





## If you worry about humanity, you should be more scared of humans than of AI

Moran Cerf  and Adam Waytz 

### ABSTRACT

Advances in artificial intelligence (AI) have prompted extensive and public concerns about this technology's capacity to contribute to the spread of misinformation, algorithmic bias, and cybersecurity breaches and to pose, potentially, existential threats to humanity. We suggest that although these threats are both real and important to address, the heightened attention to AI's harms has distracted from human beings' outsized role in perpetuating these same harms. We suggest the need to recalibrate standards for judging the dangers of AI in terms of their risks relative to those of human beings. Further, we suggest that, if anything, AI can *aid* human beings in decision making aimed at improving social equality, safety, productivity, and mitigating some existential threats.

### KEYWORDS

Artificial intelligence; existential risk; algorithmic bias; ethics; cybersecurity; nuclear decision making

The question of whether artificial intelligence (AI) poses an existential risk has received increased attention of late, with many sounding the alarm on AI's imminent threat. For example, the Future of Life Institute recently published an open letter<sup>1</sup> calling for a pause on AI research and development, and the Center for AI Safety posted an open statement<sup>2</sup> comparing the threat posed by AI to that of nuclear bombs and suggesting drastic measures to reign in the technology. These letters received wide public attention, partially because their signatories include notable technology proponents and leaders of prominent artificial intelligence-based companies.

A cynic might suggest that these public warnings serve as good PR for the technology, calling attention to the potential dangers while also signaling how remarkable and useful it is (“We built a technology so powerful that we even worry it might be *too* good and require safeguards!”) and helping the creators shape government regulation concerning future uses.

Here, we offer a less cynical but still noteworthy concern, which is that these outsized warnings about technology's existential threats serve as a red herring. Although the fears raised around AI's capacity to spread misinformation, foster unemployment, and outpace human intelligence are well founded (and we strongly advocate for taking these risks seriously), we worry these public letters distract from human beings' current proficiency at carrying out the threats attributed to technology. In reality, humans are the clear and present risk that is underscored by the AI advances.

We realize that this view requires clarity. Typical discourse asks people to take a simple binary position.

(“Are you on the side of more regulation of AI, or the side that says it is far from being a real threat?”) We argue that AI can become a modern-day imminent danger, yet that at this point it is actually the best tool to *mitigate* a far bigger threat to humanity: human decision-making. Currently, to protect the world from large-scale threats (climate change, pandemics, nuclear war, etc.), we believe the best approach involves humans working *with* AI to improve decision-making in domains as critical as those concerning life and death.

As one example, take the spread of misinformation, which the Future of Life Institute letter highlights in asking, “*Should* we let machines flood our information channels with propaganda and untruth?” Undoubtedly the spread of misinformation by AI-propagated systems is concerning, especially given the unparalleled scale of content that AI can generate. But as recent research reveals, humans are far more responsible for spreading misinformation than technology. In a study of how true and false news spreads on Twitter, researchers analyzed 126,000 stories tweeted by millions of people between 2006 and 2017 and found that false news spreads faster than true news, and that “false news spreads more than the truth because humans, not robots, are more likely to spread it” (Vosoughi, Roy, and Aral 2018). In fact, some notable signatories of the letter have themselves contributed to the spread of false conspiracy theories<sup>3</sup> and misleading information.<sup>4</sup>

A threat even more dire than misinformation is the “risk of extinction from AI” that the Center for AI Safety highlights in its open statement. Yet, in terms of whether machines or humans are more likely to initiate

extinction-level events such as nuclear war, humans still seem to have the upper hand. In recent empirical work that analyzes the decision processes employed by senior leaders in war-game scenarios involving weapons of mass destruction, humans showed an alarming tendency to err on the side of initiating catastrophic attacks.<sup>5</sup> These simulations, if implemented in reality, would pose much graver risks to humanity than machine-driven ones. Our exploration of the use of AI in critical decision-making has shown AI's superiority to human decisions in nearly all scenarios. In most cases, the AI makes the choice that humans do not make at first—but then, upon more careful consideration and deliberation, change their minds and do make, realizing it was the correct decision all along.

Other, more quotidian concerns raised about AI apply far more to human beings than to machines. Consider algorithmic bias, the phenomenon whereby algorithms involved in hiring decisions, medical diagnoses, or image detection produce outcomes that unfairly disadvantage a particular social group. For example, when Amazon implemented an algorithmic recruiting tool to score new applicants' resumes, the algorithm systematically rated female applicants worse than men, in large part because the algorithm was trained on resumes submitted over the previous 10 years that were disproportionately male.<sup>6</sup> In other

words, an algorithm trained on human bias will reproduce this bias.

Unlike humans, however, algorithmic bias can be readily deprogrammed, or as economist Sendhil Mullainathan puts it, “Biased algorithms are easier to fix than biased people.”<sup>7</sup> Mullainathan and colleagues' research showed that an algorithm used by UnitedHealth to score patients' health risks systematically underscored black patients relative to white patients because it measured illness in terms of health-care costs (which are systematically lower for black versus white individuals, given that society spends less on black patients) (Obermeyer et al. 2019). However, once identified, the researchers could easily modify this feature of the algorithm to produce risk scores that were relatively unbiased. Other work has shown that algorithms can produce less racially biased outcomes (and more effective public safety outcomes) than human judges in terms of decisions about whether or not to grant bail to defendants awaiting trial (Kleinberg et al. 2018). As biased as algorithms can be, their biases appear less ingrained and more pliable than those of humans. Compounded by recent work showing that, in hiring and lending contexts, managers reject biased algorithms in favor of more biased humans, the suggestion that humans should remain at the helm of those functions is, at best, questionable (Cowgill, Dell'acqua, and Matz 2020).

#### Where machines are better than humans

As a sobering reminder of the human-AI risk comparison, we highlight several domains where current machine intelligence seems already to challenge the performance of humans:

With regard to **traffic safety**, while much attention is given to every accident perpetuated by autonomous cars, the reality is that reports from the National Highway Traffic Safety Administration (See: <https://www.nhtsa.gov/press-releases/traffic-crash-death-estimates-2022>) and General Services Administration (See: <https://drivethru.gsa.gov/DRIVERSAFETY/DistractedDrivingPosterA.pdf>) suggest that out of over six million accidents annually (with 42,939 fatal incidents), 98 percent are due to human error, and self-driving cars are estimated to reduce this proportion by 76 percent (See: <https://web-assets.bcg.com/36/39/e80d073a4067bfe89c7482d6db69/the-european-aftermarket-in-2030.pdf>).

Similarly, in the domain of **medical diagnosis**, a meta-analysis (See: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6716335/>) of articles published across 20 years of research shows that in various domains (e.g., brain tumors) machine performance is increasingly becoming superior to that of human doctors.

Recently, AI has won competitions **for creativity** in art and advertising, surpassing human performance in **art authentication** (See: <https://www.ippi.org/ai-a-new-frontier-in-art-authentication/>) and, in legal contexts, correcting wrongful convictions (See: <https://www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time>) made by humans (resulting from false identification) and shortening trial times by over 20 percent.

Finally, it is noteworthy that current research by the corresponding author investigates the possibility of using “*digital twin*”—a reasoned and composed machine-based decision tool that replicates the key stakeholder's thinking under minimally biased conditions—to aid leaders in choices related to critical decisions (namely, nuclear and climate-related critical decisions).

Finally, consider the threat to cybersecurity. Although commentators have warned<sup>8,9,10</sup> that large language models add tools to the arsenals of hackers by democratizing cybercrime, most high-profile information leaks and hacks to date are ushered in by human beings with no reliance on AI (i.e. a disgruntled employee who knows the system's flaws and perpetrates an attack by remembering key passwords, or bad programmers who effectively enable future attacks by making wrong assumptions on their software uses—such as “no one would create a password that is 1,000,000 characters long” leading to a classical *buffer overflow* hack). In fact, AI is often the last bastion of *defense* against those hacks, identifying complex human coding mistakes early-on and correcting them.

Recently, national guardsman Jack Teixeira, who exposed highly classified material in an online chat group, did not require sophisticated technology to access sensitive documents—he was granted top secret clearance from the Pentagon. Further, a recent study conducted by IBM indicates that 95 percent of security breaches were caused by human errors such as biting on phishing scams or downloading malware.<sup>11</sup> If anything, the most concerning cybersecurity risk currently posed by AI results from its increased reliance on human trained code, which is flawed. AI takes hackable human codes and uses them to generate new codes, spreading these human-generated errors further. The only concerning current cybersecurity *attacks* by AI involve AI that simulates human communication to dupe humans into revealing key information. Cybersecurity may represent a case in which technology is more likely to be the solution rather than the problem, with research indicating, for example, that humans working *with* AI outperform humans alone in detecting machine-manipulated media such as deepfakes (Groh et al. 2021).

Even when technology contributes to unwanted outcomes, humans are often the ones pressing the buttons. Consider the effect of AI on unemployment. The Future of Life Institute letter raises concerns that AI will eliminate jobs, yet whether or not to eliminate jobs is a choice that humans ultimately make. Just because AI *can* perform the jobs of, say, customer service representatives does not mean that companies *should* outsource these jobs to bots. In fact, research indicates that many customers would prefer to talk to a human than to a bot, even if it means waiting in a queue.<sup>12</sup> Along similar lines, increasingly common statements that AI-based systems—like “the Internet,” “social media,” or the set of interconnected online functions referred to as “The Algorithm”—are destroying mental health,<sup>13</sup> causing political polarization,<sup>14</sup> or threatening democracy<sup>15</sup> neglect an obvious fact: These systems are populated

and run by human beings. Blaming technology lets people off the hook.

Although expressions of concern toward AI are invaluable in matching the excitement around new technology with caution, outsized news cycles around the threats of technology can distract from the threats of human beings. Recent research indicates that humans have a “finite pool of attention” such that “when we pay more attention to one threat, our attention to other threats decreases” (Cisco et al. 2023). So, as we contend with the rise of AI and its concomitant harms to privacy, human survival, and our relationship with truth itself, we must equally pay attention to the humans who are already well equipped to perpetrate these harms without the assistance of machines. Specifically, it has not escaped our notice that when engaging in a conversation about the risks of AI, the benchmark is often “is AI *perfect* in handling this task” (making critical decisions or guiding a self-driving car), rather than “is it *better* than humans.” The answer to the latter question in many cases, is that yes, AI can mitigate the risks to humanity.

## Notes

1. See: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
2. See: <https://www.safe.ai/statement-on-ai-risk>.
3. See: <https://www.nytimes.com/2022/10/30/business/musk-tweets-hillary-clinton-pelosi-husband.html>.
4. See: <https://www.forbes.com/sites/kenrickcai/2023/06/04/stable-diffusion-emad-mostaque-stability-ai-exaggeration/?sh=2bd38c3075c5>.
5. See: <https://www.ft.com/content/06b22337-e862-43e5-8440-d9c225e0c18d>.
6. See: <https://www.bbc.com/news/technology-45809919>.
7. See: <https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>.
8. See: <https://hbr.org/2023/04/the-new-risks-chatgpt-poses-to-cybersecurity>.
9. See: <https://www.forbes.com/sites/tonybradley/2023/02/27/defending-against-generative-ai-cyber-threats/?sh=c62032c10884>.
10. See: <https://sloanreview.mit.edu/article/from-chatgpt-to-hackgpt-meeting-the-cybersecurity-threat-of-generative-ai/>.
11. See: <https://thehackernews.com/2021/02/why-human-error-is-1-cyber-security.html>.
12. See: <https://www.userlike.com/en/blog/consumer-chat-bot-perceptions>.
13. See: <https://nypost.com/2023/02/14/the-internet-is-ruining-teens-cdc-report-is-the-latest-proof/>.
14. See: <https://www.brookings.edu/articles/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/>.
15. See: <https://www.theatlantic.com/ideas/archive/2022/07/social-media-harm-facebook-meta-response/670975/>.

## Disclosure statement

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

## Funding

Moran Cerf received funding from the Carnegie Corporation of New York (Grant ID: G-19-57248). Adam Waytz received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Notes on contributors

*Moran Cerf* is a neuroscientist and professor of business at Columbia University and a former cybersecurity expert. As a recipient of the Carnegie fellowship, he works on the applications of neuroscience and AI in nuclear decision making.

*Adam Waytz* is a professor of management and organizations at the Kellogg School of Management at Northwestern University and has consulted with Google on its chatbot, Bard.

## ORCID

Moran Cerf  <http://orcid.org/0000-0002-2012-3177>

Adam Waytz  <http://orcid.org/0000-0001-9706-6062>

## References

- Cowgill, B., F. Dell'acqua, and S. Matz. 2020. "The Managerial Effects of Algorithmic Fairness Activism." *AEA Papers & Proceedings* 110: 85–90. <https://doi.org/10.1257/pandp.20201035>.
- Groh, M., Z. Epstein, C. Firestone, and R. Picard. 2021. "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds." *PNAS* 119 (1). <https://doi.org/10.1073/pnas.2110013119>.
- Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–293. February. <https://doi.org/10.1093/qje/qjx032>.
- Obermeyer, Z., Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science: Advanced Materials and Devices* 366 (6464): 447–453. October 25. <https://www.science.org/doi/pdf/10.1126/science.aax2342>.
- Sisco, M. R., S. M. Constantino, Yu Gao, M. Tavoni, A. D. Cooperman, V. Bosetti, E. U. Weber, et al. 2023. "Examining Evidence for the Finite Pool of Worry and Finite Pool of Attention Hypotheses." *Global Environmental Change* 78: 102622. <https://www.sciencedirect.com/science/article/abs/pii/S0959378022001601?via%3Dihub>.
- Vosoughi, S., D. Roy, and S. Aral. 2018. "The Spread of True and False News Online." *Science: Advanced Materials and Devices* 359 (6380): 1146–1151. <https://www.science.org/doi/10.1126/science.aap9559>.