discussed in Chapter 27. The authors acknowledge such concerns but say they must be compared with the fairness and efficiency of rule-based methods.

> "This book comments that we tend to pay attention to possible bias—one judge may be lenient and another harsh—but do not pay enough attention to noise."

In conclusion, ironically (or perhaps not?), the amazon.com reviews of this book show a huge range of opinions, from it being overly simplistic and repetitive to overly dense and academic. Many readers were disappointed, expecting the brisk easy-to-read style of Kahneman's earlier book [1]. The initial popular science style for the curious individual reader is later transformed into an implicit appeal to organizations to change their ways. This juxtaposition is inevitably awkward and somewhat repetitive, and the actual intended audience is unclear. For minimalists, the book's own final fifteen-page Review and Conclusion would suffice, if one were willing to admit example-free arguments as evidence. However, as a teacher, I say you can't repeat central points too often, and you can't have too many examples. For a creative instructor in statistics or data science, the book could complement technical textbook material by allowing discussion of the conceptual

issues (algorithms versus human judgment) or the study of the original research articles cited. Or as an undergraduate project one could make a toy model of (say) three contractors quoting prices for a construction job whose cost to the contractor is uncertain; is it true that if the lowest price wins the job, then the expectation of profit is greatest for the contractor with least noise?

References

[1] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[2] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton, 2004.

[3] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2017.

[4] Matthew J. Salganik et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117 (2020), 8398–8403.

[5] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

[6] Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

David Aldous, Department of Statistics, 367 Evans Hall, Berkeley, CA 94720, USA. E-mail: aldousdj@berkeley.edu