PAUSEAI

“If we pursue our current approach, then we will eventually lose control.”
Stuart Russell, author of the standard textbook on AI, 2023 [1]

We have been warned of the existential risk of AI by Turing Award winners Yoshua Bengio and Geoffery Hinton; authors of the standard AI textbook, Stuart Russell and Peter Norvig; inventor of generative AI, Ian Goodfellow; and many hundreds more AI professors and researchers, not to mention the CEOs of every major AI company.[2] There is not complete consensus - notable figures like Yann LeCunn do disagree - but PauseAI was founded to advocate for a level-headed approach in the face of uncertainty. Our message is simple: to ignore such a dire warning from the experts in AI would be stunningly reckless.

"Alarm bells over the latest form of artificial intelligence – generative AI – are deafening. And they are loudest from the developers who designed it." - António Guterres, UN Secretary-General [3]

Controlling AI is currently an unsolved research problem.[4] OpenAI’s official plan to prevent human extinction is to use future AI systems to solve the ‘alignment problem’.[5] DeepMind has not published any strategy.
No one has a serious plan to safeguard humanity. So we are calling on world governments to enforce a worldwide pause on the race to superintelligence.

[1] https://news.berkeley.edu/2023/04/07/stuart-russell-calls-for-new-approach-for-ai-a-civilization-ending-technology/
[2] https://www.safe.ai/statement-on-ai-risk
[3] https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/
[4] https://arxiv.org/abs/2209.00626
[5] https://openai.com/blog/our-approach-to-alignment-research

PAUSEAI

Q+A

“Dangerously powerful AI will not be created for a very long time.”

Let’s hope so - predicting the future is hard! Recent progress in AI has been astonishingly fast. Five years ago, no one expected that we would have AI as general and useful as ChatGPT-4 by 2023. AI can now code, write academic essays and answer scientific questions better than some undergraduates.[1] It’s impossible to be sure that AI won’t exceed all humans very soon.

“AI wouldn’t want to kill us.”

Humans care about other people because we evolved that way. A powerful AI will understand human ethics, but that does not mean it will choose to act in an ethical manner. Instilling ethical goals in AI is the ‘alignment problem’ which is currently an open research question. For almost any goal an AI might pursue, it would be useful to prevent humans from opposing it .[2]

“Pausing AI is impossible. Even if we tried, China would never agree.”

Maybe ¯_(ツ)_/¯. We should still try. Humanity has effectively banned or decelerated other technological projects in the past, such as human genetic modification, CFCs and geo-engineering.[3] China is currently far behind the US in the AI race and far more worried about the destabilising effects.[4]

“AIs don’t have bodies. How could they hurt us?”

The AI could design a deadly pandemic virus and order a print online. The AI could hack into infrastructure such as oil pipelines and water systems. It could use the autonomous drones currently deployed in Ukraine.[5] The smarter the AI gets, the more creative ways to take control it will think of.

[1] https://arxiv.org/abs/2303.08774
[2] https://arxiv.org/abs/2209.00626
[3] https://forum.effectivealtruism.org/posts/WfodoyjePTTuaTjLe/efficacy-of-ai-activism-have-we-ever-said-no
[4] https://www.humanetech.com/podcast/the-ai-race-china-vs-the-us-with-jeffrey-ding-and-karen-hao
[5] https://www.newscientist.com/article/2397389-ukrainian-ai-attack-drones-may-be-killing-without-human-oversight/