TechScape: AI is feared to be apocalyptic or touted as world-changing – maybe it's neither

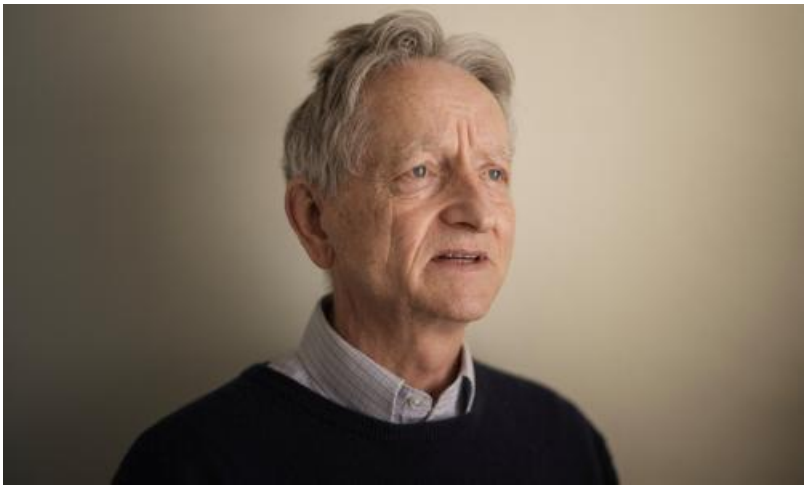
Publication Date: 2023-05-09

Author: Alex Hern

Section: Technology

Tags: Artificial intelligence (AI), TechScape newsletter, Computing, Consciousness, Google, Alphabet, ChatGPT, newsletters

Article URL: https://www.theguardian.com/technology/2023/may/09/techscape-artificial-intelligence-risk



What if AI doesn't fundamentally reshape civilisation? This week, I spoke to Geoffrey Hinton, the English psychologist-turned-computer scientist whose work on neural networks in the 1980s set the stage for the explosion in AI capabilities over the last decade. Hinton wanted to speak to deliver a message to the world: he is afraid of the technology he helped create. You need to imagine something more intelligent than us by the same difference that we're more intelligent than a frog. And it's going to learn from the web, it's going to have read every single book that's ever been written on how to manipulate people, and also seen it in practice." He now thinks the crunch time will come in the next five to 20 years, he says. "But I wouldn't rule out a year or two. And I still wouldn't rule out 100 years – it's just that 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." Hinton is not the first big figure in AI development to sound the alarm, and he won't be the last. The undeniable – and accelerating – improvement in the underlying technology lends itself easily to visions of unending progress. The clear possibility of a flywheel effect, where progress itself begets further progress, adds to the potential. Researchers are already seeing good results, for instance, on using AI-generated data to train new AI models, while others are incorporating AI systems into everything from chip design to data-centre operations. Another cohort of AI workers agree with the premise, but deny the conclusion. Yes, AI will change the world, but there's nothing to fear from that. This view – broadly lumped under the "singularitarian" label – is that AI development represents a massive leap in human capability but not necessarily a scary one. A world in which powerful AIs end human suffering is within grasp, they say, whether that's because we upload ourselves to a digital heaven or simply allow the machines to handle all the drudgework of human existence and live in a utopia of their creation. (A minority view is that AI will indeed wipe out humanity and that's good, too. Just as a parent doesn't fear their child inheriting the world, so should we be happy, rather than fearful, that an intelligence created by humans will surpass us and outlive us. "effective accelerationists" see their role as midwifes for a god. It isn't always clear how sincere they're being.) One response is to simply deny everything. If AI progress is overstated, or if the technological gains are likely to stall out, then we don't need to worry. History is littered with examples of progress that seemed unending but instead hit hard limits that no one had foreseen. You cannot take a steam engine to the moon, you do not have a flying car and a nuclear-powered washing machine is a bad idea for many reasons. We can already see potential stumbling blocks on the horizon: if GPT-4 is trained on an appreciable portion of all digitised text in existence, what is left for GPT-5? But I'm more interested in the middle ground. Most technologies do not end the world. (In fact, so far, humanity has a 100% hit

rate for not destroying itself, but past results may not be indicative of future performance.) Many technologies do change the world. How might that middle ground shake out for AI? ‘Small’ AI v ‘giants’ For me, the answer clicked when I read a leaked document purportedly from a Google engineer assessing the company’s hopes of winning the AI race. From our article: A document from a Google engineer leaked online said the company had done “a lot of looking over our shoulders at OpenAI”, referring to the developer of the ChatGPT chatbot. “The uncomfortable truth is, we aren’t positioned to win this arms race and neither is OpenAI. While we’ve been squabbling, a third faction has been quietly eating our lunch,” the engineer wrote. The engineer went on to state that the “third faction” posing a competitive threat to Google and OpenAI was the open-source community. The document is online, and I’d encourage you to give a read if you’re interested in the nuts and bolts of AI competition. There’s lots of granular detail about why the anonymous author thinks that Google and OpenAI might be on a losing path, from breakthroughs in “fine-tuning” and distribution to the ease with which one can adapt an open-source model to a hyper-specific use case. One particular passage caught my eye: Giant models are slowing us down. In the long run, the best models are the ones which can be iterated upon quickly. We should make small variants more than an afterthought, now that we know what is possible in the <20B parameter regime. The author of the memo is focused on one possibility – that “small” AI models will, by virtue of being distributed among many users and more easily retrained for specific niches, eventually catch up to and overtake the “giant” models like GPT-4 or Google’s own LaMDA, which represent the state of the art in the field. But there’s another possibility worth exploring: That they won’t, and they’ll “win” anyway. A large language model like GPT-4 is incredibly powerful yet laughably flawed. Despite the literal billions thrown at the system, it is still prone to basic errors like hallucination, will still misunderstand simple instructions and continues to stumble over basic concepts. The tale of the next decade of investment in large language models is going to be shovelling money in a pit to shave away ever more of those failure modes. Spending a billion dollars will get you from 99% to 99.9% accurate. Spending another 10 billion might get you to 99.99%. Spending a further 100 billion might get you to 99.999%. Meanwhile, the 99% OK version of the AI system, which once was gated behind a paywall on OpenAI’s website, filters down through the open source community until it’s sitting on your iPhone, running locally and being retrained on your personal communications every morning, learning how you talk and think without any data being shared with OpenAI or Google. AI A-OK? This vision of the future puts “super-intelligent AI” as a similar class of problem to “self-driving car”, but with a very different landscape. The problem plaguing the tech industry is that a self-driving car that is 99% safe is useless. You have no choice but to continue development, throwing ever more money at the problem, until you finally develop a system that is not only safer than a human driver, but so safe that no one alive will see the inexplicable moments when it does fail horribly and drives full-speed into a wall for no apparent reason. A generative AI isn’t like that. No one dies if your AI-powered music search engine labels Taylor Swift as “electroclash”. No property is destroyed if the poem you ask GPT to write for a colleague’s leaving card has a garbage metre. No one will sue if the cartoon character on the AI-generated poster for your kid’s birthday party has two thumbs. There will still be motivation for throwing bundles of money at the hard problems. But for the day-to-day use, small, cheap and nimble could beat large, expensive and flawless. And at the scale of the consumer tech industry, that could be enough to bend the arc of the future in a very different way. Think, perhaps, of supersonic flight. There’s no purely technological reason why the fastest transatlantic crossing is several hours slower now than it was when I was born. But a combination of consumer behaviour, the economics of the industry, the regulatory state and the plain difficulty of perennially increasing flight speed means that it is. Instead, the world optimised other things: comfort, fuel efficiency, safety and flexibility took the lead. There’s still the potential for disaster in that vision of the world. Perhaps the accumulation of small, cheap improvements to the light and nimble AI models still inexorably takes us towards superintelligence. Or perhaps there are still enough customers who are willing to throw a trillion dollars at adding another fraction of a percent of reliability to an AI system that the world faces existential risk anyway. But the central scenario for any new technology, I think, has to start with the assumption that the world next year will still look a lot like the world this year. I’ve not woken up dead yet, after all. If you want to read the complete version of the newsletter please subscribe to receive TechScape in your inbox every Tuesday