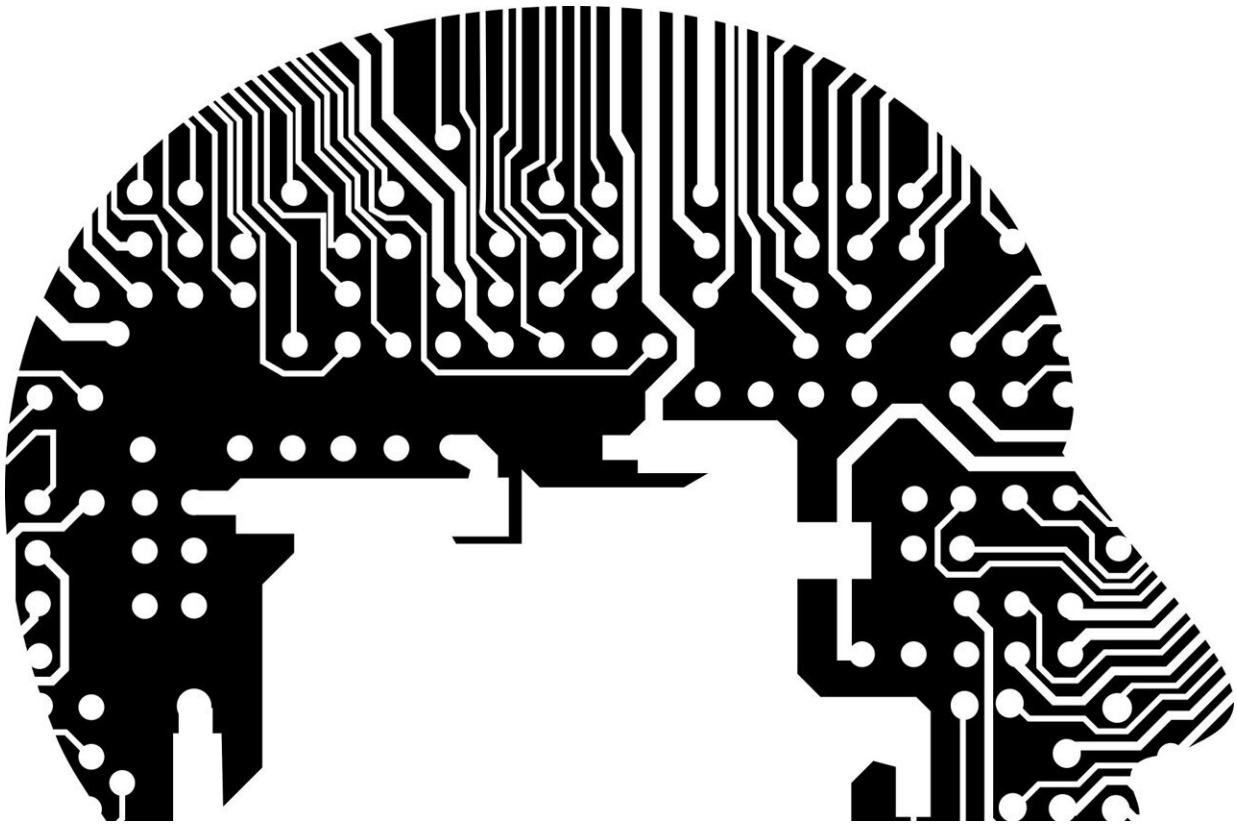
What is 'AI alignment'? Silicon Valley's favorite way to think about AI safety misses the real issues

July 13 2023, by Aaron J. Snoswell



Credit: Pixabay/CC0 Public Domain

As increasingly capable artificial intelligence (AI) systems become widespread, the question of the risks they may pose has taken on new

urgency. Governments, researchers and developers have highlighted AI safety.

The EU is moving on AI regulation, the UK is convening an AI safety summit, and Australia is seeking input on supporting safe and responsible AI.

The current wave of interest is an opportunity to address concrete AI safety issues like bias, misuse and labor exploitation. But many in Silicon Valley view safety through the speculative lens of "AI alignment", which misses out on the very real harms current AI systems can do to society—and the pragmatic ways we can address them.

What is 'AI alignment'?

"AI alignment" is about trying to make sure the behavior of AI systems matches what we *want* and what we *expect*. Alignment research tends to focus on hypothetical future AI systems, more advanced than today's technology.

It's a challenging problem because it's hard to predict how technology will develop, and also because humans aren't very good at knowing what we want—or agreeing about it.

Nevertheless, there is no shortage of alignment research. There are a host of technical and philosophical proposals with esoteric names such as "cooperative inverse reinforcement learning" and "iterated amplification".

There are two broad schools of thought. In "top-down" alignment, designers explicitly specify the values and ethical principles for AI to follow (think Asimov's three laws of robotics), while "bottom-up" efforts try to reverse-engineer human values from data, then build AI

systems aligned with those values. There are, of course, difficulties in defining "human values", deciding who chooses which values are important, and determining what happens when humans disagree.

OpenAI, the company behind the ChatGPT chatbot and the DALL-E image generator among other products, recently outlined its plans for "superalignment". This plan aims to sidestep tricky questions and align a future superintelligent AI by first building a merely human-level AI to help out with alignment research.

But to do this they must first align the alignment-research AI…

Why is alignment supposed to be so important?

Advocates of the alignment approach to AI safety say failing to "solve" AI alignment could lead to huge risks, up to and including the extinction of humanity.

Belief in these risks largely springs from the idea that "Artificial General Intelligence" (AGI)—roughly speaking, an AI system that can do anything a human can— could be developed in the near future, and could then keep improving itself without human input. In this narrative, the super-intelligent AI might then annihilate the human race, either intentionally or as a side-effect of some other project.

In much the same way the mere possibility of heaven and hell was enough to convince the philosopher Blaise Pascal to believe in God, the possibility of future super-AGI is enough to convince some groups we should devote all our efforts to "solving" AI alignment.

There are many philosophical pitfalls with this kind of reasoning. It is also very difficult to make predictions about technology.

Even leaving those concerns aside, alignment (let alone "superalignment") is a limited and inadequate way to think about safety and AI systems.

Three problems with AI alignment

First, **the concept of "alignment" is not well defined**. Alignment research typically aims at vague objectives like building "provably beneficial" systems, or "preventing human extinction".

But these goals are quite narrow. A super-intelligent AI could meet them and still do immense harm.

More importantly, **AI safety is about more than just machines and software**. Like all technology, AI is both technical and social.

Making safe AI will involve addressing a whole range of issues including the political economy of AI development, exploitative labor practices, problems with misappropriated data, and ecological impacts. We also need to be honest about the likely uses of advanced AI (such as pervasive authoritarian surveillance and social manipulation) and who will benefit along the way (entrenched technology companies).

Finally, **treating AI alignment as a technical problem puts power in the wrong place**. Technologists shouldn't be the ones deciding what risks and which values count.

The rules governing AI systems should be determined by public debate and democratic institutions.

OpenAI is making some efforts in this regard, such as consulting with users in different fields of work during the design of ChatGPT. However, we should be wary of efforts to "solve" AI safety by merely

gathering feedback from a broader pool of people, without allowing space to address bigger questions.

Another problem is a lack of diversity—ideological and demographic—among alignment researchers. Many have ties to Silicon Valley groups such as effective altruists and rationalists, and there is a lack of representation from women and other marginalized people groups who have historically been the drivers of progress in understanding the harm technology can do.

If not alignment, then what?

The impacts of technology on society can't be addressed using technology alone.

The idea of "AI alignment" positions AI companies as guardians protecting users from rogue AI, rather than the developers of AI systems that may well perpetrate harms. While safe AI is certainly a good objective, approaching this by narrowly focusing on "alignment" ignores too many pressing and potential harms.

So what is a better way to think about AI safety? As a social and technical problem to be addressed first of all by acknowledging and addressing existing harms.

This isn't to say that alignment research won't be useful, but the framing isn't helpful. And hare-brained schemes like OpenAI's "superalignment" amount to kicking the meta-ethical can one block down the road, and hoping we don't trip over it later on.

Provided by The Conversation