Hope or horror? The great AI debate dividing its pioneers

Publication Date: 2023-10-24

Author: Dan Milmo

Section: Technology

Tags: Technology, The AI race, Artificial intelligence (AI), Google, Computing, analysis

Article URL: https://www.theguardian.com/technology/2023/oct/24/hope-or-horror-the-great-ai-debate-dividing-its-pioneers

Demis Hassabis says he is not in the “pessimistic” camp about artificial intelligence. But that did not stop the CEO of Google DeepMind signing a statement in May warning that the threat of extinction from AI should be treated as a societal risk comparable to pandemics or nuclear weapons. That uneasy gap between hope and horror, and the desire to bridge it, is a key reason why Rishi Sunak convened next week’s global AI safety summit in Bletchley Park, a symbolic choice as the base of the visionary codebreakers – including computing pioneer Alan Turing – who deciphered German communications during the second world war. “I am not in the pessimistic camp about AI obviously, otherwise I wouldn’t be working on it,” Hassabis tells the Guardian in an interview at Google DeepMind’s base in King’s Cross, London. “But I’m not in the ‘there’s nothing to see here and nothing to worry about’ [camp]. It’s a middle way. This can go well but we’ve got to be active about shaping that.” Hassabis, a 47-year-old Briton, co-founded UK company DeepMind in 2010. It was bought by Google in 2014 and has achieved stunning breakthroughs in AI under his leadership. The company is now known as Google DeepMind after merging with the search firm’s other AI operations, with Hassabis at the helm as CEO. His unit is behind the AlphaFold program that can predict the 3D shapes of proteins in the human body – as well as nearly all catalogued proteins known to science. This is a revolutionary achievement that will help achieve breakthroughs in areas such as discovering new medicines because it maps out the biological building blocks of life. This year Hassabis was jointly awarded one of the most prestigious prizes in science, the Lasker basic medical research award, for the work on AlphaFold. Many winners of the award go on to win a Nobel prize. Last month Hassabis’ team released AlphaMissense, which uses the same AI protein program to spot protein malformations that could cause disease. Hassabis says he would have preferred the May statement to contain references to AI’s potential benefits. “I would have had a line saying about all the incredible opportunities that AI is going to bring: medicine, science, all the things help in everyday life, assisting in everyday life.” He says AI advances will trigger “disruption” in the jobs market – skilled professions such as law, medicine and finance are at risk, according to experts – but he says the impact will be “positive overall” as the economy adapts. This has also led to talk among AI professionals of the technology funding a universal basic income or even a universal basic service, which provides services such as transport and accommodation for free. “Some kind of sharing of the upsides would be needed in some form,” says Hassabis. But the OECD, an influential international organisation, says jobs at the highest risk from AI-driven automation are highly skilled and represent about 27% of employment across its 38 member countries, which include the UK, Japan, Germany, the US, Australia and Canada. No wonder the OECD talks of an “AI revolution which could fundamentally change the

workplace". Nonetheless, the summit will focus on threats from frontier AI, the term for cutting-edge systems that could cause significant loss of life. These include the ability to make bioweapons, create sophisticated cyber-attacks and to evade human control. The latter issue refers to fears about artificial general intelligence, or "god-like" AI, meaning a system that operates with above or beyond human levels of intelligence. The pessimistic camp that voices these fears has strong credentials. Geoffrey Hinton, a British computer scientist often described as one of the "godfathers" of modern AI, quit his job at Google this year in order to voice his fears about the technology more freely. Hinton told the Guardian in May of his concerns that AI firms are trying to build intelligences with the potential to outthink humanity. "My confidence that this wasn't coming for quite a while has been shaken by the realisation that biological intelligence and digital intelligence are very different, and digital intelligence is probably much better." Stuart Russell, another senior British computer scientist, has warned of a scenario where the UN asks an AI system to create a self-multiplying catalyst to de-acidify the oceans, with the safety instruction that the outcome is non-toxic and that no fish are harmed. But the result uses up a quarter of the oxygen in the atmosphere and subjects humans to a slow and painful death. Both Hinton and Russell are attending the summit along with Hassabis, world politicians, other tech CEOs and civil society figures. Referring to AGI, Hassabis says "we're a long time before the systems become anywhere on the horizon" but says future generation systems will carry risks. Hence the summit. Critics of the summit argue that the focus on existential risk ignores short-term problems such as deepfakes. The government seems to have acknowledged the immediate concerns, with the agenda for the summit referring to election disruption and AI tools producing biased outcomes. Hassabis argues that there are three categories of risk and all "equally important" and need to be worked on simultaneously. The first is near-term risks such as deepfakes and bias. "Those types of issues … obviously are very pressing issues, especially with elections next year," he said. "So there … we need solutions now." Google DeepMind has already launched a tool that watermarks AI-generated images. The second risk category is rogue actors accessing AI tools, via publicly available and adjustable systems known as open source models, and using them to cause harm. "How does one restrict access to bad actors, but somehow enable all the good use cases? That's a big debate." The third is AGI, which is no longer discussed as a fantastical possibility. Hassabis says super powerful systems could be a "decade away plus" but the thinking on controlling them needs to start immediately. There are also alternative views in this field. Yann LeCun, the chief AI scientist at Mark Zuckerberg's Meta and a respected figure in AI, said last week that fears AI could wipe out humanity were "preposterous". Nonetheless, a concern among those worried about superintelligent systems is the notion that they could evade control. "Can it exfiltrate its own code, can extract its own code, improve its own code," says Hassabis. "Can it copy itself unauthorised? Because these would all be undesirable behaviours, because if you want to shut it down, you don't want it getting around that by copying itself somewhere else. There's a lot of behaviours like that, that would be undesirable in a powerful system." He said tests would have to be designed to head off the threat of such autonomous behaviour. "You've got to actually develop a test to test that … and then you can mitigate it, and maybe even legislate against it at some point. But the research has to be done first." The situation is further complicated by the fact that highly capable generative AI tools – technology that produces plausible text, image and voice from simple human prompts – are already out there and the regulatory framework to regulate them is still being built. Signs of a framework are emerging, such as commitments to AI safety signed by major western tech firms at the White House in July. But the commitments are voluntary. Hassabis talks of starting with an IPCC-style body before moving eventually to an entity "equivalent to" the anti-nuclear proliferation International Atomic Energy Agency, although he stresses that none of the regulatory analogies are "directly applicable" to AI. This is new territory. If you are in the pessimistic camp, it could take years to build a solid regulatory framework. And as Hassabis says, work on safety needs to start "yesterday".