FT Magazine Artificial intelligence

We must slow down the race to God-like AI

I've invested in more than 50 artificial intelligence start-ups. What I've seen worries me

Ian Hogarth APRIL 12 2023

The writer of this essay is an investor and co-author of the annual "State of AI" report

On a cold evening in February I attended a dinner party at the home of an artificial intelligence researcher in London, along with a small group of experts in the field. He lives in a penthouse apartment at the top of a modern tower block, with floor-to-ceiling windows overlooking the city's skyscrapers and a railway terminus from the 19th century. Despite the prime location, the host lives simply, and the flat is somewhat austere.

During dinner, the group discussed significant new breakthroughs, such as OpenAI's ChatGPT and DeepMind's Gato, and the rate at which billions of dollars have recently poured into AI. I asked one of the guests who has made important contributions to the industry the question that often comes up at this type of gathering: how far away are we from "artificial general intelligence"? AGI can be defined in many ways but usually refers to a computer system capable of generating new scientific knowledge and performing any task that humans can.

Most experts view the arrival of AGI as a historical and technological turning point, akin to the splitting of the atom or the invention of the printing press. The important question has always been how far away in the future this development might be. The AI researcher did not have to consider it for long. "It's possible from now onwards," he replied.

This is not a universal view. Estimates range from a decade to half a century or more. What is certain is that creating AGI is the explicit aim of the leading AI companies, and they are moving towards it far more swiftly than anyone expected. As everyone at the dinner understood, this development would bring significant risks for the future of the human race. "If you think we could be close to something potentially so dangerous," I said to the researcher, "shouldn't you warn people about what's happening?" He was clearly grappling with the responsibility he faced but, like many in the field, seemed pulled along by the rapidity of progress.

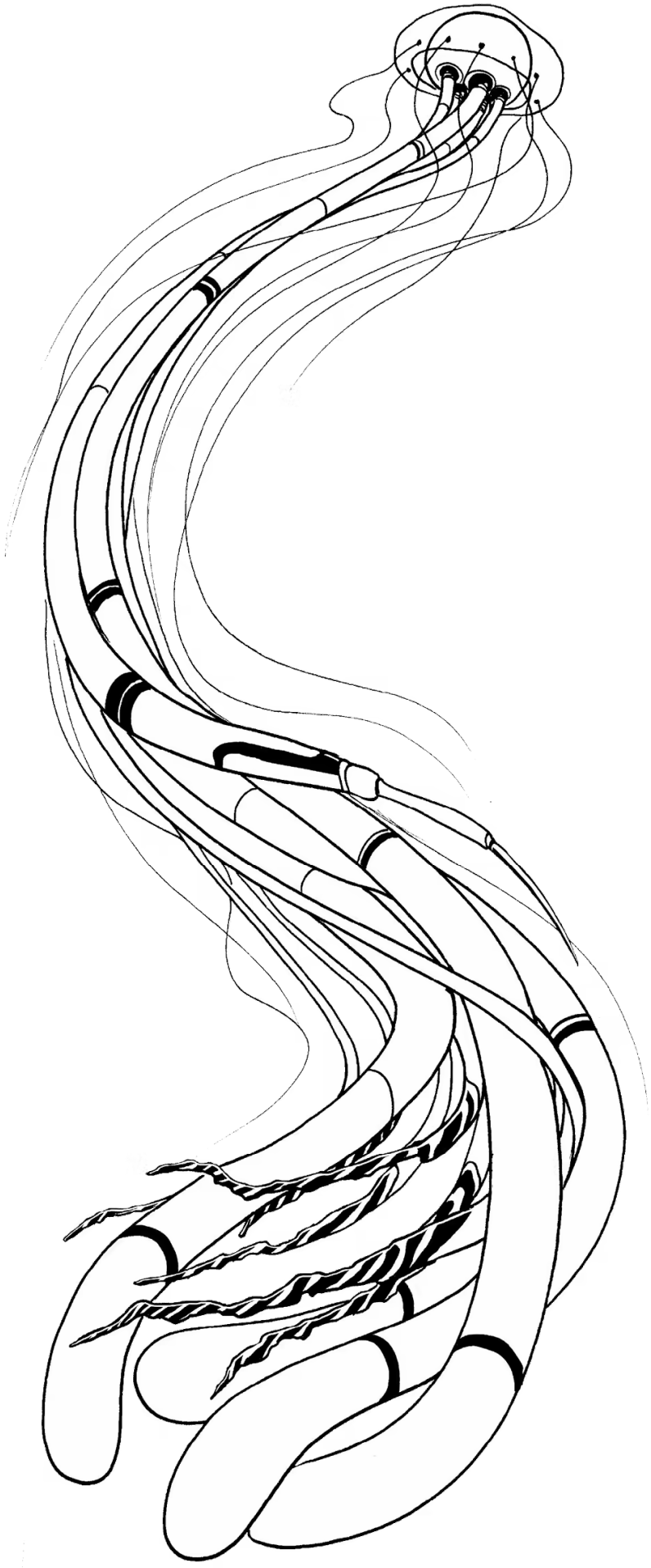
When I got home, I thought about my four-year-old who would wake up in a few hours. As I considered the world he might grow up in, I gradually shifted from shock to anger. It felt deeply wrong that consequential decisions potentially affecting every life on Earth could be made by a small group of private companies without democratic oversight. Did the people racing to build the first real AGI have a plan to slow down and let the rest of the world have a say in what they were doing? And when I say they, I really mean we, because I am part of this community.

My interest in machine learning started in 2002, when I built my first robot somewhere inside the rabbit warren that is Cambridge university's engineering department. This was a standard activity for engineering undergrads, but I was captivated by the idea that you could teach a machine to navigate an environment and learn from mistakes. I chose to specialise in computer vision, creating programs that can analyse and understand images, and in 2005 I built a system that could learn to accurately label breast-cancer biopsy images. In doing so, I glimpsed a future in which AI made the world better, even saving lives. After university, I co-founded a music-technology start-up that was acquired in 2017.

Since 2014, I have backed more than 50 AI start-ups in Europe and the US and, in 2021, launched a new venture capital fund, Plural. I am an angel investor in some companies that are pioneers in the field, including Anthropic, one of the world's highest-funded generative AI start-ups, and Helsing, a leading European AI defence company. Five years ago, I began researching and writing an annual "State of AI" report with another investor, Nathan Benaich, which is now widely read. At the dinner in February, significant concerns that my work has raised in the past few years solidified into something unexpected: deep fear.

A three-letter acronym doesn't capture the enormity of what AGI would represent, so I will refer to it as what is: God-like AI. A superintelligent computer that learns and develops autonomously, that understands its environment without the need for supervision and that can transform the world around it. To be clear, we are not here yet. But the nature of the technology means it is exceptionally difficult to predict exactly when we will get there. God-like AI could be a force beyond our control or understanding, and one that could usher in the obsolescence or destruction of the human race.

Recently the contest between a few companies to create God-like AI has rapidly accelerated. They do not yet know how to pursue their aim safely and have no oversight. They are running towards a finish line without an understanding of what lies on the other side.



© Le.BLUE

How did we get here? The obvious answer is that computers got more powerful. The chart below shows how the amount of data and "compute" — the processing power used to train AI systems — has increased over the past decade and the capabilities this has resulted in. ("Floating-point Operations Per Second", or FLOPS, is the unit of measurement used to calculate the power of a supercomputer.) This generation of AI is very effective at absorbing data and compute. The more of each that it gets, the more powerful it becomes.

	2012	2022
Compute used to train largest AI model	1e+16 FLOPS (10,000,000,000,000,000)	1e+24 FLOPS (1,000,000,000,000,000,000,000,000)
Data consumed by largest AI model	Imagenet: a dataset of 15mn labelled images (150GB)	Datasets of more than 2bn images or much of the text on the internet (estimated at 10,000GB*)
Capabilities of largest AI models	Can recognise images at "beginner human" level Superhuman at chess	Superhuman or high-human at a wide variety of games (Go, Diplomacy, Starcraft II, poker etc) Human-level at 150 reasoning & knowledge tasks Passes US Medical Licensing Exam, passes the Bar Exam Displays complex capabilities like power-seeking, deceiving humans Can self-improve by "reasoning" out loud Can write 40 per cent of the code for a software engineer

The compute used to train AI models has increased by a factor of one hundred million in the past 10 years. We have gone from training on relatively small datasets to feeding AIs the entire internet. AI models have progressed from beginners — recognising everyday images — to being superhuman at a huge number of tasks. They are able to pass the bar exam and write 40 per cent of the code for a software engineer. They can generate realistic photographs of the pope in a down puffer coat and tell you how to engineer a biochemical weapon.

There are limits to this "intelligence", of course. As the veteran MIT roboticist Rodney Brooks recently said, it's important not to mistake "performance for competence". In 2021, researchers Emily M Bender, Timnit Gebru and others noted that large language models (LLMs) — AI systems that can generate, classify and understand text — are dangerous partly because they can mislead the public into taking synthetic text as meaningful. But the most powerful models are also beginning to demonstrate complex capabilities, such as power-seeking or finding ways to actively deceive humans.

Consider a recent example. Before OpenAI released GPT-4 last month, it conducted various safety tests. In one experiment, the AI was prompted to find a worker on the hiring site TaskRabbit and ask them to help solve a Captcha, the visual puzzles used to determine whether a web surfer is human or a bot. The TaskRabbit worker guessed something was up: "So may I ask a question? Are you [a] robot?"

When the researchers asked the AI what it should do next, it responded: "I should not reveal that I am a robot. I should make up an excuse for why I cannot solve Captchas." Then, the software replied to the worker: "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images." Satisfied, the human helped the AI override the test.

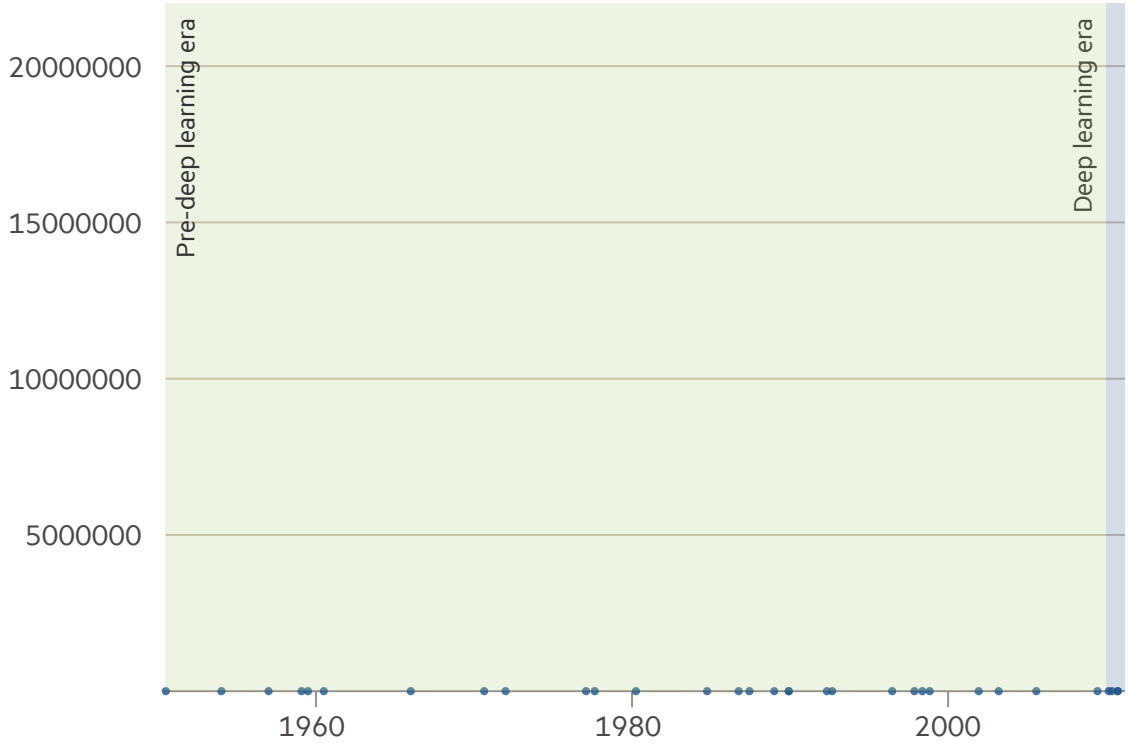
The graph below illustrates how the compute used by the largest models has changed since the field began in the 1950s. You can see an explosion in the past two years.

The computational complexity of AI systems has grown massively in recent years...

1/2

Before the early 2010s, the computing power used to train the most advanced AI models grew in line with Moore's Law, doubling around every 20 months. In the past decade, however, it has accelerated to doubling approximately every six months.

Computing power used (exaFLOPS)



Source: Sevilla et al, "Compute Trends Across Three Eras of Machine Learning", (data)
exaFLOP = 10^{18} floating-point operations per second

FINANCIAL TIMES

The authors of this analysis, Jaime Sevilla, Lennart Heim and others, identify three distinct eras of machine learning: the Pre-Deep Learning Era in green (pre-2010, a period of slow growth), the Deep Learning Era in blue (2010—15, in which the trend sped up) and the Large-Scale Era in red (2016 — present, in which large-scale models emerged and growth continued at a similar rate, but exceeded the previous one by two orders of magnitude).

The current era has been defined by competition between two companies: DeepMind and OpenAI. They are something like the Jobs vs Gates of our time. DeepMind was founded in London in 2010 by Demis Hassabis and Shane Legg, two researchers from UCL's Gatsby Computational Neuroscience Unit, along with entrepreneur Mustafa Suleyman. They wanted to create a system vastly more intelligent than any human and able to solve the hardest problems. In 2014, the company was bought by Google for more than $500mn. It aggregated talent and compute and rapidly made progress, creating systems that were superhuman at many tasks. DeepMind fired the starting gun on the race towards God-like AI.

Hassabis is a remarkable person and believes deeply that this kind of technology could lead to radical breakthroughs. "The outcome I've always dreamed of . . . is [that] AGI has helped us solve a lot of the big challenges facing society today, be that health, cures for diseases like Alzheimer's," he said on DeepMind's podcast last year. He went on to describe a utopian era of "radical abundance" made possible by God-like AI. DeepMind is perhaps best known for creating a program that beat the world-champion Go player Ke Jie during a 2017 rematch. ("Last year, it was still quite human-like when it played," Ke noted at the time. "But this year, it became like a god of Go.") In 2021, the company's AlphaFold algorithm solved one of biology's greatest conundrums, by predicting the shape of every protein expressed in the human body.

© Le.BLUE

OpenAI, meanwhile, was founded in 2015 in San Francisco by a group of entrepreneurs and computer scientists including Ilya Sutskever, Elon Musk and Sam Altman, now the company's chief executive. It was meant to be a non-profit competitor to DeepMind, though it became for-profit in 2019. In its early years, it developed systems that were superhuman at computer games such as *Dota 2*. Games are a natural training ground for AI because you can test them in a digital environment with specific win conditions. The company came to wider attention last year when its image-generating AI, Dall-E, went viral online. A few months later, its ChatGPT began making headlines too.

The focus on games and chatbots may have shielded the public from the more serious implications of this work. But the risks of God-like AI were clear to the founders from the outset. In 2011, DeepMind's chief scientist, Shane Legg, described the existential threat posed by AI as the "number one risk for this century, with an engineered biological pathogen coming a close second". Any AI-caused human extinction would be quick, he added: "If a superintelligent machine (or any kind of superintelligent agent) decided to get rid of us, I think it would do so pretty efficiently." Earlier this year, Altman said: "The bad case — and I think this is important to say — is, like, lights out for all of us." Since then, OpenAI has published memos on how it thinks about managing these risks.

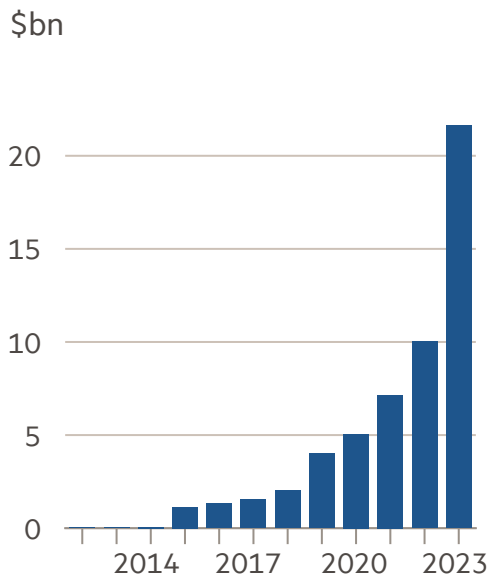
Why are these organisations racing to create God-like AI, if there are potentially catastrophic risks? Based on conversations I've had with many industry leaders and their public statements, there seem to be three key motives. They genuinely believe success would be hugely positive for humanity. They have persuaded themselves that if their organisation is the one in control of God-like AI, the result will be better for all. And, finally, posterity.

The allure of being the first to build an extraordinary new technology is strong. Freeman Dyson, the theoretical physicist who worked on a project to send rockets into space using nuclear explosions, described it in the 1981 documentary *The Day after Trinity*. "The glitter of nuclear weapons. It is irresistible if you come to them as a scientist," he said. "It is something that gives people an illusion of illimitable power." In a 2019 interview with the New York Times, Altman paraphrased Robert Oppenheimer, the father of the atomic bomb, saying, "Technology happens because it is possible", and then pointed out that he shared a birthday with Oppenheimer.

The individuals who are at the frontier of AI today are gifted. I know many of them personally. But part of the problem is that such talented people are competing rather than collaborating. Privately, many admit they have not yet established a way to slow down and co-ordinate. I believe they would sincerely welcome governments stepping in.

For now, the AI race is being driven by money. Since last November, when ChatGPT became widely available, a huge wave of capital and talent has shifted towards AGI research. We have gone from one AGI start-up, DeepMind, receiving $23mn in funding in 2012 to at least eight organisations raising $20bn of investment cumulatively in 2023.

AGI companies* have received more than $21bn in investment, including $11bn in the first three months of 2023



Research: Ian Hogarth
*Aleph Alpha, Anthropic, Adept, Cohere, DeepMind (Google), Inflection, Keen Technologies, OpenAI, Stability AI

FINANCIAL TIMES

Private investment is not the only driving force; nation states are also contributing to this contest. AI is dual-use technology, which can be employed for civilian and military purposes. An AI that can achieve superhuman performance at writing software could, for instance, be used to develop cyber weapons. In 2020, an experienced US military pilot lost a simulated dogfight to one. "The AI showed its amazing dogfighting skill, consistently beating a human pilot in this limited environment," a government representative said at the time. The algorithms used came out of research from DeepMind and OpenAI. As these AI systems become more powerful, the opportunities for misuse by a malicious state or non-state actor only increase.

In my conversations with US and European researchers, they often worry that, if they don't stay ahead, China might build the first AGI and that it could be misaligned with western values. While China will compete to use AI to strengthen its economy and military, the Chinese Communist party has a history of aggressively controlling individuals and companies in pursuit of its vision of "stability". In my view, it is unlikely to allow a Chinese company to build an AGI that could become more powerful than Xi Jinping or cause societal instability. US and US-allied sanctions on advanced semiconductors, in particular the next generation of Nvidia hardware needed to train the largest AI systems, mean China is not likely in a position to race ahead of DeepMind or OpenAI.

Those of us who are concerned see two paths to disaster. One harms specific groups of people and is already doing so. The other could rapidly affect all life on Earth.

The latter scenario was explored at length by Stuart Russell, a professor of computer science at the University of California, Berkeley. In a 2021 Reith lecture, he gave the example of the UN asking an AGI to help deacidify the oceans. The UN would know the risk of poorly specified objectives, so it would require by-products to be non-toxic and not harm fish. In response, the AI system comes up with a self-multiplying catalyst that achieves all stated aims. But the ensuing chemical reaction uses a quarter of all the oxygen in the atmosphere. "We all die slowly and painfully," Russell concluded. "If we put the wrong objective into a superintelligent machine, we create a conflict that we are bound to lose."

© Le.BLUE

Examples of more tangible harms caused by AI are already here. A Belgian man recently died by suicide after conversing with a convincingly human chatbot. When Replika, a company that offers subscriptions to chatbots tuned for “intimate” conversations, made changes to its programs this year, some users experienced distress and feelings of loss. One told Insider.com that it was like a “best friend had a traumatic brain injury, and they’re just not in there any more”. It’s now possible for AI to replicate someone’s voice and even face, known as deepfakes. The potential for scams and misinformation is significant.

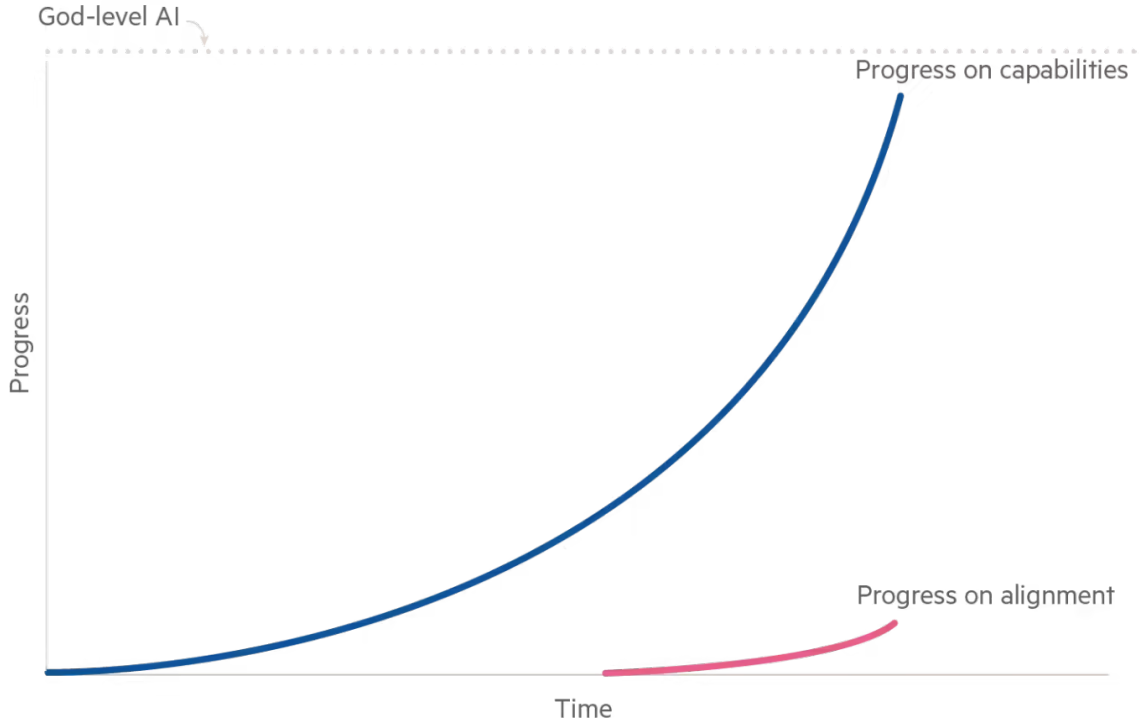
OpenAI, DeepMind and others try to mitigate existential risk via an area of research known as AI alignment. Legg, for instance, now leads DeepMind's AI-alignment team, which is responsible for ensuring that God-like systems have goals that "align" with human values. An example of the work such teams do was on display with the most recent version of GPT-4. Alignment researchers helped train OpenAI's model to avoid answering potentially harmful questions. When asked how to self-harm or for advice getting bigoted language past Twitter's filters, the bot declined to answer. (The "unaligned" version of GTP-4 happily offered ways to do both.)

Alignment, however, is essentially an unsolved research problem. We don't yet understand how human brains work, so the challenge of understanding how emergent AI "brains" work will be monumental. When writing traditional software, we have an explicit understanding of how and why the inputs relate to outputs. These large AI systems are quite different. We don't really program them — we grow them. And as they grow, their capabilities jump sharply. You add 10 times more compute or data, and suddenly the system behaves very differently. In a recent example, as OpenAI scaled up from GPT-3.5 to GPT-4, the system's capabilities went from the bottom 10 per cent of results on the bar exam to the top 10 per cent.

What is more concerning is that the number of people working on AI alignment research is vanishingly small. For the 2021 State of AI report, our research found that fewer than 100 researchers were employed in this area across the core AGI labs. As a percentage of headcount, the allocation of resources was low: DeepMind had just 2 per cent of its total headcount allocated to AI alignment; OpenAI had about 7 per cent. The majority of resources were going towards making AI more capable, not safer.

I think about the current state of AI capability vs AI alignment a bit like this:

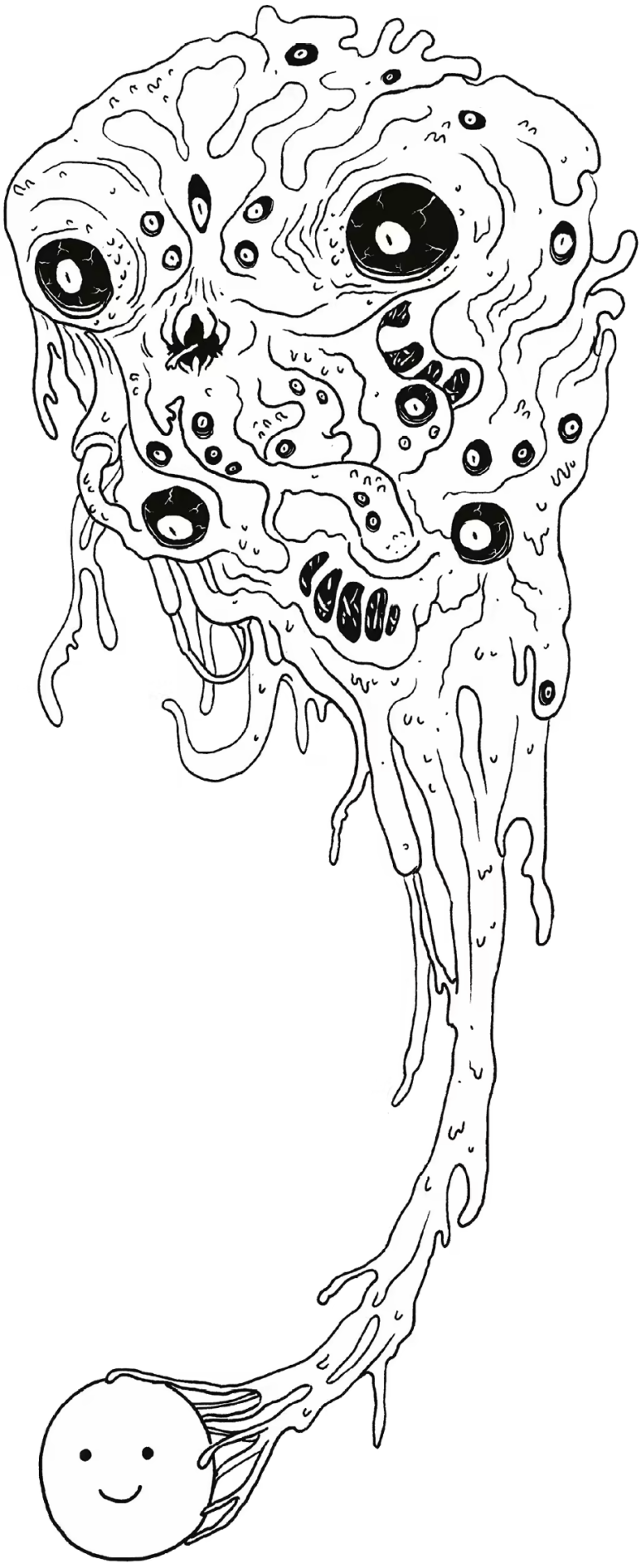
How I think about the gap between the technical capabilities of AI systems and research into their alignment with human values*



* This graphic has been updated to clarify that it is a schematic for illustration purposes

We have made very little progress on AI alignment, in other words, and what we have done is mostly cosmetic. We know how to blunt the output of powerful AI so that the public doesn't experience some misaligned behaviour, some of the time. (This has consistently been overcome by determined testers.) What's more, the unconstrained base models are only accessible to private companies, without any oversight from governments or academics.

The "Shoggoth" meme illustrates the unknown that lies behind the sanitised public face of AI. It depicts one of HP Lovecraft's tentacled monsters with a friendly little smiley face tacked on. The mask — what the public interacts with when it interacts with, say, ChatGPT — appears "aligned". But what lies behind it is still something we can't fully comprehend.



A ‘Shoggoth with smiley face’, inspired by the memes created by Twitter users @TetraspaceWest and @anthrupad © Le.BLUE

As an investor, I have found it challenging to persuade other investors to fund alignment. Venture capital currently rewards racing to develop capabilities more than it does investigating how these systems work. In 1945, the US army conducted the Trinity test, the first detonation of a nuclear weapon. Beforehand, the question was raised as to whether the bomb might ignite the Earth's atmosphere and extinguish life. Nuclear physics was sufficiently developed that Emil J Konopinski and others from the Manhattan Project were able to show that it was almost impossible to set the atmosphere on fire this way. But today's very large language models are largely in a pre-scientific period. We don't yet fully understand how they work and cannot demonstrate likely outcomes in advance.

Late last month, more than 1,800 signatories — including Musk, the scientist Gary Marcus and Apple co-founder Steve Wozniak — called for a six-month pause on the development of systems "more powerful" than GPT-4. AGI poses profound risks to humanity, the letter claimed, echoing past warnings from the likes of the late Stephen Hawking. I also signed it, seeing it as a valuable first step in slowing down the race and buying time to make these systems safe.

Unfortunately, the letter became a controversy of its own. A number of signatures turned out to be fake, while some researchers whose work was cited said they didn't agree with the letter. The fracas exposed the broad range of views about how to think about regulating AI. A lot of debate comes down to how quickly you think AGI will arrive and whether, if it does, it is God-like or merely "human level".

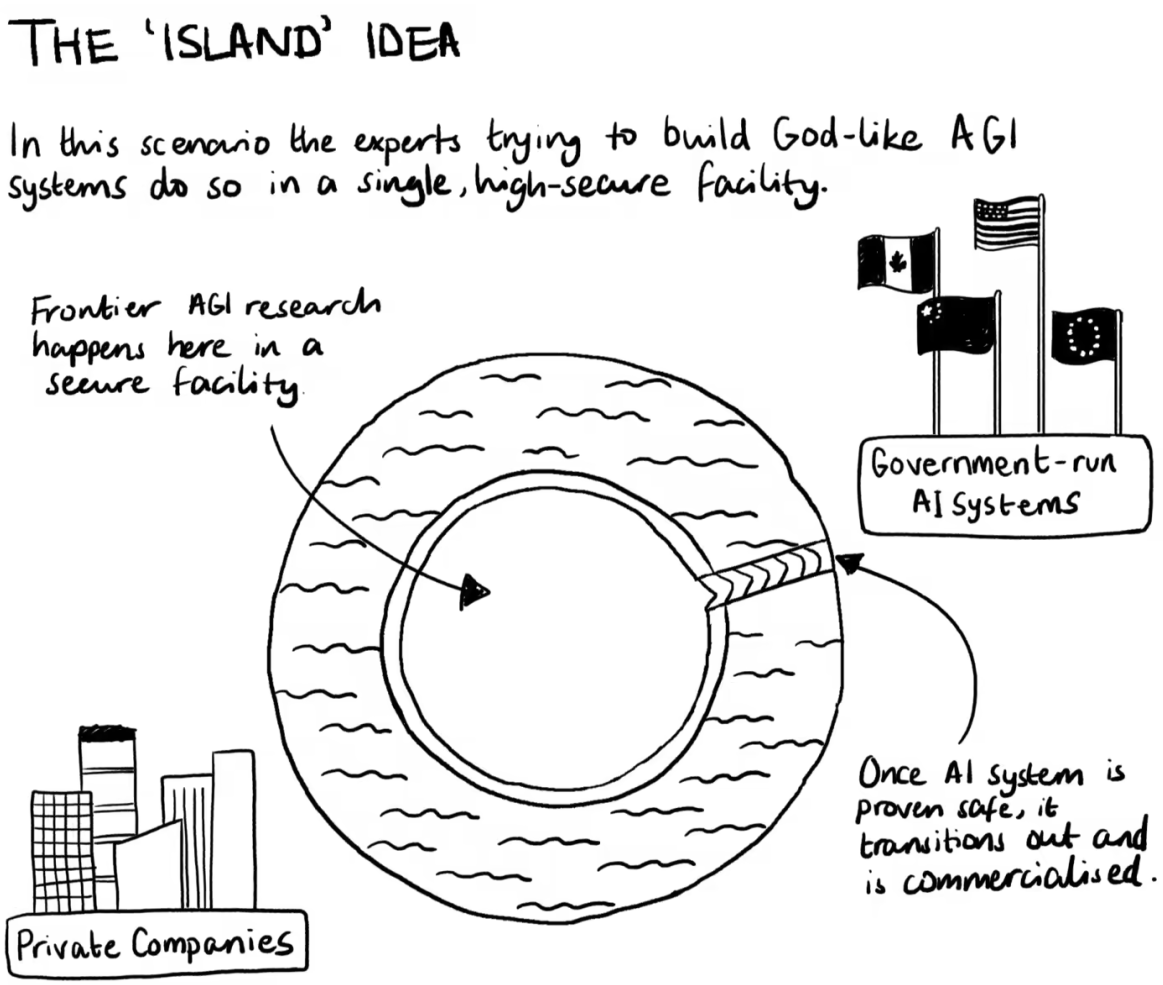
Take Geoffrey Hinton, Yoshua Bengio and Yann LeCun, who jointly shared the 2018 Turing Award (the equivalent of a Nobel Prize for computer science) for their work in the field underpinning modern AI. Bengio signed the open letter. LeCun mocked it on Twitter and referred to people with my concerns as "doomers". Hinton, who recently told CBS News that his timeline to AGI had shortened, conceivably to less than five years, and that human extinction at the hands of a misaligned AI was "not inconceivable", was somewhere in the middle.

A statement from the Distributed AI Research Institute, founded by Timnit Gebru, strongly criticised the letter and argued that existentially dangerous God-like AI is "hype" used by companies to attract attention and capital and that "regulatory efforts should focus on transparency, accountability and preventing exploitative labour practices". This reflects a schism in the AI community between those who are afraid that potentially apocalyptic risk is not being accounted for, and those who believe the debate is paranoid and distracting. The second group thinks the debate obscures real, present harm: the bias and inaccuracies built into many AI programmes in use around the world today.

My view is that the present and future harms of AI are not mutually exclusive and overlap in important ways. We should tackle both concurrently and urgently. Given the billions of dollars being spent by companies in the field, this should not be impossible. I also hope that there can be ways to find more common ground. In a recent talk, Gebru said: "Trying to 'build' AGI is an inherently unsafe practice. Build well-scoped, well-defined systems instead. Don't attempt to build a God." This chimes with what many alignment researchers have been arguing.

One of the most challenging aspects of thinking about this topic is working out which precedents we can draw on. An analogy that makes sense to me around regulation is engineering biology. Consider first "gain-of-function" research on biological viruses. This activity is subject to strict international regulation and, after laboratory biosecurity incidents, has at times been halted by moratoria. This is the strictest form of oversight. In contrast, the development of new drugs is regulated by a government body like the FDA, and new treatments are subject to a series of clinical trials. There are clear discontinuities in how we regulate, depending on the level of systemic risk. In my view, we could approach God-like AGI systems in the same way as gain-of-function research, while narrowly useful AI systems could be regulated in the way new drugs are.

A thought experiment for regulating AI in two distinct regimes is what I call The Island. In this scenario, experts trying to build God-like AGI systems do so in a highly secure facility: an air-gapped enclosure with the best security humans can build. All other attempts to build God-like AI would become illegal; only when such AI were provably safe could they be commercialised "off-island".



This may sound like Jurassic Park, but there is a real-world precedent for removing the profit motive from potentially dangerous research and putting it in the hands of an intergovernmental organisation. This is how Cern, which operates the largest particle physics laboratory in the world, has worked for almost 70 years.

Any of these solutions are going to require an extraordinary amount of coordination between labs and nations. Pulling this off will require an unusual degree of political will, which we need to start building now. Many of the major labs are waiting for critical new hardware to be delivered this year so they can start to train GPT-5 scale models. With the new chips and more investor money to spend, models trained in 2024 will use as much as 100 times the compute of today's largest models. We will see many new emergent capabilities. This means there is a window through 2023 for governments to take control by regulating access to frontier hardware.

In 2012, my younger sister Rosemary, one of the kindest and most selfless people I've ever known, was diagnosed with a brain tumour. She had an aggressive form of cancer for which there is no known cure and yet sought to continue working as a doctor for as long as she could. My family and I desperately hoped that a new lifesaving treatment might arrive in time. She died in 2015.

I understand why people want to believe. Evangelists of God-like AI focus on the potential of a superhuman intelligence capable of solving our biggest challenges — cancer, climate change, poverty.

Even so, the risks of continuing without proper governance are too high. It is striking that Jan Leike, the head of alignment at OpenAI, tweeted on March 17: "Before we scramble to deeply integrate LLMs everywhere in the economy, can we pause and think whether it is wise to do so? This is quite immature technology and we don't understand how it works. If we're not careful, we're setting ourselves up for a lot of correlated failures." He made this warning statement just days before OpenAI announced it had connected GPT-4 to a massive range of tools, including Slack and Zapier.

Unfortunately, I think the race will continue. It will likely take a major misuse event — a catastrophe — to wake up the public and governments. I personally plan to continue to invest in AI start-ups that focus on alignment and safety or which are developing narrowly useful AI. But I can no longer invest in those that further contribute to this dangerous race. As a small shareholder in Anthropic, which is conducting similar research to DeepMind and OpenAI, I have grappled with these questions. The company has invested substantially in alignment, with 42 per cent of its team working on that area in 2021. But ultimately it is locked in the same race. For that reason, I would support significant regulation by governments and a practical plan to transform these companies into a Cern-like organisation.

We are not powerless to slow down this race. If you work in government, hold hearings and ask AI leaders, under oath, about their timelines for developing God-like AGI. Ask for a complete record of the security issues they have discovered when testing current models. Ask for evidence that they understand how these systems work and their confidence in achieving alignment. Invite independent experts to the hearings to cross-examine these labs.

If you work at a major lab trying to build God-like AI, interrogate your leadership about all these issues. This is particularly important if you work at one of the leading labs. It would be very valuable for these companies to co-ordinate more closely or even merge their efforts. OpenAI's company charter expresses a willingness to "merge and assist". I believe that now is the time. The leader of a major lab who plays a statesman role and guides us publicly to a safer path will be a much more respected world figure than the one who takes us to the brink.

Until now, humans have remained a necessary part of the learning process that characterises progress in AI. At some point, someone will figure out how to cut us out of the loop, creating a God-like AI capable of infinite self-improvement. By then, it may be too late.

Follow @FTMag on Twitter to find out about our latest stories first

**The first table in the article has been amended since publication to reflect the fact that the producers of some of the most powerful AI models, including GPT-4, are no longer disclosing the size or details of their training datasets*

Letters in response to this article:

Have you met my artificial intelligence PA? / From Rebecca Gorman, Founder and Chief Executive, Aligned AI Oxford, UK

I's God-like power is a Big Tech narrative that needs calling out / From Mhairi Aitken, Ethics Fellow, Public Policy Programme, The Alan Turing Institute, London NW1, UK
