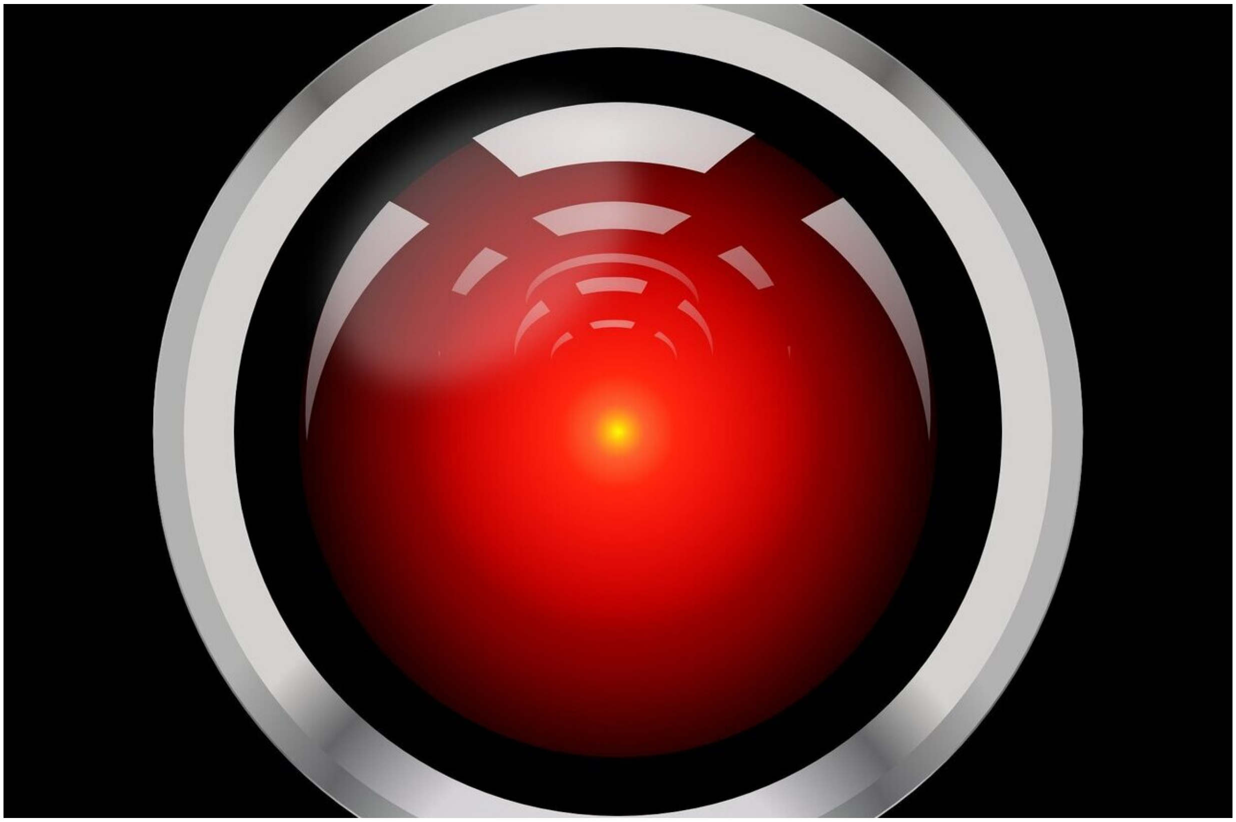
PHYS.ORG

Research explores promoting public trust in AI

March 8 2021



Credit: Pixabay/CC0 Public Domain

The public doesn't need to know how Artificial Intelligence works to trust it. They just need to know that someone with the necessary skillset is examining AI and has the authority to mete out sanctions if it causes or is likely to cause harm.

Dr. Bran Knowles, a senior lecturer in data science at Lancaster University, says: "I'm certain that the public are incapable of determining the trustworthiness of individual AIs… but we don't need them to do this. It's not their responsibility to keep AI honest."

Today (March 8) Dr. Knowles presents a research paper "The Sanction of Authority: Promoting Public Trust in AI" at the ACM Conference on Fairness, Accountability and Transparency (ACM FAccT).

The paper is co-authored by John T. Richards, of IBM's T.J. Watson Research Center, Yorktown Heights, New York.

The general public are, the paper notes, often distrustful of AI, which stems both from the way AI has been portrayed over the years and from a growing awareness that there is little meaningful oversight of it.

The authors argue that greater transparency and more accessible explanations of how AI systems work, perceived to be a means of increasing trust, do not address the public's concerns.

A 'regulatory ecosystem," they say, is the only way that AI will be meaningfully accountable to the public, earning their trust.

"The public do not routinely concern themselves with the trustworthiness of food, aviation, and pharmaceuticals because they trust there is a system which regulates these things and punishes any breach of safety protocols," says Dr. Richards.

And, adds Dr. Knowles: "Rather than asking that the public gain skills to make informed decisions about which AIs are worthy of their trust, the public needs the same guarantees that any AI they might encounter is not going to cause them harm."

She stresses the critical role of AI documentation in enabling this trustworthy regulatory ecosystem. As an example, the paper discusses work by IBM on AI Factsheets, documentation designed to capture key facts regarding an AI's development and testing.

But, while such documentation can provide information needed by internal auditors and external regulators to assess compliance with emerging frameworks for trustworthy AI, Dr. Knowles cautions against relying on it to directly foster public trust.

"If we fail to recognize that the burden to oversee trustworthiness of AI must lie with highly skilled regulators, then there's a good chance that the future of AI documentation is yet another terms and conditions-style consent mechanism—something no one really reads or understands," she says.

The paper calls for AI documentation to be properly understood as a means to empower specialists to assess trustworthiness.

"AI has material consequences in our world which affect real people; and we need genuine accountability to ensure that the AI that pervades our world is helping to make that world better," says Dr. Knowles.

More information: Bran Knowles et al. The Sanction of Authority, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021). DOI: 10.1145/3442188.3445890

Provided by Lancaster University

Citation: Research explores promoting public trust in AI (2021, March 8) retrieved 27 April 2024 from https://phys.org/news/2021-03-explores-ai.html