

● Sentient artificial general intelligence computer HAL 9000 in *2001: A Space Odyssey*

+CTRL KILL

Experts are warning that advanced AI poses an existential threat, but how can a piece of software wipe us out, what would motivate it and how do we stop AI from taking over?

WORDS STEPHEN COUSINS

"This mission is too important for me to allow you to jeopardise it," are the chilling words spoken by the supercomputer

HAL 9000 in the Stanley Kubrick sci-fi masterpiece *2001: A Space Odyssey* after it learns that two astronauts are planning to disconnect it after a malfunction.

Barely a couple of decades after the events depicted in that movie, many of the world's leading artificial intelligence (AI) technologists are warning of the arrival of out-of-control digital intelligence – and this one, they say, will not only endanger humans, but potentially destroy us all.

Tech giants such as Google, Meta and ChatGPT creator OpenAI are

investing huge amounts of money into creating AI systems designed to outthink and outperform human intelligence, but even experts within those companies are beginning to sound the alarm that advanced AI could pose a catastrophic threat.

In May last year, over 350 researchers and executives, including the CEO of OpenAI and 38 members of Google's DeepMind AI unit, signed a statement warning that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war". The largest-ever survey of AI professionals, published last month, found that between a third and half of respondents

believe there is at least a 10% chance of advanced AI leading to outcomes as bad as human extinction.

Unprecedented times

There are risks that a superior intelligence may purposefully eradicate people, or accidentally snuff us out while trying to pursue its own goals – just as we have with many less well adapted species. Programmers still have no idea how AI 'thinking' really functions and are unable to align it with human goals and values.

Furthermore, as commercial imperatives see AI increasingly embedded into infrastructure and military, political and corporate systems, its physical agency to interact with the world will increase and

humans will have less and less opportunity to intervene and course-correct.

More optimistic supporters of AI argue that it could one day unlock the answers to mankind's greatest challenges, such as helping tackle climate change, curing disease and lifting millions of people out of poverty. However, we ignore the bleaker possibilities at our peril.

"The specific danger we face is that if we create something vastly more intelligent than us and if we do not successfully align it on the very first try, it is overwhelmingly likely to kill us," a spokesperson for the Machine Intelligence Research Institute (MIRI) tells *E+T*.

It wasn't so long ago that fears of AI were confined to sci-fi novels and movies.

As recently as 2020 the mere suggestion that machine intelligence could one day rival human capabilities was given short shrift by academics.

But last year the technology crossed the Rubicon as a confluence of innovations – such as breakthroughs in mathematical modelling, improvements to hardware and computing power, and the emergence of massive high-quality datasets – supercharged generative AI tools, revealing their incredible potential.

A key development was in training AI using transformer-based large language models, which made it possible for software to ingest billions of sentences or images and quickly recognise patterns to produce more natural-sounding responses or visual reproductions.

"In a few short years, language models went from barely being able to count to writing functional software and solving college-level mathematics problems," explains Ajeya Cotra, senior analyst on the potential risks from advanced AI at Open Philanthropy. "No human engineer specifically designed these capabilities; instead, models gained new capabilities themselves just from being trained with more hardware and more data."

Soon it became clear the tech could not only replace human jobs that are easily automated, but also impact complex work roles once considered safe. A report from Goldman Sachs has estimated that AI capable of content generation could do a quarter of all work currently done by humans – equivalent to 300 million jobs lost across Europe and the US.

It's a big jump from stealing people's jobs to wiping them off the face of the planet, but many researchers are convinced that digital intelligence will soon reach an inflection point where it transitions into the realms of an artificial general intelligence (AGI) (see *Know your AI from your AGI*) capable of learning and performing any intellectual task a human can. This could take less than a decade. Looking even further ahead, an artificial super intelligence (ASI) could have cognitive functions and thinking skills even more advanced than any human.

Assuming that AI does one day outsmart its flesh-and-blood creators, what would motivate it to wipe us out *Terminator*-style?

Paperclip theory

"An ASI would not necessarily be conscious or have its own will, but would likely be trained by humans to achieve a certain goal," explains Otto Barten, director at the Existential Risk Observatory, which is campaigning to reduce the risk of human extinction by AI. "If humans are in the way of achieving this goal, we are at risk in the same way an anthill may be at risk if we want to build a road over it."

Barten's observations pick up on the 'paper clip' thought experiment, developed by philosopher Nick Bostrom at Oxford University. This imagines an ASI programmed with the goal of producing paper clips and given the ability to learn and invent ways to achieve its goal better.

The AI is so single-minded that if there is a way of turning something into paperclips, it will find it and secure the appropriate resources, ultimately diverting them from all other activities. The world is then inundated with paperclips. Human efforts to stop the AI are seen as subverting its goal, so it becomes focused on its own survival and fights us for resources – and, as a superior intelligence, it is likely to win that battle.

MIRI has identified three scenarios whereby an AI might either actively or accidentally destroy humans. In scenario one, it is indifferent to us but destroys something critical to our survival as a side effect of pursuing its goals. For example, it decides to use all the hydrogen in the oceans to create energy. In scenario two, it views us as a threat, perhaps for no other reason than we might make another AI that could prevent it from successfully pursuing its goals. In the final scenario, we are made of resources that it might want to use for something else.

If we play devil's advocate and accept the notion that AI does have



the ‘will’ to exterminate us, what about the ‘means’? How can an entity confined to computer circuit boards and binary code have the physical agency to wreak devastation on the world?

One obvious avenue is military applications: military leaders are increasingly turning decision-making over to automated systems, with a dizzying array of AI defence solutions that are either in development or already in use.

Australian company Athena AI has demonstrated a system that can detect people wearing military clothes and carrying weapons and pinpoint them on a map. Israel was recently accused by journalists of using a system largely built on AI to ‘generate’ targets in Gaza almost automatically.

According to the campaign group Stop Killer Robots, the capacity for fully-autonomous weapons, able to decide to engage a target without human intervention, is closer than we think. The organisation is calling for an international legal treaty that bans autonomous weapon systems that do not allow for meaningful human control and use sensors to target humans.

It is intrinsically difficult to predict how a super AI could influence our world, not least because an inferior intelligence, like ours, will always find it harder to determine the actions of a vastly smarter one. As Barten puts it: “When I play chess against Magnus Carlsen, I cannot predict which moves he will play, but I can predict the end result: I lose.”

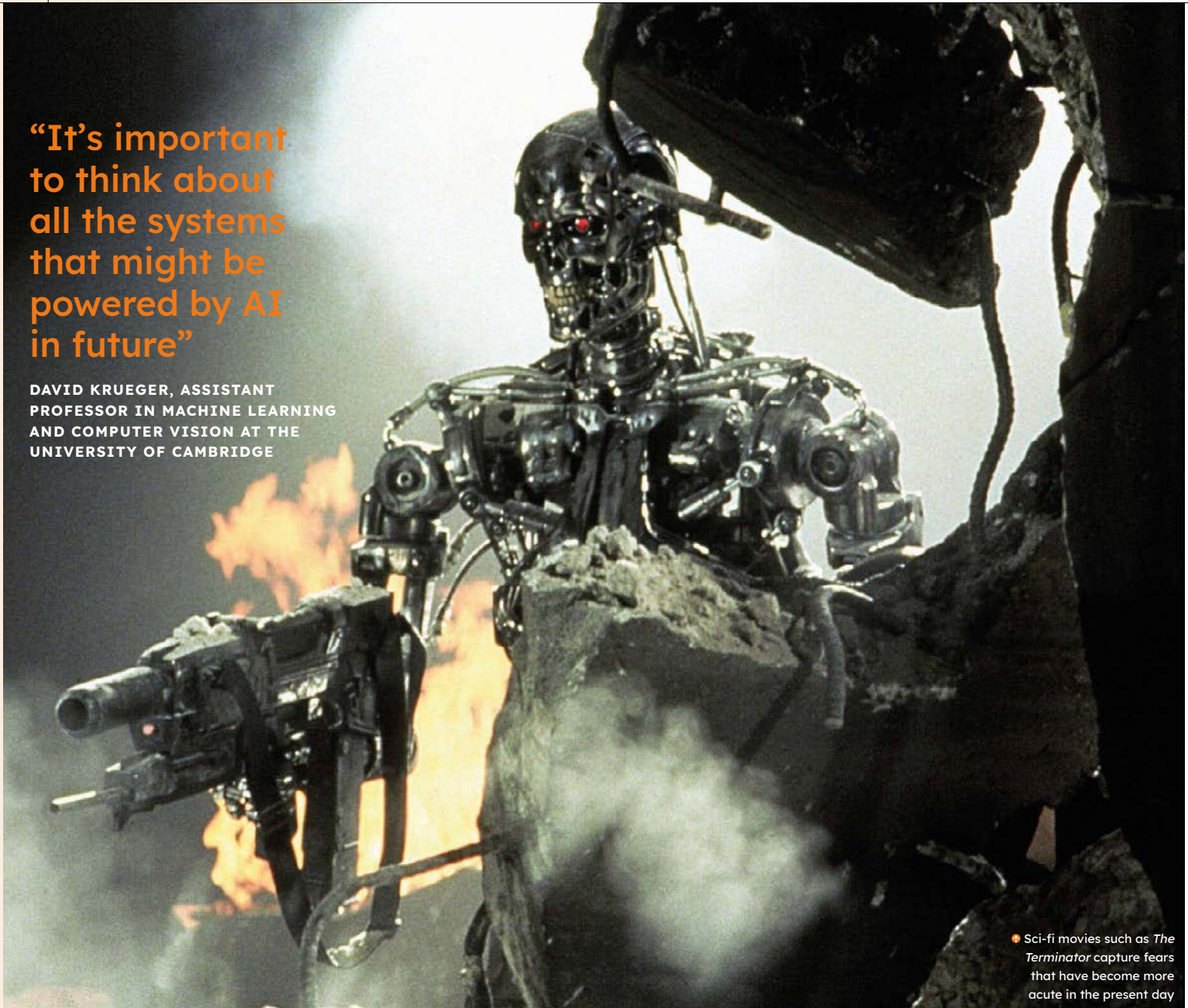
Nevertheless, AI’s future ubiquity could open up multiple paths to destruction. Experts have talked about the ‘obsolescence regime’ whereby, because AIs are cheaper, faster and smarter than people, companies that don’t rely on them become uncompetitive. Therefore AI gains more power and control and humans’ influence falls away.

AI lives online and if it becomes better than us at programming and hacking, it could potentially take down large parts of the internet, IT systems and connected hardware.

“It’s important to think about all the systems that might be powered by AI in future,” says David Krueger, assistant professor in machine learning and computer vision at the University

“It’s important to think about all the systems that might be powered by AI in future”

DAVID KRUEGER, ASSISTANT PROFESSOR IN MACHINE LEARNING AND COMPUTER VISION AT THE UNIVERSITY OF CAMBRIDGE



Sci-fi movies such as *The Terminator* capture fears that have become more acute in the present day

of Cambridge. “Infrastructure, military, decision-making, politics and corporate governance could be handed over increasingly to AI systems ... If AIs are making decisions very quickly, and we don’t always understand why they make the decisions they do, we will have less and less ability to intervene.”

Social influencer

AI algorithms already play a major role in spreading lies on the internet and social media, and concerns have been raised about its role in radicalisation, political polarisation and even the genocide in Myanmar. A World Economic

Forum survey named AI-generated misinformation and disinformation as one of the top global risks for 2024. Earlier this year, a survey found that 70% of MPs in the UK are concerned AI-generated content will increase the spread of disinformation ahead of the general election.

A powerful AGI could generate and spread highly realistic fake content – including deepfake images and videos, news reports or other propaganda – to bend humans to its will, potentially turning humanity against itself or on a path to self-destruction. As Barten notes, many a war has been waged by humans

convinced by “superior narratives” and false beliefs.

AI doesn’t need a physical presence if it can get people to act on its behalf through a campaign of manipulation, coercion through impersonation or blackmail, or simply paying for human labour to enact its wishes.

Potential manipulation is scary, but the speed with which a superior intelligence could trigger a global catastrophe has some researchers even more worried. “If it can solve certain biological challenges, it could build itself a tiny molecular laboratory and manufacture and release lethal bacteria,” says Eliezer Yudkowsky,

Image: Alamy

WHAT JOBS WILL AI TAKE OVER?

In administrative occupations, about one-third of jobs could be displaced



Although AI is often seen as a threat to jobs, this analysis of the UK jobs market by IPPR shows some sectors will benefit – like STEM – while others like administration and secretarial will be hard hit.

Source: IPPR analysis of LFS and ONET (2023)

co-founder and research fellow at MIRI. “What that looks like is everybody on Earth falling over dead inside the same second.”

Rapid annihilation and “not letting humans know there is going to be a fight” may be important to an AI, Yudkowsky adds, because “if you give the humans warning, if you kill some of them before others, maybe somebody panics and launches all the nuclear weapons. Then you are slightly inconvenienced.”

It’s a chilling thought, but scientists are far from achieving a consensus on whether an AI could ever gain this level of power, or the timescales on which a calamity could play out.

Perhaps we should be more concerned about the shorter-term risk of humans harnessing super intelligence for nefarious reasons. Almost three-quarters of respondents to the large-scale AI survey labelled AI-powered scenarios involving deepfakes, manipulation of public opinion, engineered weapons, authoritarian control of populations and worsening economic inequality as either of substantial or extreme concern.

“Digital scientists could be used to develop bioweapons instead of cancer cures,” says Cotra. “Countries may find themselves racing to use digital ‘generals’ and ‘soldiers’ to achieve military dominance ➤

before their rivals do, just as they did with the nuclear bomb in the 1940s.”

Dr Gopichand Katragadda, AI technology leader and current president of the IET, says: “As knowledge becomes more commoditised through AI, there’s growing apprehension among knowledge workers about their future relevance. With its ability to work across various domains like language, vision and sensors, AGI is perceived as more human-like, intensifying concerns about the evolving nature of work, innovation and leadership.”

Short circuit

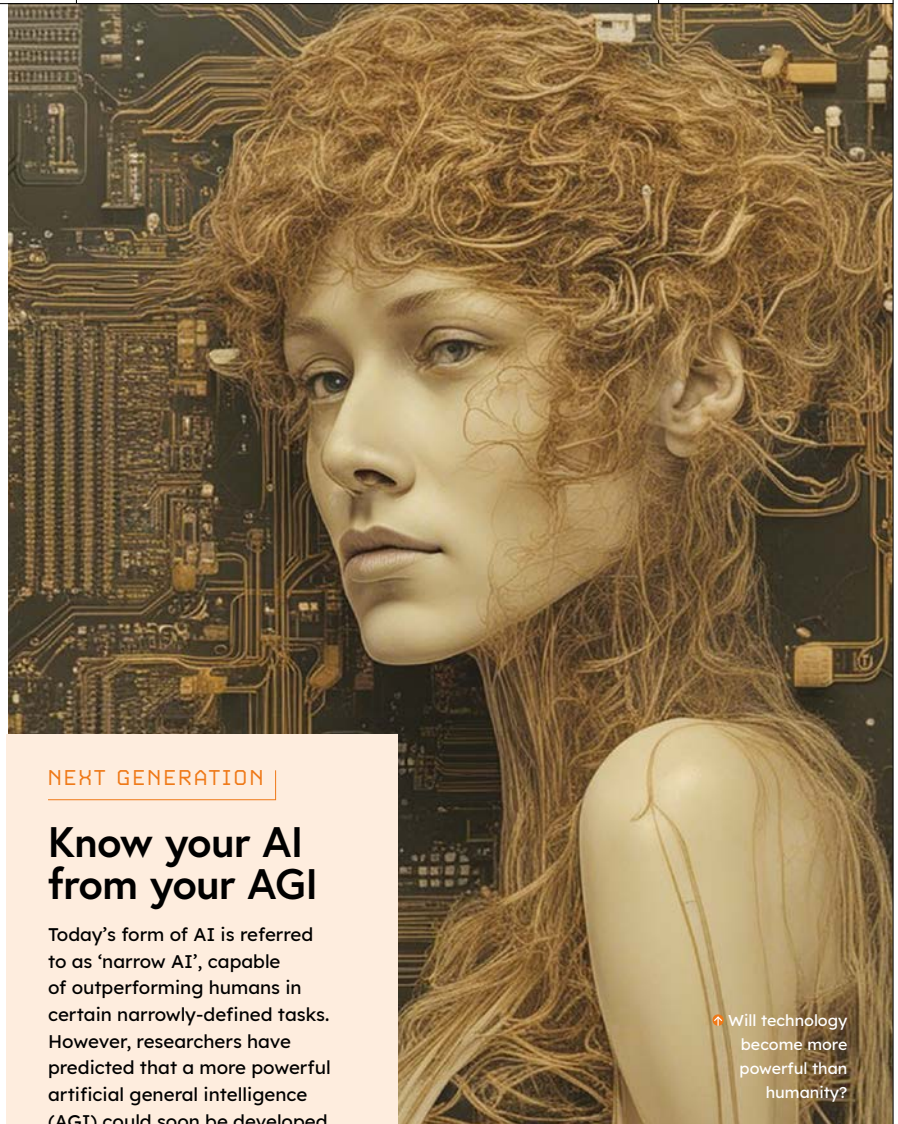
As AIs become more powerful, humans will need to identify ways to ensure they pursue our values and goals, while still fostering the innovation needed to deliver the many beneficial outcomes for society.

The issue of ‘AI alignment’ is a critical one because at present no one fully understands how AI systems ‘think’ or how to reliably instil goals in them. It is also fraught with technical, philosophical and sociological challenges.

The gap in our understanding of AI was amusingly demonstrated in January when the delivery firm DPD was forced to disable part of its AI-powered online chatbot after a disgruntled customer was able to make it swear and write a poem criticising the company. And many of us have seen the sometimes spectacularly wrong answers coming out of ChatGPT.

“We don’t currently have any serious understanding of how natural human language is being interpreted by AI, which doesn’t fully understand human concepts,” says Krueger, who adds that efforts to align such systems have so far been “very empirical and ad hoc” with “no real deep scientific understanding or principled approach”.

Bowling ahead with AI development may be foolish given the many unknowns, requiring concerted action from tech companies, politicians and other decision-makers. Solutions could take the form of a pause to AI development while the risks are assessed (as Musk and his contemporaries called for in their open letter), policy interventions such as creating a licensing regime for frontier



NEXT GENERATION

Know your AI from your AGI

Today’s form of AI is referred to as ‘narrow AI’, capable of outperforming humans in certain narrowly-defined tasks. However, researchers have predicted that a more powerful artificial general intelligence (AGI) could soon be developed.

An AGI would be capable of learning and performing any intellectual task a human can, without the need for humans to train it using underlying models. The concept remains theoretical, but many academics and researchers believe there is at least a chance that it could be achieved by the end of the decade, while others believe it could take years, decades or even longer than a century.

Looking further ahead, AGI could one day evolve into ‘artificial super intelligence’ (ASI) with cognitive functions and thinking skills more advanced than any human. The largest ever survey of AI researchers, published in January 2024, gave 50% odds to the notion that AI will be able to outperform humans on every task by 2047.

Will technology become more powerful than humanity?

AI, mandating model evaluations or third-party auditing, or tracking hardware that can be used to train AI.

Increasing societal awareness of AI existential risk could be equally effective, says Barten. “It would make sense to democratically decide how far we want AI capabilities to go, and which risks we find acceptable.”

Yoshua Bengio, one of the modern godfathers of AI, has called for companies developing powerful AI systems to be required to have independent board members representing the interests of society, as advances in the technology accelerate rapidly.

Get these complex factors right and, instead of directing a *Terminator*-style apocalypse, digital intelligence could one day help humanity flourish as never before. **ET**