MITSloan
Management Review

SPRING 2022
ISSUE

# What Humans Lose When We Let AI Decide

Why you should start worrying about artificial intelligence now.

Christine Moser
Frank den Hond
Dirk Lindebaum

Vol. 63, No. 3 • Reprint #63307 • sloanreview.mit.edu

[DECISION MAKING]

# What Humans Lose When We Let AI Decide

Why you should start worrying about artificial intelligence now.

BY CHRISTINE MOSER, FRANK DEN HOND, AND DIRK LINDEBAUM

YAREK WASZUL/THEISPOT.COM

It's been more than 50 years since HAL, the malevolent computer in the movie *2001: A Space Odyssey*, first terrified audiences by turning against the astronauts he was supposed to protect. That cinematic moment captures what many of us still fear in AI: that it may gain superhuman powers and subjugate us. But instead of worrying about futuristic sci-fi nightmares, we should instead wake up to an equally alarming scenario that is unfolding before our eyes: We are increasingly, unsuspectingly yet willingly, abdicating our power to make decisions based on our own judgment, including our moral convictions. What we believe is "right" risks becoming no longer a question of ethics but simply what the "correct" result of a mathematical calculation is.

Day to day, computers already make many decisions for us, and on the surface, they seem to be doing a good job. In business, AI systems execute financial transactions and help HR departments assess job applicants. In our private lives, we rely on personalized recommendations when shopping online, monitor our physical health with wearable devices, and live in homes equipped with "smart" technologies that control our lighting, climate, entertainment systems, and appliances.

Unfortunately, a closer look at how we are using AI systems today suggests that we may be wrong in assuming that their growing power is mostly for the good. While much of the current critique of AI is still framed by science fiction dystopias, the way it is being used now is increasingly dangerous. That's not because Google and Alexa are breaking bad but because we now rely on machines to make decisions for us and thereby increasingly substitute data-driven calculations for human judgment. This risks changing our morality in fundamental, perhaps irreversible, ways, as we argued in our recent essay in *Academy of Management Learning & Education* (which we've drawn on for this article).[1]

When we employ judgment, our decisions take into account the social and historical context and different possible outcomes, with the aim, as philosopher John Dewey wrote, "to carry an incomplete situation to its fulfilment."[2] Judgment relies not only on reasoning but also, and importantly so, on capacities such as imagination, reflection, examination, valuation, and empathy. Therefore, it has an intrinsic moral dimension.

Algorithmic systems, in contrast, output decisions after processing data through an accumulation of calculus, computation, and rule-driven rationality — what we call *reckoning*.[3] The problem is that having processed our data, the answers these systems give are constrained by the narrow objectives for which they were designed, without regard for potentially harmful consequences that violate our moral standards of justice and fairness. We've seen this in the error-prone, racially biased predictive analytics used by many American judges in sentencing.[4]

And in the Netherlands, some 40,000 families unfairly suffered profound financial harm and other damages due to the tax authorities' reliance on a flawed AI system to identify potential fraudulent use of a child benefit tax-relief program. The ensuing scandal forced the Dutch government to resign in January 2021.

## Mind the Gap

These kinds of unintended consequences should come as no surprise. Algorithms are nothing more than "precise recipes that specify the exact sequence of steps required to solve a problem," as one definition puts it.[5] The precision and speed with which the problem is solved makes it easy to accept the algorithm's answer as authoritative, particularly as we interact with more and more such systems designed to learn and then act autonomously.

But as the examples above suggest, this gap between our inflated expectation of what an algorithm can do and its actual capabilities can be dangerous. Although computers can now learn independently, draw inferences from environmental stimuli, and then act on what they have learned, they have limits. The common interpretation of these limits as "bias" is missing the point, because to call it bias implies that there is the possibility that the limits of AI reckoning can be overcome, that AI will eventually be able to eliminate bias and render a complete and truthful representation of reality. But that is not the case. The problem lies deeper.

AI systems base their decisions on what philosopher Vilém Flusser called "technical images": the abstract patterns of pixels or bytes — the digital images of the "world" that the computer has been trained to produce and process. A technical image is the *computed* transformation of digitalized data about some object or idea, and, as such, it is a *representation* of the world. However, as with any representation, it is incomplete; it makes discrete what is continuous and renders static what is fluid, and therefore one must learn how to use it intelligently. Like a two-dimensional map, it is a selective representation, perhaps even a misrepresentation. During the making of the map, a multidimensional reality is reduced by forcing it to fit onto a two-dimensional surface with a limited number of colors and symbols. Many of us still know how to use a map. The trouble with AI systems is that we cannot fully understand how the machine "draws" its technical images — what it emphasizes, what it omits, and how it connects pieces of information in transitions toward the technical image. As statistician George Box once noted, "All models are wrong, but some are useful." We add, "If we don't know *how* a model is wrong, it is not only useless, but using it can be dangerous." This is because reckoning based on a technical image can never reveal the full truth, simply because this is not what it has been made for.

Through a process of reckoning, the machines make decisions — which are essentially based on technical images — that have an air of precision, and of being indisputably correct. The risk is that we end up fashioning ourselves and our society based on the image that the technology has formed of us. If we rely too much on these algorithms, we risk mistaking the map for the territory, like the unfortunate motorists in Marseille, France, who let their GPS direct them straight off a quayside street and into the harbor.

Consider, for instance, the way people use wearable electronic devices that monitor bodily functions, including pulse rate, steps taken, temperature, and hours of sleep, as indicators of health. Instead of asking yourself how you feel, you can check your wearable. It may tell you that you should be concerned because you don't take the minimum number of steps generally recommended for a healthy life — a target that may make sense for many but could be counterproductive if the air quality is bad or you have weak lungs.

How can we capture the undeniable benefits of these smart machines without mistaking the machine's abstraction for the whole story and delegating to AI the complicated and consequential decisions that *should* be made by humans?

The problem is less the machines than we ourselves. Our trust in AI leads us to confuse reckoning — decision-making based on the summing up of various kinds of data and technical images — with judgment. Too much faith in the machine — and in our ability to program and control that machine — can produce untenable, and perhaps even irreversible, situations. We see(k) in AI's confident answers the kind of certainty that our ancestors sought in vain in entrails, tarot cards, or the stars above.

Some observers — scholars, managers, policy makers — believe that following responsible AI development principles is a valid and effective approach to injecting ethical considerations into AI systems. We

Through a process of reckoning, the machines make decisions that have an air of precision, and of being indisputably correct. The risk is that we end up fashioning ourselves and our society based on the image that the technology has formed of us.

**What Humans Lose When We Let AI Decide** (Continued from page 13)

agree when they argue that as a cultural product, AI is bound to reflect the outlooks of people who commissioned the algorithm, wrote its code, and then used the program. We disagree when they say that therefore careful attention to the overall project and its programming is all that is needed to keep the results mostly positive. Critical voices about our ability to instill ethics into AI are increasing, asking the basic question of whether we *can* "teach" ethics to AI systems in the first place. Going back to the roots of AI, we should realize that the very fundaments of judgment and reckoning are different and cannot be reconciled. This means that AI systems will never be capable of judgment, only of reckoning. Any attempt to inject ethics into AI therefore means that the ethics will be straightjacketed, distorted, and impoverished to fit the characteristics of AI's reckoning.

Those who believe in ethical AI consider the technology to be a tool, on par with other tools that are essentially extensions of the human body, such as lenses (to extend the view of the eyes), or spoons and screwdrivers (to extend the dexterity of the hands). But we are skeptical. In our view, as with wearables, it's crucial to consider the user as part of the tool: We need to think hard about how algorithms shape us. If the substitution of data-driven reckoning for human judgment is something to be very cautious about, then following the rallying cry to develop responsible AI will not help us break the substitution. What can we do?

## A Call to Action (and Inaction)

We see two ways forward.

First, we want to make a call for *in*action: We need to stop trying to automate everything that can be automated just because it is technically feasible. This trend is what Evgeny Morozov describes as *technological solutionism*: finding solutions to problems that are not really problems at all.[6] Mesmerized by the promise of eternal improvement, technological solutionism blunts our ability to think deeply about whether, when, how, and why to use a given tool.

> In the absence of vigilance and doubt, we might miss the moment when our decision-making frame has transitioned from judgment to reckoning. This would mean that we would hand over our power to make moral decisions to the machine's reckoning.

Second, we also make a call to action. As a society, we need to learn what AI really is and how to work with it. Specifically, we need to learn how to understand and use its technical images, just as we needed to learn how to work with models and maps — not just in a general, abstract, and idealized sense, but for each specific project where it is envisioned.

In practice, that means that managers should judge on a case-by-case basis whether, how, and why to use AI. Avoiding blanket applications of the technology means taking AI's limits seriously. For example, while algorithms gain greater predictive power when fed more data, they will always assume a static model of society, when, in fact, time and context are ever changing. And once managers have judged AI to be a useful tool for a given project, we see essential merit in managers developing and maintaining an attitude of vigilance and doubt. This is because, in the absence of vigilance and doubt, we might miss the moment when our decision-making frame has transitioned from judgment to reckoning. This would mean that we would hand over our power to make moral decisions to the machine's reckoning.

Given the rapid advances of AI, there isn't much time. We all must relearn how to make decisions informed by our own judgment rather than let ourselves be lulled by the false assurance of algorithmic reckoning.

***Christine Moser*** *(@tineadam) is an associate professor of organization theory at Vrije Universiteit Amsterdam in the Netherlands.* ***Frank den Hond*** *is the Ehrnrooth Professor in Management and Organisation at the Hanken School of Economics in Finland and is affiliated with Vrije Universiteit Amsterdam.* ***Dirk Lindebaum*** *is a senior professor in organization and management at Grenoble Ecole de Management. Comment on this article at https://sloanreview.mit.edu/x/63307.*

### REFERENCES

**1.** C. Moser, F. den Hond, and D. Lindebaum, "Morality in the Age of Artificially Intelligent Algorithms," Academy of Management Learning & Education, April 7, 2021, https://journals.aom.org.

**2.** J. Dewey, "Essays in Experimental Logic" (Chicago: University of Chicago Press, 1916), 362.

**3.** B.C. Smith, " The Promise of Artificial Intelligence: Reckoning and Judgment" (Cambridge, Massachusetts: MIT Press, 2019).

**4.** K.B. Forrest, "When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence" (Singapore: World Scientific Publishing, 2021).

**5.** J. MacCormick, "Nine Algorithms That Changed the Future" (Princeton, New Jersey: Princeton University Press, 2012), 3.

**6.** E. Morozov, "To Save Everything, Click Here: The Folly of Technological Solutionism" (New York: PublicAffairs, 2013).

**Reprint 63307.**