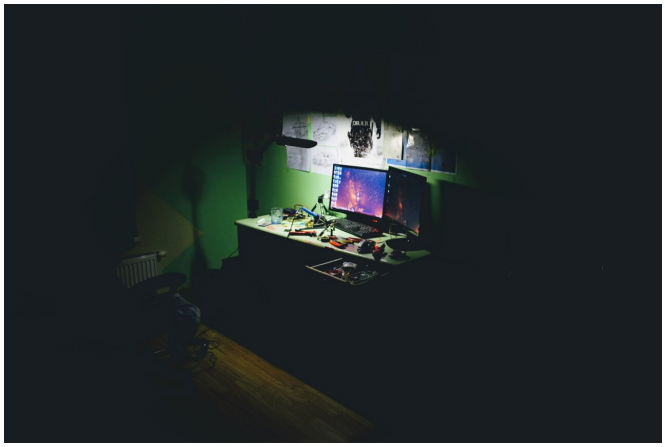
Community of ethical hackers needed to prevent AI's looming 'crisis of trust', experts argue

December 9 2021



Credit: Unsplash/CC0 Public Domain

The Artificial Intelligence industry should create a global community of hackers and "threat modelers" dedicated to stress-testing the harm potential of new AI products in order to earn the trust of governments and the public before it's too late.

This is one of the recommendations made by an international team of risk and machine-learning experts, led by researchers at the University of Cambridge's Center for the Study of Existential Risk (CSER), who have authored a new "call to action" published today in the journal *Science*.

They say that companies building intelligent technologies should harness techniques such as "red team" hacking, audit trails and "bias bounties"—paying out rewards for revealing ethical flaws—to prove their integrity before releasing AI for use on the wider public.

Otherwise, the industry faces a "crisis of trust" in the systems that increasingly underpin our society, as public concern continues to mount over everything from driverless cars and autonomous drones to secret social media algorithms that spread misinformation and provoke political turmoil.

The novelty and "black box" nature of AI systems, and ferocious competition in the race to the marketplace, has hindered development and adoption of auditing or third party analysis, according to lead author Dr. Shahar Avin of CSER.

The experts argue that incentives to increase trustworthiness should not be limited to regulation, but must also come from within an industry yet to fully comprehend that public trust is vital for its own future—and trust is fraying.

The new publication puts forward a series of "concrete" measures that they say should be adopted by AI developers.

"There are critical gaps in the processes required to create AI that has earned public trust. Some of these gaps have enabled questionable behavior that is now tarnishing the entire field," said Avin.

"We are starting to see a public backlash against technology. This 'tech-lash' can be all encompassing: either all AI is good or all AI is bad.

"Governments and the public need to be able to easily tell apart between the trustworthy, the snake-oil salesmen, and the clueless," Avin said. "Once you can do that, there is a real incentive to be trustworthy. But while you can't tell them apart, there is a lot of pressure to cut corners."

Co-author and CSER researcher Haydn Belfield said: "Most AI developers want to act responsibly and safely, but it's been unclear what concrete steps they can take until now. Our report fills in

some of these gaps."

The idea of AI "red teaming"—sometimes known as white-hat hacking—takes its cue from cyber-security.

"Red teams are ethical hackers playing the role of malign external agents," said Avin. "They would be called in to attack any new AI, or strategise on how to use it for malicious purposes, in order to reveal any weaknesses or potential for harm."

While a few big companies have internal capacity to "red team"—which comes with its own ethical conflicts—the report calls for a third-party community, one that can independently interrogate new AI and share any findings for the benefit of all developers.

A global resource could also offer high quality red teaming to the small start-up companies and research labs developing AI that could become ubiquitous.

The new report, a concise update of more detailed recommendations published by a group of 59 experts last year, also highlights the potential for bias and safety "bounties" to increase openness and public trust in AI.

This means financially rewarding any researcher who uncovers flaws in AI that have the potential to compromise public trust or safety—such as racial or socioeconomic biases in algorithms used for medical or recruitment purposes.

Earlier this year, Twitter began offering bounties to those who could identify biases in their image-cropping algorithm.

Companies would benefit from these discoveries, say researchers, and be given time to address them before they are publicly revealed. Avin points out that, currently, much of this "pushing and prodding" is done on a limited, ad-hoc basis by academics and investigative journalists.

The report also calls for auditing by trusted external agencies—and for open standards on how to document AI to make such auditing possible—along with platforms dedicated to sharing "incidents": cases of undesired AI behavior that could cause harm to humans.

These, along with meaningful consequences for failing an external audit, would significantly contribute to an "ecosystem of trust" say the researchers.

"Some may question whether our recommendations conflict with commercial interests, but other safety-critical industries, such as the automotive or pharmaceutical industry, manage it perfectly well," said Belfield.

"Lives and livelihoods are ever more reliant on AI that is closed to scrutiny, and that is a recipe for a crisis of trust. It's time for the industry to move beyond well-meaning ethical principles and implement real-world mechanisms to address this," he said.

Added Avin: "We are grateful to our collaborators who have highlighted a range of initiatives aimed at tackling these challenges, but we need policy and public support to create an ecosystem of trust for AI."

More information: Shahar Avin, Filling gaps in trustworthy development of AI, *Science* (2021). DOI: 10.1126/science.abi7176. www.science.org/doi/10.1126/science.abi7176

Provided by University of Cambridge

APA citation: Community of ethical hackers needed to prevent AI's looming 'crisis of trust', experts argue (2021, December 9) retrieved 8 December 2022 from https://techxplore.com/news/2021-12-ethical-hackers-ai-looming-crisis.html