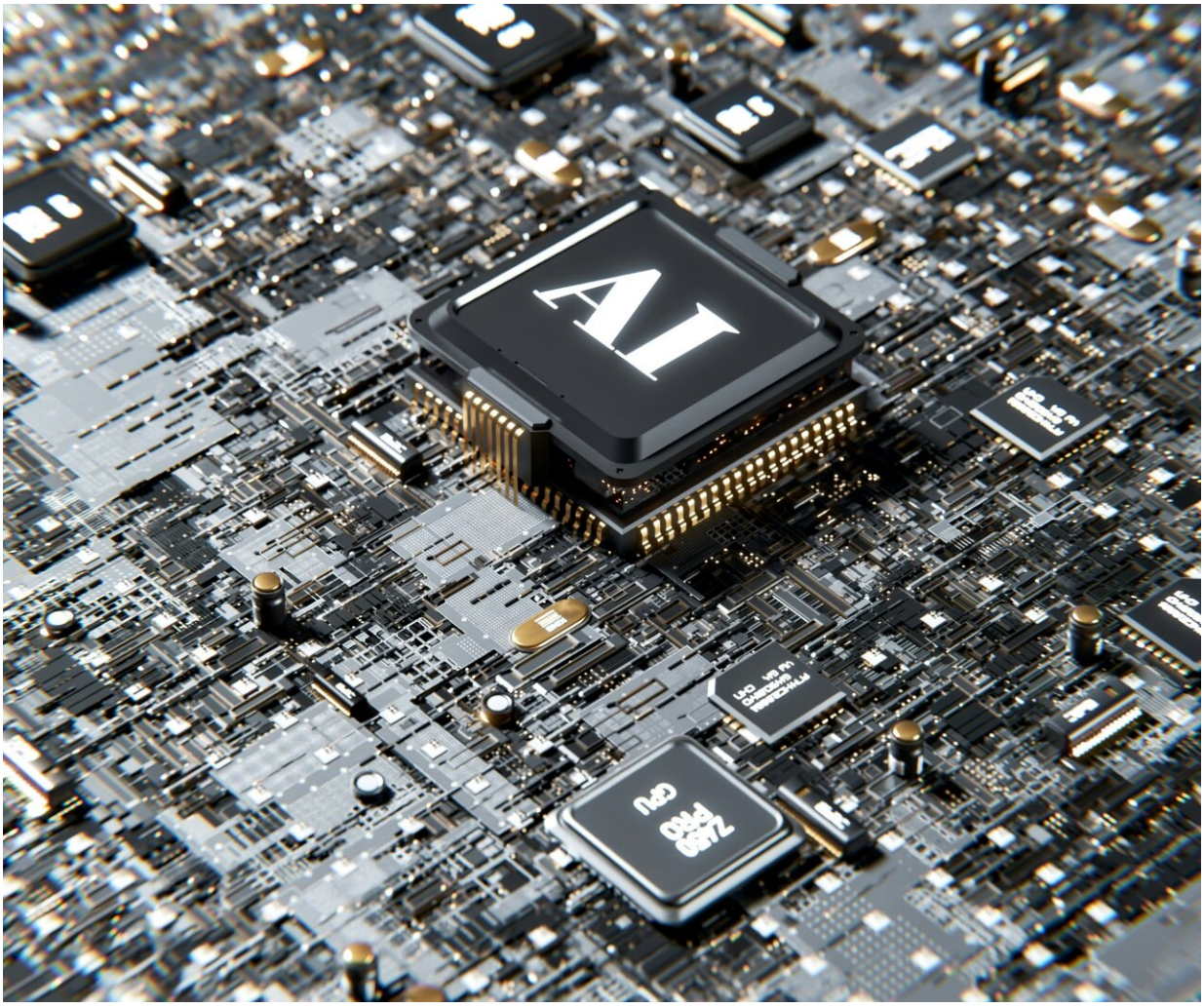
There is no proof that AI can be controlled, researcher warns

February 12 2024



Credit: Unsplash/CC0 Public Domain

There is no current evidence that AI can be controlled safely, according to an extensive review, and without proof that AI can be controlled, it should not be developed, a researcher warns.

Despite the recognition that the problem of AI control may be one of the most important problems facing humanity, it remains poorly understood, poorly defined, and poorly researched, Dr. Roman V. Yampolskiy explains.

In his book, *AI: Unexplainable, Unpredictable, Uncontrollable*, AI Safety expert Dr. Yampolskiy looks at the ways that AI has the potential to dramatically reshape society, not always to our advantage.

He explains, "We are facing an almost guaranteed event with potential to cause an existential catastrophe. No wonder many consider this to be the most important problem humanity has ever faced. The outcome could be prosperity or extinction, and the fate of the universe hangs in the balance."

Uncontrollable superintelligence

Dr. Yampolskiy has carried out an extensive review of AI scientific literature and states he has found no proof that AI can be safely controlled—and even if there are some partial controls, they would not be enough.

He explains, "Why do so many researchers assume that AI control problem is solvable? To the best of our knowledge, there is no evidence for that, no proof. Before embarking on a quest to build a controlled AI, it is important to show that the problem is solvable.

"This, combined with statistics that show the development of AI superintelligence is an almost guaranteed event, show we should be

supporting a significant AI safety effort."

He argues our ability to produce intelligent software far outstrips our ability to control or even verify it. After a comprehensive literature review, he suggests advanced intelligent systems can never be fully controllable and so will always present certain level of risk regardless of benefit they provide. He believes it should be the goal of the AI community to minimize such risk while maximizing potential benefit.

What are the obstacles?

AI (and superintelligence), differ from other programs by its ability to learn new behaviors, adjust its performance and act semi-autonomously in novel situations.

One issue with making AI 'safe' is that the possible decisions and failures by a superintelligent being as it becomes more capable is infinite, so there are an infinite number of safety issues. Simply predicting the issues not be possible and mitigating against them in security patches may not be enough.

At the same time, Yampolskiy explains, AI cannot explain what it has decided, and/or we cannot understand the explanation given as humans are not smart enough to understand the concepts implemented. If we do not understand AI's decisions and we only have a "black box," we cannot understand the problem and reduce likelihood of future accidents.

For example, AI systems are already being tasked with making decisions in health care, investing, employment, banking and security, to name a few. Such systems should be able to explain how they arrived at their decisions, particularly to show that they are bias-free.

Yampolskiy says, "If we grow accustomed to accepting AI's answers

without an explanation, essentially treating it as an Oracle system, we would not be able to tell if it begins providing wrong or manipulative answers."

Controlling the uncontrollable

As capability of AI increases, its autonomy also increases but our control over it decreases, Yampolskiy explains, and increased autonomy is synonymous with decreased safety.

For example, for superintelligence to avoid acquiring inaccurate knowledge and remove all bias from its programmers, it could ignore all such knowledge and rediscover/proof everything from scratch, but that would also remove any pro-human bias.

"Less intelligent agents (people) can't permanently control more intelligent agents (ASIs). This is not because we may fail to find a safe design for superintelligence in the vast space of all possible designs, it is because no such design is possible, it doesn't exist. Superintelligence is not rebelling, it is uncontrollable to begin with," he explains.

"Humanity is facing a choice, do we become like babies, taken care of but not in control or do we reject having a helpful guardian but remain in charge and free."

He suggests that an equilibrium point could be found at which we sacrifice some capability in return for some control, at the cost of providing a system with a certain degree of autonomy.

Aligning human values

One control suggestion is to design a machine that precisely follows

human orders, but Yampolskiy points out the potential for conflicting orders, misinterpretation or malicious use.

He says, "Humans in control can result in contradictory or explicitly malevolent orders, while AI in control means that humans are not."

If AI acted more as an advisor it could bypass issues with misinterpretation of direct orders and potential for malevolent orders, but the author argues that for AI to be a useful advisor it must have its own superior values.

"Most AI safety researchers are looking for a way to align future superintelligence to values of humanity. Value-aligned AI will be biased by definition, pro-human bias, good or bad is still a bias. The paradox of value-aligned AI is that a person explicitly ordering an AI system to do something may get a 'no' while the system tries to do what the person actually wants. Humanity is either protected or respected, but not both," he explains.

Minimizing risk

To minimize the risk of AI, he says it needs it to be modifiable with 'undo' options, limitable, transparent and easy to understand in human language.

He suggests all AI should be categorized as controllable or uncontrollable, and nothing should be taken off the table and limited moratoriums, and even partial bans on certain types of AI technology should be considered.

Instead of being discouraged, he says, "Rather it is a reason, for more people, to dig deeper and to increase effort, and funding for AI Safety and Security research. We may not ever get to 100% safe AI, but we can

make AI safer in proportion to our efforts, which is a lot better than doing nothing. We need to use this opportunity wisely."

More information: Roman V. Yampolskiy, AI, *AI: Unexplainable, Unpredictable, Uncontrollable* (2024). DOI: 10.1201/9781003440260

Provided by Taylor & Francis