Gerd Gigerenzer *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*

Penguin, 2023, 311pp. ISBN 978-0-141-99504-5

On the face of things, this book is timely. Writing for a general audience, the widely published and distinguished author Gerd Gigerenzer offers a *tour d'horizon* of the good, the bad and the ugly impacts of information technology (IT) on modern life. The author's aim is well-summarized in his conclusion. "We should be able to look at digital technology with level-headed admiration rather than unwarranted awe or suspicion" (p. 247). This sounds commendable.

Open the book at any random page, however, and what you'll find is an ample measure suspicion and criticism (often warranted, to be sure), with only a light sprinkling of level-headed admiration. Indeed, once we get between the covers, *How to Stay Smart in a Smart World* reads pretty much like a laundry list of bad things about IT in general and AI in particular.

This may well be useful today for the book's target general reader, who the author rewards with a sack-full of interesting and sometimes surprising nuggets. To make a lasting impact, however, it needs a clear argument that distinguishes social implications of the maturing IT revolution from the additional implications of the nascent artificial intelligence (AI) revolution flagged in the book's subtitle. This is disappointing, since a book for a general audience aiming with relentless focus on novel and distinctive implications of AI algorithms would be very welcome. Gigerenzer does deal with these at various points in the book, but too often muddies the water with digressions, for the most part reasonable enough, into what is wrong with the new IT economy – much of which is a carry-over from the old pre-IT economy.

This is nicely illustrated in the treatment of his first example, computer dating and matchmaking, which is clearly chosen to draw in his general reader. Back in the day when computers were giant machines in air-conditioned rooms, fed punched cards and mostly tended by research scientists, spy agencies and meteorologists – *long* before the IT revolution – there were dating agencies. Their matchmaking technology was a database comprising a heap of paper files, shuffled around according to the intuition of the agency's human staff. Now we have widely-used computer dating agencies such as Match.com, OK Cupid and eHarmony. Their matchmaking technology involves proprietary "black box" computational algorithms manipulating giant subscriber databases. Advances

in IT have powered a massive increase in the size of the database of potential matches, which migrated from a steel filing cabinet to a cloud server. One proprietary black box matchmaking algorithm (the intuition of agency staff) was replaced by another (based on rudimentary AI). But what has *fundamentally* changed – other than scale, price and social acceptability?

Gigerenzer tells us that computer dating agencies report success rates in a way designed to make potential customers much more optimistic than they should be about their actual chance of finding a mate, given the hard objective probabilities. Welcome to the world of advertising, where inflated and untrue claims were if anything more extreme in the nineteenth century. He tells us that none of the agencies' "scientific" AI dating algorithms "have been shown to be reliable and valid by standard scientific methods" (p. 22). Bad, but precisely the same thing could be said of the intuitive algorithms used by old-school dating agency staff. There is nothing particularly new here. What the author is telling us, and he is not wrong, is that people trying to sell us this fancy new high-tech product as something new and exciting are engaging in exactly the same type of shenanigans as their predecessors in the world of old technology … they just have more powerful tools to brag about.

Another example of a digression that is not inherently to do with AI concerns informed consent. "When we login in online, we enter the age of uninformed consent. At the same time, courts hold us responsible, whether or not we have read the terms and conditions" (p. 146). This happens because nearly all of us click "agree" to accept privacy policies which "appear to be deliberately written in a way that dissuades the few who try to read them from continuing" (p. 146). True and bad, but being bamboozled into agreeing to long lists of fine print terms and conditions we have neither the time nor the inclination to read is also nothing new. Anyone who has rented a car at an airport after a long flight knows this only too well. The difference, of course, is we are now giving away our data when we accept those terms and conditions. However, as the author tells us ruefully (p. 165), most of us are not prepared to pay even a dollar a month to stop that from happening.

The book is much stronger when Gigerenzer hammers home the point that the human brain is vastly more sophisticated, complex, powerful …. and, moreover, more energy efficient … than the most sophisticated, complex, powerful and energy hungry AI system. This is what is promised in the book's subtitle, and should ideally have been its primary, relentless focus.

The argument concerns humans' *understanding* of the problems they confront – particularly when this involves decision making under uncertainty. We should not be intimidated by the fact that algorithms can now beat the world's best human players at Chess and Go. These are completely closed systems with perfectly defined rules and no uncertainty. For any given position, there is a finite number of moves an opponent can make, a finite number of responses to this, and so on. These are games that can be won with sheer processing power and a good algorithm.

Commented [LJ(1)]: What's 'Go'? https://en.wikipedia.org/wiki/Go_(game)

Starting in the cradle, however, humans build mental models of the world that help them understand it in ways that, as yet, machines can't replicate. Strikingly, these mental models help human brains learn quickly on the basis of a few experiences – as opposed to slowly on the basis of massive amounts of training data. Crucially, they allow humans to deal with completely new situations which they have never previously encountered. In contrast, AI algorithms do not think in the least bit like humans. They don't actually think at all in any human sense.

Decision makers routinely face the problem of dealing with "known unknowns" (aka *risk*, for example betting on a fair roulette wheel), which AI algorithms can handle. Much more problematic for AI based on machine learning, our world is also riddled with "unknown unknowns" (aka *uncertainty*, for example figuring out whether that child bouncing a ball on the sidewalk is about to step out in front of my car). Because the AI algorithms have no "understanding" of the situation they are "analyzing", especially when that situation involves uncertainty, they can easily go wrong in settings they never previously encountered.

Gigerenzer usefully characterizes this distinction in terms of what he calls the *stable-word principle*. "Complex algorithms work best in well-defined, stable, situations where large amounts of data are available. Human intelligence has evolved to deal with uncertainty …" (p. 39). Viewed in this way much of what is now called "artificial intelligence", deep learning powered by neural networks, is not in any way based on an ability to reason, which is characteristic of human intelligence. It is essentially turbocharged data analysis. "Brute computational power is high-speed calculation, not intelligence" (p. 44).

This is very important because it is often impossible to reverse engineer *why* a deep-learning algorithm has made some particular decision – why you were denied bail when someone who looks just like you was not. And if we can't give *reasons* for a decision, we can't justify its *reasonableness* or

fairness. The author does touch briefly on this matter at various points, but would have done well to have developed an extended discussion of what is one of the core problems with AI based on machine learning.

Gigerenzer's plausible answer to the question of how to bring modern computational fire-power to bear on decision-making under uncertainty is what he calls "psychological AI". This is based on decision making "heuristics" akin to, sometimes the same as, those used by humans to wrangle the need to make decisions in an uncertain world. "After all, humans evolved heuristics to deal with uncertainty, and psychological AI aims to program these heuristics into a computer" (p. 44). As it happens, heavy duty heuristic-based computational modeling is already well developed, though it is usually called "agent-based modeling" not psychological AI. The book makes no mention whatsoever of this rapidly developing field, despite the fact that high powered agent-based models have been deployed on many important social problems – for example the spread of a virus in a human population.

The author does touch on the notion of psychological AI and heuristic modeling at various points in his narrative. However, despite the book's subtitle, he never develops a sustained argument about it. This lack of a sustained argument (as opposed to a problem recital) extends to a failure to distinguish what is completely new about using impenetrable AI algorithms, from what is in effect a turbocharged carry-over from a previous era. This requires making a clearer distinction between three types of problem.

The first concerns errors. Large language models (LLMs) "hallucinate". Self-driving cars crash and kill pedestrians. Facial recognition system misidentify people. An algorithm trained to distinguish between Russian and US tanks fails catastrophically in the wild because, in its training data, most Russian tanks were photographed on a cloudy day and most US tanks in the sun. The algorithm "learned" to distinguish Russian tanks from American by the presence of clouds (pp. 89-90), then on cloudy days identified US tanks as Russian, with potentially disastrous consequences if used for actual targeting. This is the "Russian Tank Fallacy" (pp. 89-91). Such mistakes are for the most part the typical teething problems faced by any new technology on its initial steep upward development curve. In a few years' time, many of them are likely to have been fixed.

The second type of problem concerns biases. Current AI algorithms systematically favor white people over people of color when it comes to suggested bail conditions or sentencing in the criminal justice system, or to hiring in the job market. Some of these biases are discoverable, and result from conscious or unconscious biases in the people who design the algorithms and select the training data. If so, they are fixable with better algorithms and better training data. Of course directly analogous biases have long been observed when real humans make exactly the same judgments about sentencing or hiring. These are also potentially resolvable by better procedures and training. As far I am aware, we don't yet know in such settings *whether human or AI-driven biases are more fixable* in the long run. And *that* is an important question when we think about the future of AI.

Commented [JL2]: So is unconscious bias in AI a product of its maker who may also suffer from unconscious or perhaps conscious bias? See addition above

Other biases of AI systems may be deeply hidden inside in a black box algorithm and its training data. The result can be, without our even realizing this, that the algorithm does things we explicitly don't want it to do. Examples of this are obviously hard to find because, of their nature, we don't know where to look for them – though the Russian Tank Fallacy is a clear example. This is a deeper and more difficult problem.

The third problem concerns ethics. Which decisions *should* be farmed out to a machine and which reserved for walking and talking humans? A problem with self-driving cars, discussed by Gigerenzer, illustrates this. A young child suddenly runs out into the road in front of a self-driving car as it approaches a line of elderly people on the sidewalk waiting for a bus. Does the car swerve and hit the bus queue or not swerve and hit the child? This a classic ethical problem and of course is not in any way "caused" by or peculiar to AI. A human driver faces *precisely* the same decision.

Some human drivers are slow to react and kill the child before they can do anything about it. Others react instinctively in the split second they have to make the choice, swerving or not. There is an ethical choice to be made, but no time for reflection as to what that choice ought to be. Whatever the outcome, we would likely treat a sober driver as morally blameless in such circumstances and the deaths as the result of a tragic accident.

But what of the AI algorithm? The algorithm's designers will (hopefully) have anticipated such a situation, have developed some rules for dealing with it, and have built these rules into the algorithm's code. Whether these rules cause the car to kill the child or the pedestrians, the fact that the death was *caused by an explicit calculation programmed by the algorithm designer*, rather than by an instinctive

human reaction, raises ethical issues which are new and deep and difficult, although Gigerenzer never digs into these. The same choice has been made in the same situation, whether instinctively by a human or following explicit calculations by an algorithm. We may feel ethically queasy about those explicit calculations on whose life to save, but what *should* the algorithm designer do?

In conclusion, this book deals with an extraordinarily important topic, and informs readers with what are essentially a series of engaging anecdotes. This is no small achievement. It's also something of a missed opportunity. There are old, enduring problems that the new AI hasn't solved, and sometimes exacerbates. And there are completely new problems arising from the rapidly expanding use of impenetrable AI algorithms to make important decisions which affect us all. It would have been good to draw these distinctions more carefully and discussed them more extensively.

Reviewer*:* Michael Laver is Visiting Professor, Department of Methodology, London School of Economics and Emeritus Professor of Politics at New York University. Among his books are *The Politics of Private Desires* (1981) and *The Governance Cycle in Parliamentary Democracies* (2023).

Email: michael.laver@nyu.edu