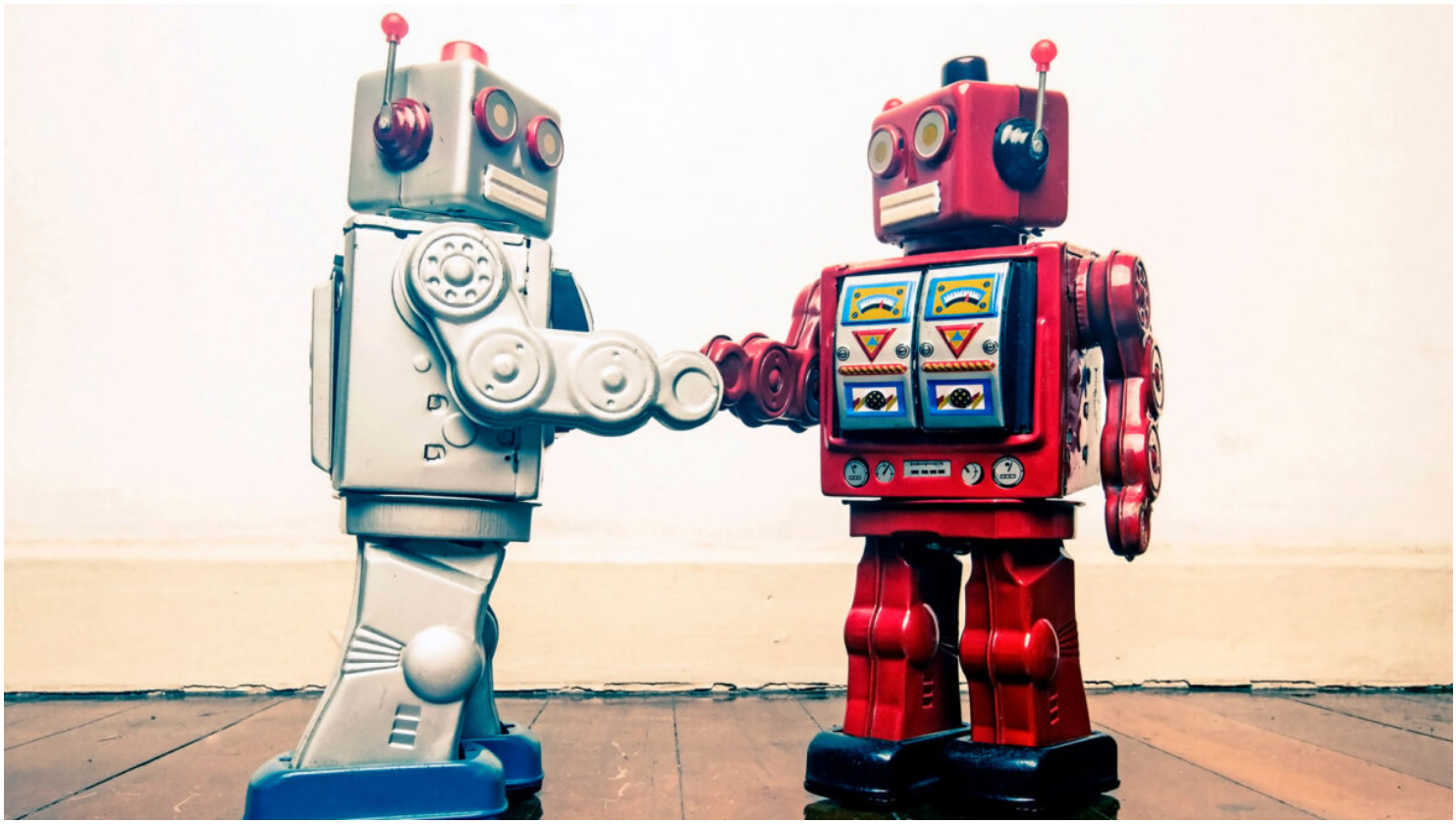
INNOVATION

What Does Building a Fair AI Really Entail?

by David De Cremer

September 03, 2020



Charles Taylor/EyeEm/Getty Images

Artificial intelligence (AI) is rapidly becoming integral to how organizations are run. This should not be a surprise; when analyzing sales calls and market trends, for example, the judgments of computational algorithms can be considered superior to those of humans. As a result, AI techniques are increasingly used to make decisions. Organizations are employing algorithms to allocate valuable resources, design work schedules, analyze employee performance, and even decide whether employees can stay on the job.

This creates a new set of problems even as it solves old ones. As algorithmic decision-making's role in calculating the distribution of limited resources increases, and as humans become more dependent on and vulnerable to the decisions of AI, anxieties about fairness are rising. How unbiased can an automated decision-making process with humans as the recipients really be?

To address this issue, computer scientists and engineers are focusing primarily on how to govern the use of data provided to help the algorithm learn (that is, data mining) and how to use guiding principles and techniques that can promote *interpretable AI:* systems that allow us to understand how the results emerged. Both approaches rely, for the most part, on the development of computational methods that factor in certain features believed to be related to fairness.

At the heart of the problem is the fact that algorithms calculate optimal models from the data they're given — meaning they can end up replicating the problems they're meant to correct. A 2014 effort to remove human bias in recruitment at Amazon, for example, rated candidates in gender-biased ways; the historical job performance data it was given showed that the tech industry was dominated by men, so it assessed hiring men to be a good bet. The Correctional Offender Management Profiling for Alternative Sanctions, an AI-run program, offered biased predictions for recidivism that wrongly forecast that Black defendants (incorrectly judged to be at higher risk of recidivism) would reoffend at a much greater rate than white defendants (incorrectly flagged as low-risk).

Organizations and governments have tried to establish guidelines to help AI developers refine technical aspects so that algorithmic decisions will be more interpretable — allowing humans to understand clearly how decisions were reached — and thus fairer. For example, Microsoft has launched programs that identify high-level principles such as fairness, transparency, accountability, and ethicality to guide computer scientists and engineers in their coding efforts. Similar efforts are underway on the government level, as demonstrated by the European Union's Ethics Guidelines for Trustworthy AI and Singapore's Model AI Governance Framework.

But neither the efforts of computer scientists to factor in technological features nor the efforts of companies and governments to develop principle-based guidelines quite solves the issue of trust. To do that, designers need to account for the information needs and expectations of the people facing the results of the models' outputs. This is important ethically and also practically: An abundance of research in management shows that the fairer decisions are perceived to be, the more that employees accept them, cooperate with others, are satisfied with their jobs, and perform better. Fairness matters greatly to organizational functioning, and there's no reason to think that will change when AI becomes the decision maker.

So, how can businesses that want to implement AI persuade users that they're not compromising on fairness? Put simply, they need to stop thinking about fairness — a complicated concept — as something they can address with the right automated processes and start thinking about an interdisciplinary approach in which computer and social sciences work together. Fairness is a social construct that humans use to coordinate their interactions and subsequent contributions to the collective good, and it is subjective. An AI decision maker should be evaluated on how well it helps people connect and cooperate; people will consider not only its technical aspects but also the social forces operating around it. An interdisciplinary approach allows for identifying three types of solutions that are usually not discussed in the context of AI as a fair decision maker.

Solution 1: Treat AI fairness as a cooperative act.

Algorithms aim to reduce error rates as much as possible in order to reveal the optimal solution. But while that process can be shaped by formal criteria of fairness, algorithms leave the *perceptual* nature of fairness out of the equation and do not cover aspects such as whether people feel they have been treated with dignity and respect and have been taken care of — important justice concerns. Indeed, algorithms are largely designed to create optimal prediction models that factor in technical features to enhance formal fairness criteria, such as interpretability and transparency, despite the fact that those features do not necessarily meet the expectations and needs of the human end user. As a result, and as the Amazon example shows, algorithms may predict outcomes that society perceives as unfair.

There's a simple way to address this problem: The model produced by AI should be evaluated by a human devil's advocate. Although people are much less rational than machines and are to some extent blind to their own inappropriate behaviors, research shows that they are less likely to be biased when evaluating the behaviors and decisions of others. In view of this insight, the strategy for achieving AI fairness must involve a cooperative act between AI and humans. Both parties can bring their best abilities to the table to create an optimal prediction model adjusted for social norms.

Recommendation: Organizations need to invest significantly in the ethical development of their managers. Being a devil's advocate for algorithmic decision makers requires managers to develop their common sense and intuitive feel for what is right and wrong.

Solution 2: Regard AI fairness as a negotiation between utility and humanity.

Algorithmic judgment is demonstrated to be more accurate and predictive than human judgment in a range of specific tasks, including the allocation of jobs and rewards on the basis of performance evaluations. It makes sense that in the search for a better-functioning business, algorithms are increasingly preferred over humans for those tasks. From a statistical point of view, that preference may appear valid. However, managing workflow and resource allocation in (almost) perfectly rational and consistent ways is not necessarily the same as building a humane company or society.

No matter how you may try to optimize their workdays, humans don't work in steady, predictable ways. We have good and bad days, afternoon slumps, and bursts of productivity — all of which presents a challenge for the automated organization of the future. Indeed, if we want to use AI in ways that promote a humane work setting, we have to accept the proposition that we should not optimize the search for utility to the detriment of values such as tolerance for failure, which allows people to learn and improve — leadership abilities considered necessary to making our organizations and society humane. The optimal prediction model of fairness should be designed with a negotiation mindset that strives for an acceptable compromise between utility and humane values.

Recommendation: Leaders need to be clear about what values the company wants to pursue and what moral norms they would like to see at work. They must therefore be clear about *how* they want to do business and *why*. Answering those questions will make evident the kind of organization they would like to see in action.

Solution 3: Remember that AI fairness involves perceptions of responsibility.

Fairness is an important concern in most (if not all) of our professional interactions and therefore constitutes an important responsibility for decision makers. So far, organizations and governments — because of their adherence to matrix structures — have tackled the question of fair AI decision-making by developing checklists of qualities to guide the development of algorithms. The goal is to build AIs whose outputs match a certain definition of what's fair.

That's only half of the equation, however: AI's fairness as a decision maker really depends on the choices made by the organization adopting it, which is responsible for the outcomes its algorithms generate. The perceived fairness of the AI will be judged through the lens of the organization employing it, not just by the technical qualities of the algorithms.

Recommendation: An organization's data scientists need to know and agree with the values and moral norms leadership has established. At most organizations, a gap exists between what data scientists are building and the values and business outcomes organizational leaders want to achieve. The two groups need to work together to understand what values cannot be sacrificed in the use of algorithms. For example, if the inclusiveness of minority groups, which are usually poorly represented in available data, is important to the company, then algorithms need to be developed that include that value as an important filter and ensure that outliers, not just commonalities, are learned from.

***

Organizations need to recognize that their stakeholders will perceive them, not the algorithms they deploy, as responsible for any unfair outcomes that may emerge. What makes for AI fairness will then also be a function of how fair stakeholders perceive the company generally to be. Research has shown that fairness perceptions, in addition to distributive fairness that algorithms have mastered to some extent, may entail how fairly the organization treats its employees and customers, whether it communicates in transparent ways, and whether it is regarded as respectful toward the community at large. Organizations adopting AI to participate in decision-making are advised to put the necessary time and energy into building the right work culture: one with a trustworthy and fair organizational image.

David De Cremer is a Provost's chair and professor in management and organizations at NUS Business School, National University of Singapore. He is the founder and director of the Center on AI Technology for Humankind at NUS Business school. Before moving to NUS, he was the KPMG endowed chaired professor in management studies at Judge Business School, University of Cambridge. He is named one of the World's top 30 management gurus and speakers in 2020 by the organization GlobalGurus and recently published the book *"Leadership by algorithm: Who leads and who follows in the AI era?"*

This article is about INNOVATION

⊕ Follow This Topic

Related Topics: Managing Organizations | Technology

Comments

Leave a Comment

Post Comment

1 COMMENTS

SARAH ALT 10 days ago

Very good contribution and valuable for leaders wondering where they start when it comes to their ethical AI responsibilities. Thank you.

Reply 0 0

Join The Conversation

POSTING GUIDELINES

We hope the conversations that take place on HBR.org will be energetic, constructive, and thought-provoking. To comment, readers must sign in or register. And to ensure the quality of the discussion, our moderating team will review all comments and may edit them for clarity, length, and relevance. Comments that are overly promotional, mean-spirited, or off-topic may be deleted per the moderators' judgment. All postings become the property of Harvard Business Publishing.