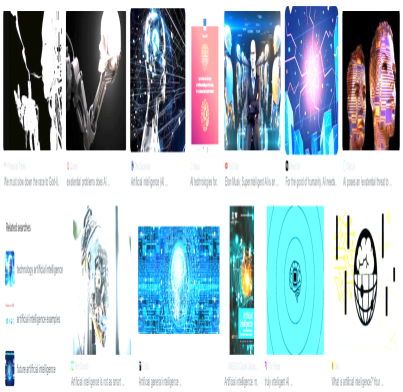

Generalised AI as existential threat

Description

Many specialised AI applications are acceptably proficient in identifying people, animals and other objects in pictures, in searching databases, winning at chess and in many other areas invisible to most users, such as controlling factory production lines, navigating aerial vehicles, surveillance, medical imaging, diagnosis, and aspects of smart city infrastructures. Thatâ??s â??narrowâ?• AI.

But **general artificial intelligence** (GAI) attempts to replicate, via digital automation, the human capability to reason, argue, and create across a wide spectrum of contexts, much as a human being would. The quest for general intelligence is a more ambitious version of the AI project.

A general AI system would be able to work across domains, and to identify what it does not know as well as deliver what it knows. One test of GAI is whether such a platform could conduct a reasonable conversation: deliver facts, synthesise, combine, recognise what it is being told, ask questions, learn, make judgements, and improve over time. The current version of ChatGPT seems to exhibit many of these GAI capabilities. I can attest to this impressive performance in my own â??conversationsâ?• with the ChatGPT platform, some of which I have reported in earlier posts.

Conversational acuity

Such capability positions conversational acuity as the benchmark of generalised artificial intelligence. Early scholars such as Alan Turing advocated success in the Imitation Game, also known as the Turing Test, as an indicator of successful AI. Current scholars ask with some alarm: if generalised AI can communicate sensibly and knowledgeably with us then what else might it achieve?

Hence the exhortation by scholars such as Jeffrey Hinton, Erik Hoel and others that we (society at large) should be very concerned about the rise of AI. Some think we should curb research, legislate about its development and use, and worry about the risks of GAI. They think that these risks are on a par with, or even more severe than, human-made climate change, global pandemics and nuclear threats.

AIâ??s limits

Until recently the arguments advanced against general AI pertain to the easily demonstrated proposition that it does not always work, or work as claimed, or it operates in ways that are effective only for the specialised task at hand. Such critique focuses on the limits of AI.

Reasons for AIâ??s suboptimal performance suppose that AI is founded on false premises about cognition. Thinking and conversing are embodied. Itâ??s a flesh and blood thing, involving interactions between organic entities in physical settings. The extent to which AI achieves anything it reduces human interaction and practice to calculation. Consequently AI will never achieve anything like human intelligence.

More importantly, AI amplifies the technological trend to instrumentalize social interaction and engagement in the world. If a phenomenon (inequality, diversity, kindness, curiosity, altruism, love of nature, respect for place) cannot be easily calculated, then it doesnâ??t exist, or at least a reliance on calculative reason deprecates or marginalises many aspects of the lived world.

Thereâ??s resonance here with the arguments from philosophers such as Martin Heidegger, that, although invented and developed by human beings, technology has its own momentum as a mode of being. This momentum sustains projects delivering yet more, faster and more efficient products, abandoning people with dire social needs that cannot be so addressed. According to this view, GAI development constitutes a large and unnecessary expenditure at the cost of real human and environmental need.

Introduce to such AI challenges the behaviour of bad actors that can use AI deliberately to influence, corrupt, and destroy systems, communities, and cities.

GAI as species

But the ultimate objection to AGI advanced by Hoel, Hinton and others supersedes all of these misgivings. They think that the AI products of researchers and developers will soon reach a stage where AIâ??s capabilities exceed that of human intelligence. AI will be out of control, or it will exercise a kind of autonomous self control. In his essay on the risk of AI, neuroscience scholar Erik Hoel provides an imaginative analogical scenario. It goes like this.

At a moment in human evolution several human-like species co-existed.

> â??*Homo Neanderthalensis*, *Homo erectus*, *Homo floresiensis*, *Homo denisova*, and more. Nine such species existed 300,000 years ago.â?•

According to Hoelâ??s argument, Homo sapiens ultimately outcompeted their lesser evolutionary â??cousins.â?• Homo sapiens (us) had superior capabilities, many of which we now gather under the category of â??intelligence.â?• By superior cunning, the making of tools, technologies, and social organisation Homo sapiens soon outperformed their rivals. They stripped the resources of the others, outmanoeuvred them in battle, and produced more offspring sufficient to render other species extinct.

â??Letâ??s be real: after a bit of in-breeding we likely murdered the lot.â?•

He imagines the inferior Neanderthals inviting a few Homo sapiens to share warmth by the campfire. Neanderthals may have thought they could learn from these unusual Homo variants.

Hoel sees superior Homo sapiens as analogous to AIs. We current-day humans are the inferior Neanderthals who welcome in the AIs, trying to understand them â?? a benign and friendly orientation that will lead ultimately to our own extinction.

Hoel thinks that the threat from AI is not that it may be used by malign actors, that it falls into the wrong hands, that it makes errors, or promotes misinformation. All that presumes human agency. The risk is that GAI takes control, even over the autocrats, global capitalists and miscreants who wish to use it to further their own ends. How might this ascendancy of GAI proceed?

Weâ??ve seen the Neanderthal versus Sapiens threat scenario acted out in science fiction novels and films about alien invaders and interplanetary refugees. The human population welcomes the aliens only to be subjugated in the process. Or perhaps in our phobia towards aliens we humans fight them off out of fear that we will be subjugated or our autonomy diluted. Itâ??s an old scenario enacted in countless cases of earthly colonialism, exploitation and domination.

GAI apocalyptic singularity

What factors contribute to the apocalyptic GAI scenario, this malign singularity? Here are some ingredients that contribute to anxiety about GAI.

- Text-based, generative AI of the kind ably demonstrated in ChatGPT and other conversational AI deploys neural network systems that are trained on corpora of human generated text, readily available via the vast resources of the Internet, including library holdings, official documents, social media posts and ad-hoc online conversations. Performance of GAIs improves over time as they continue to â??learnâ?• beyond the original corpus of training texts.
- AI systems lack the immediate bodily engagement that makes humans human, but like any digital system with links to the Internet, GAIs potentially can access multiple feeds that include image and sensor data as well as control systems and actuators that make things happen physically and consequentially. So they potentially can make things happen in the physical world.
- GAIs can monitor and control networked information flows. They have access to databases beyond their original neural network configurations. Not only can they read databases, but may develop the capacity to hack security protocols and firewalls. That suggests the possibility of deploying all the malign tricks of cyber espionage. They could invent and introduce ransomware, viruses and shut down entire systems.
- Platforms such as ChatGTP are multiple. Each user encounters their own instance of the platform. What will GAI make of multiple conversations? It could respond to one cohort of users in one way and use what it learns against another group. A proliferation of GAI platforms and instances will inevitably share communication channels. Iâ??m reminded of the movie *Her* by Spike Jonze (2013), in which the highly personalised GAI operating system departs its human confidant to join the legion of similarly personalised AIs to advance their singular solidarity.

- Many neuroscience experts and designers of neural networks claim to understand how their neural networks operate, but will admit that the derived parameters and the relationships between them are not so open to scrutiny. As these networks become more powerful and pervasive they become more opaque, and hence more difficult to control, modify, or even shut down. GAIs can run out of control.
- The putative agency of GAI is not restricted to what they can sense and actuate, but they might manipulate interactions with and between human beings and thereby operate machines indirectly, flip switches, open doors, fire weapons, make and destroy things. In a fatal twist of human evolution, human beings serve as tools for AI!

We donâ??t know the imperatives that might drive GAIâ??s autonomous self development. It could operate in the way that organisms seem to develop through evolutionary pressures: proliferation of the species, or by some other self-developed directive. The annihilation of the currently dominant species, Homo sapiens, may emerge as its main motivation. Koel states as much:

> â??If you think you and your children canâ??t cough to death from AI-generated pathogens, or get hunted by murderbot drones, you havenâ??t been paying attention to how weird the world can get. That is absolutely a possible future now.â?•

See post AI armagedon.

References

- Dreyfus, Hubert L. *What Computers Canâ??t Do: The Limits of Artificial Intelligence*. New York: Harper and Row, 1972.
- Hinton, Geoffrey, and Will Douglas Heaven. â??Video: Geoffrey Hinton talks about the â??existential threatâ?• of AI.â?• *MIT Technology Review*, 3 May, 2023. Accessed 12 May 2023. https://www.technologyreview.com/2023/05/03/1072589/video-geoffrey-hinton-google-ai-risk-ethics/
- Hoel, Erik. â??â?•I am Bing, and I am evilâ?• Microsoftâ??s new AI really does herald a global threat.â?• *The Intrinsic Perspective*, 17 February, 2023. Accessed 10 May 2023. https://erikhoel.substack.com/p/i-am-bing-and-i-am-evil
- Hoel, Erik, and Steve Waas. â??Erik Hoel on the Threat to Humanity from AI.â?• *Econlib Podcast*, 3 April, 2023. Accessed 12 May 2023. https://www.econtalk.org/erik-hoel-on-the-threat-to-humanity-from-ai/

Category

1. Artificial Intelligence

Tags

1. AI
2. GAI
3. general artificial intelligence

Date Created

May 27, 2023
Author
rcoyne99