Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Mind as magic eight ball: A review of Kahneman, Sibony, and Sunstein's *Noise: A flaw in human judgment*

Little, Brown, Spark, 2021, 464 pp., $14.00 (hardback), ISBN 0316451401

Michael Brownstein

Routledge
Taylor & Francis Group

BOOK REVIEW

Mind as magic eight ball: A review of Kahneman, Sibony, and Sunstein's *Noise: A flaw in human judgment*, Little, Brown, Spark, 2021, 464 pp., $14.00 (hardback), ISBN 0316451401

Different doctors make different judgments about whether the same patient has breast cancer, tuberculosis, depression, and many other illnesses. Some case managers in child protective service agencies are far more likely than others to remove a child from her family. Asylum decisions made by judges are so random that they amount to "refugee roulette." A great deal of psychiatric diagnoses are about as reliable as you'd expect a dart-throwing chimp to be. So too it goes – apparently – with hiring, bail decisions, forensic science, insurance adjusters, patent applications, financial forecasts, and more. All of this illustrates the problem of *noise*, according to Daniel Kahneman, Olivier Sibony, & Cass Sunstein. Noise is randomness in our judgments and decisions. It's a basic, pervasive, dangerous, and too often unnoticed problem.

Noise can be thought of as a successor to Kahneman's *Thinking Fast and Slow* and the biases and heuristics research program on which it's built. The basic formula is that error = bias + noise. A doctor who is more likely to diagnose breast cancer in white women than black women – assuming there are no race-based differences in the actual prevalence of breast cancer – is biased. There is a pattern in this doctor's diagnoses, leading to a predictable error. A doctor whose diagnoses are scattered – who over- and underdiagnoses patients randomly – is making noisy decisions. Most doctors' (and other people's) errors are due to a combination of bias and noise.

A thought experiment from the book illustrates the difference. You're looking at four targets on a shooting range. On target A, the holes are clumped together in the center. On target B, the holes are clumped together in the lower left corner. On target C, the holes are scattered all over. And on target D, the holes are scattered from top to bottom but are all on the right side. Your job is to diagnose the performance of the shooter for each target.

This is easy if you're looking at the front of the targets. Target A's shooter is accurate. Her shots cluster around the bullseye.

Target B's shooter is biased. Her shots consistently err low and to the left. The doctor who is biased against black women makes errors of this kind. She deviates systematically from the goal of making accurate diagnoses in a consistent and predictable way.

Target C's shooter is noisy. Her shots are all over the place. Analogously, a noisy doctor might give different diagnoses to the same patient on different days. If several doctors examined the same patient and arrived at different conclusions, their collective judgment would be noisy.

Target D's shooter is both biased and noisy. She has a rightward bias but is noisy in terms of shooting high and low. Our target D doctor analogue might be reliably biased against black women, but noisy with respect to types of cancer diagnosis, some days "identifying" too many lobular carcinomas and other days too few.

Now imagine you are only allowed to look at the backside of the targets. A crucial difference between bias and noise, the authors argue, is that you can only diagnose bias when you know what counts as "correct" (i.e., when you're looking at the front of the target). Targets A and B are indistinguishable in terms of accuracy if you don't know where the marksman is aiming. But you *can* diagnose noise without knowing where the bullseye is. You can see that the shots on targets C and (to a lesser extent) D are random. Just so, you can tell that something has gone wrong when different doctors treat the same patient differently, or different judges assign wildly different penalties for the same crime. You don't need to know the true diagnosis or the correct sentence to see the error.

Noise explores in depth the sources of and potential solutions to this kind of problem. The authors distinguish "level noise," for example, which refers to variability in average judgments, from "pattern noise," which describes the tendency for a particular cue to induce randomness in one's judgments. Suppose you and I are both asked to rate, on a scale of 1–10, how happy we think a mutual friend is. What you take an "8" to mean and what I take an "8" to mean might be different. This would produce different averages between us; this is what Kahneman and colleagues mean by level noise. Now imagine the mutual friend is from Cleveland, and I just happen to believe that people from Cleveland are always unhappy. This idiosyncratic cue would drive my ratings down in this specific case, giving rise to pattern noise. There are two kinds of pattern noise: "stable pattern noise" and "occasion noise." If, over time, my belief about Clevelanders is steady, then it would produce stable pattern noise. But if my belief that Clevelanders are unhappy is ephemeral – maybe I just happened to see a movie about depressed Clevelanders and the idea stuck in my mind on the day that I filled out the ratings scale – then it would count as occasion noise. More common causes of occasion noise than movies about depressed Clevelanders are bad weather, time of day, and one's mood.

Much of the research about noisy decision-making focuses on experts and authority figures – doctors, judges, wealth managers, fingerprint analysts, and so on. A book about how depressingly noisy their judgments can be is certainly valuable. We ought to be on the lookout for false prophets, from stock pickers who parlay a string of good luck into a reputation for financial future-telling to forensic fingerprinting experts who send innocent people to prison. On this front, the book's discussion of psychiatry is devastating.

Noise is more difficult to see and explain than bias. The latter has a kind of "explanatory charisma" which noise lacks. The prejudiced doctor I described above makes predictably bad decisions; the badness is directional and patterned. Because noise is unpatterned, its specific manifestations are difficult to predict. Bias is also easier to be outraged by, while noise feels merely unfortunate. But as the authors show in detail, the costs of noise are manifold – financial, epistemic, moral – and they often outweigh the costs of bias.

Another virtue of *Noise* is its many specific and detailed suggestions, summarized with six principles (371–4) for better "decision hygiene." Here are some specific examples: because comparative judgments are more sensitive than absolute judgments, professors ought to rank students' papers from best to worst, rather than read and grade them one by one. For similar reasons, to avoid scaling without an anchor, jury deliberations about punitive damages ought to be made with information about damages awarded in relevantly similar cases. The law disallows this, for reasons the authors describe as "psychological nonsense." Some forms of decision hygiene are simple and well-known within certain fields, like using checklists to guide medical decision-making and ensuring forecasters make independent judgments before aggregating their predictions. Such strategies may not be widely known, though, and might be transferable to new contexts. *Noise* is pitched to firms and organizations, so most of the outlined strategies focus on specific group-based decisions, such as hiring, whether to accept a deal to acquire a company, etc. A set of appendices detail how companies and other groups can perform "noise audits" on their decision-making processes. Not all these suggestions are useful if you happen not to work in the corporate world, but some are. Professors in search of better processes for evaluating candidates for tenure and promotion, for example, would be well-served to consider Chapter 25′s "mediating assessments protocol." Indeed, I've suggested as much at my university, and been told the exact nonsense Kahneman and colleagues have been hearing for decades: "our process works just fine," "we know a good candidate when we see one," etc.

Noise is a long book. Despite its rich depth, the book's central goal wasn't always clear to me. It's part treatise on the sources of error in judgment, part manual for improving professional judgment in organizational contexts, part call to arms to distrust intuition and learn to love algorithms, and part introduction to statistics. I finished the book a little suspicious that the press's editor couldn't stand up to these titans in the field and force them to kill some of their darlings.

The introduction to statistics is sometimes terrific and sometimes maddening. I benefitted from *Noise's* lucid discussion of confidence intervals, for example. On the other hand, the authors sometimes slip back into academicese, like when defining multiple regression: "this technique . . . 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" (113). I also found myself a little cranky about the neologisms the authors create for the varieties of noise. I struggled to keep track of which was which, in part because the authors introduce them several times, each time a little bit differently. It wasn't obvious to me as well that their neologisms are any more intuitive than their extant scholarly names (e.g., is test-retest reliability really a worse name than "occasion noise?").

Given the book's length, I found it surprising – maybe even shocking – to find no discussion of replication. It's not that every trade book about science needs to talk about p hacking and file drawers and so on, but a book built around claims that judges are more lenient when it's hot outside, that doctors prescribe more opiates at the end of a long day and fewer cancer screenings in the afternoon than in the morning, and that college admissions officers pay more attention to applicants' academic credentials on cloudy days, ought to. I have no idea if any or all of these assertions are true. But *Noise* didn't give me confidence that they were especially well-vetted. References to correlational and posthoc observational studies aren't distinguished from studies with more robust experimental designs. A great many of the studies cited in the book come from pre-replication crisis days as well, and the lack of discussion of how we might temper our inferences based on them is sorely missing.

Another thing missing from the book – not shocking for sure, and somewhat par for the course in psychological science – was discussion of the cultural context of noise. I have little trouble believing that human judgment and decision-making are always noisy, to some degree. But always in the same amount? Are there perhaps relatively noisy cultures? The lack of discussion of historical and cultural context in *Noise* is particularly striking given current cultural politics surrounding expertise in the United States and other Western democracies. As Michael Lewis has put it so well, never have experts been so good at their jobs.[1] Weather forecasters, NBA referees, doctors, forensic scientists, and so on are *vastly* better today than they were 50 years ago at predicting the weather, spotting the foul, accurately diagnosing cancer, and distinguishing arson from accidents. And yet, compared with a half-decade ago, our distrust of these experts is at an all-time high. Compared with before, these experts have virtually eliminated noise from their judgment, and we hate them for it!

Why? One possibility is that expertise today requires the very kind of statistical thinking *Noise* explains, and most of us just don't have the tools to understand probabilistic reasoning. (Nate Silver's "failure" to forecast Donald Trump's 2016 Presidential victory is a case in point; it was only a failure if you think 3-in-10 odds never land.) On this interpretation, *Noise* is a welcome effort to improve our collective statistical reasoning. But the takeaway from *Noise* is not that experts are vastly better today than they used to be; the takeaway is that expert judgment is way worse than you think. So not only did I want more context – how does noise vary cross-culturally? how has it changed over time? – but I found myself wondering about *Noise's* timing. What does it mean, and what does it do, to write an ahistorical book about the failures of expert judgment today, in this historical moment?

Another question I had was whether accuracy is always the goal of judgment, as the authors have it. Kahneman and colleagues are careful to define their terms. They mean "judgment" in a narrow sense, as a verifiable conclusion of thinking, not as a synonym for thinking itself. But it seems to me there are times when judgment even in this narrow sense might aim at something other than accuracy. A silly but telling example is the anecdote the authors tell about George Lucas during the production of *Return of the Jedi*. Lucas' collaborator Lawrence Kasdan apparently recommended killing off Luke, explaining, "the movie has more emotional weight if someone you love is lost along the way." Lucas rejected the advice: "I don't like that and I don't believe that." This is an example, according to the authors, of "conclusion bias," or forming beliefs based on what you like. Lucas might have made the right call by not killing Luke, but he was lucky, they say, because he displayed an error in reasoning. To which I thought: *come on!* Artists make judgments about what to do on the basis of what they like all the time, and thank goodness they do. I'm no expert on the subject, but aesthetic judgment does not seem to me to aim at accuracy. Creativity seems tied to thinking and acting in unpredictable, noisy ways.

More serious examples of a kind of ballooning of the value of accuracy are ubiquitous in *Noise*, depending on how you look at it. Consider bail reform. The authors show convincingly that machine learning outperforms human judges in predicting who is a flight risk while awaiting trial. The point of the example is to illustrate one of the many ways that diminishing noise serves the cause of justice. Now, suppose for the sake of argument that you believe the bail *system* is inherently unjust, as it privileges wealthy people who can more easily make bail. If you think keeping people in jail because they're poor is fundamentally unjust, you might wonder why need to use an algorithm here at all. You might even worry that a more accurate bail system perpetuates injustice, by making the practice more palatable. The point isn't that *Noise's* analysis is wrong; noise does add to the flaws of the bail system. But one wonders whether this is a case of everything looking like a nail when you have a hammer. Accuracy is a big hammer, but some injustices aren't well-smashed by it.

The authors take a bold stand about algorithms. They write, "much of this book might be taken as an argument for greater reliance on algorithms" (334). I happen to agree with a lot of what the authors say about the advantages of algorithmic decision-making over human judgment, but they discuss virtually none of the voluminous literature exploring challenges for ethical AI and aligning algorithmic decision-making with what we want it to do. The authors do take the problem of biased algorithms seriously – e.g., recidivism prediction algorithms that use zip codes as proxies for race – but their response to this worry is merely promissory. Sophisticated algorithms can be both more accurate and less biased, they assure us, and the question is just how we choose to use them. This may be true in the sense in which it is true of all technologies, from nuclear fission to social media. We *could* use them for good, if only we choose to. But this will reassure nobody. Machine learning that is powerful enough to discover new medical drugs is equally powerful to discover new biochemical weapons (Urbina et al., 2022), and little progress on difficult questions about algorithmic governance is made by saying we can have it all if only we choose to use it well.

There is much to be learned from reading *Noise*. Oddly, though, it is a rather noisy book. A good bit too much of some ideas, and too little of others, obscure the signal. Perhaps the authors did too much of their writing on hot days. Or was that cloudy days?

Note

1. See especially Lewis (2019, 2021) and his podcast "Against the Rules." From his early books *Liar's Poker* and *The Big Short* to these more recent titles, Lewis explores the revolution in statistical thinking that is very much indebted to Kahneman and Amos Tversky's research. This origin story is the subject of Lewis' intellectual biography of Kahneman and Tversky, *The Undoing Project* (2016).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Lewis, M. (2016). *The undoing project*. W.W. Norton.
Lewis, M. (2019). *The fifth risk*. W.W. Norton.
Lewis, M. (2021). *The premonition*. W.W. Norton.
Urbina, F., Lentzos, F., Invernizzi, C., & Elkins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Michael Brownstein
Departments of Philosophy, John Jay College of Criminal Justice and The Graduate Center, CUNY, New York, NY, USA
msbrownstein@gmail.com

https://doi.org/10.1080/09515089.2022.2093708