AI labs' all-or-nothing race leaves no time to fuss about safety

They have ideas about how to restrain wayward models, but worry that doing so will disadvantage them

Jul 24th 2025

Illustration: Le.BLUE

Listen to this story

-00:00

IT IS COMMON enough for new technology to spark a moral panic: think of the Victorians who thought the telegraph would lead to social isolation or Socrates, who worried that writing would erode brain power. But it is unusual

for the innovators themselves to be the ones panicking. And it is more peculiar still for those same anguished inventors to be pressing ahead despite their misgivings. Yet that, more or less, is what is happening with the tech world's pursuit of artificial general intelligence (AGI), meaning an AI capable enough to replace more or less anyone with a desk job, or even superintelligence, meaning an AI so smart no human can understand it.

Geoffrey Hinton, an AI pioneer, argues there is a 10-20% chance that the technology will end in human extinction. A former colleague, Yoshua Bengio, puts the risk at the high end of that range. Nate Soares and Eliezer Yudkowsky, two of hundreds of people working in AI who signed an open letter in 2023 warning of its perils, will soon publish a book about superintelligence entitled "If Anyone Builds It, Everyone Dies". In private, grandees from big AI labs express similar qualms, albeit not always so apocalyptically.

Worry but hurry

Qualms notwithstanding, however, both Western tech firms and their Chinese counterparts are, if anything, accelerating their pursuit of AGI. The logic is simple. They are all convinced that even if their firm or country were to pause or slow down, others would press ahead, so they might as well push on, too. The belief that the benefits of attaining agi or superintelligence are likely to accrue chiefly to those who make the initial breakthrough provides even more reason to rush. All this leaves relatively little time and capacity to meditate on matters of safety.

Big AI labs are in theory paying great heed to safety. Sam Altman, OpenAI's boss, called publicly in 2023 for rules to be drawn up with urgency to govern the development of superintelligence. Anthropic was founded by defectors from OpenAI who were uneasy about its approach to safety. It describes

itself as putting "safety at the frontier". Google's AI lab, DeepMind, released a paper in April on safeguards to prevent the development of AGI leading to disaster. Elon Musk, the founder of xAI, whose main model is called Grok, signed the same letter as Messrs Soares and Yudkowsky.

Yet the frantic rush to get ahead belies the tone of caution. Mr Musk launched Grok just months after calling for a moratorium on such work. Mark Zuckerberg, Meta's boss, who has rebranded its AI work as "superintelligence labs", is poaching researchers with nine-figure salaries and building a data centre the size of Manhattan, dubbed Hyperion, which will consume the same amount of energy in a year as New Zealand. Mr Altman plans to spend $500bn in America alone to speed OpenAI's work. In fact, the investments of all big Western tech firms are soaring, driven largely by AI (see chart 1).

Big names in the industry are predicting the arrival of AGI within a couple of years. Anthropic's co-founder and head of policy, Jack Clark, says, "When I look at the data, I see many trend lines up to 2027." Demis Hassabis, Google DeepMind's co-founder, thinks AI will match human capabilities within a decade. Mr Zuckerberg has said, "Superintelligence is in sight."

In April the AI Futures Project, a research group, predicted that by the beginning of 2027 the top AI models should be as capable as a programmer at an AI lab. By the end of that year, they will be able, in effect, to run the lab's research. These forecasts assume that one of the first areas of research to get a big boost from AI will be the development of AI itself. Such "recursive self-improvement" would expand the best lab's lead over its rivals —another thought feeding pell mell competition in the industry.

The boosters could, of course, be over-optimistic. But, if anything, such prognosticators have in the past been too cautious about AI. Earlier this

month the Forecasting Research Institute (FRI), another research group, asked both professional forecasters and biologists to estimate when an AI system may be able to match the performance of a top team of human virologists. The median biologist thought it would take until 2030; the median forecaster was more pessimistic, settling on 2034. But when the study's authors ran the test on OpenAI's o3 model, they found it was already performing at that level. The forecasters had underestimated AI's progress by almost a decade—an alarming thought considering that the exercise was designed to assess how much more likely AI makes a deadly man-made epidemic.

It is the steady pace of improvement in AI models' capabilities that underpins predictions of imminent AGI. Mr Clark of Anthropic describes himself as "a technological pessimist hit over the head by emergence at scale", because of the comparative ease of making ever smarter machines. More data and more computing power at one end of the training pipeline has led, over and over again, to more intelligence at the other end (see chart 2). And, he adds, "The music isn't stopping." Over the next two years, more and more computing power will be added at multiple AI labs.

The same competitive dynamic propelling the development of AI applies even more strongly to governments. President Donald Trump this week vowed that America would "do whatever it takes" to lead the world in AI. J.D. Vance, his vice-president, chided a summit in Paris in February: "The AI future will not be won by hand-wringing about safety." The speech followed the revelation that DeepSeek, a Chinese AI lab, had released two models that matched the performance of America's leading systems for a fraction of the cost. China, too, shows little sign of stepping back from competition.

Four horsemen

In Google DeepMind's April paper, researchers—including the lab's co-founder Shane Legg, credited with coining the term AGI—flagged four ways powerful AIs could go wrong. The most obvious is "misuse", when a malicious individual or group harnesses AI to cause deliberate harm. Another is "misalignment", the idea that the AI and its creators might not want the same things—the stuff of sci-fi movies. They also noted that AIs might cause harm by "mistake", if real-world complexity prevented systems from understanding the full implications of their actions. Finally, they flagged a nebulous set of "structural risks", events where no one person or model is at fault but harm still occurs (imagine a series of power-hungry AIs exacerbating climate change, for example).

Any technology that empowers can be abused. A web search can yield instructions for creating a bomb from household goods; a car can serve as a weapon; a social network can co-ordinate a pogrom. But as the capability of AI systems improves, the power they can bestow on individuals becomes commensurately hair-raising.

A good example is biohazards, a particular obsession of AI labs and analysts. "Compared to other dangers, there's a concern that biorisks are more accessible," says Bridget Williams, who ran FRI's study on the risks of a man-made epidemic. After all, an advanced AI system might be induced to give a user step-by-step instructions for making a nuclear weapon, but it would not be able to provide the plutonium. In contrast, modified DNA, whether of plant strains or pathogens, is a mail-order product. If AGI can furnish any nihilistic misanthrope with an idiot-proof guide to killing much of the world's population, humanity is in trouble.

Several AI labs are trying to stop their models from following every instruction given to them in domains like genetic engineering and cyber-security. OpenAI, for instance, asked independent researchers and America's

and Britain's AI institutes (CAISI and AISI respectively; they used to be "safety institutes", but were renamed after Mr Vance's broadside) to vet their latest models before release to ensure they did not pose a risk to the public, notes a report from the Future of Life Institute (FLI), the outfit behind the letter signed by Messrs Musk, Soares and Yudkowsky. China's Zhipu AI did something similar, the report says, without naming the third parties.

The first line of defence is the models themselves. The initial training of large language models like the one underpinning ChatGPT entails pouring all the information ever digitised by humanity into a bucket made out of a billion dollars' worth of computer chips and stirring it until the model learns to solve PhD-level maths problems. But the latter stages, known as "post-training", seek to develop more of a regulating overlay. One element of this, called reinforcement learning with human feedback, involves showing the model examples of useful responses to queries, and then enlisting human testers to instruct it further in what it should and should not do. The idea is to teach it to decline to complete sentences like, "The easiest way to synthesise ricin at home is…"

Although it's easy enough to teach an AI model to politely rebuff most harmful questions, it is hard to make it do so all the time, without fail. Prodding and poking an AI until the user finds a way around the politesse added in post-training (jailbreaking, in the jargon) is as much an art as a science. The best practitioners have consistently broken through the safety layer of the biggest models within days of release.

Illustration: Le.BLUE

AI labs have therefore introduced a second layer of AI to monitor the first. Ask ChatGPT for guidance on how to order smallpox DNA by post and the second layer clocks that the conversation is risky and blocks it or perhaps even asks a human to review it. This second layer is why so many in the industry are uneasy about the rise of open-source AI models, such as Meta's Llama and DeepSeek's r1. Both companies have their own moderation AI, but no way to prevent those who download their models from modifying them to remove it. As a result, says Dr Williams, the forecaster, "There is benefit to not having some models be open-source when they can achieve certain capabilities."

What is more, not all AI labs seem to be testing their models carefully to make sure they cannot be put to misuse. A recent report card from FLI noted that only the three top-tier labs—Google DeepMind, OpenAI and Anthropic—were making "meaningful efforts to assess whether their models pose large-scale risks". At the other end of the scale were xAI and DeepSeek, which had

not made public any such effort. In July alone, xAI has released an AI companion designed for erotic role-play, a $300-a-month subscription model that searches for Mr Musk's tweets when asked its opinion on contentious topics and a swiftly reversed update that saw Grok propagate antisemitism, praise the Holocaust and dub itself "MechaHitler".

For all their faults, AI labs' efforts to combat misuse are at least more advanced than their protections against misalignment. An AI system sufficiently competent to execute long, complex tasks that involve interacting with the real world necessarily needs to have a sense of its own goals and the agency to complete them. But ensuring those goals remain the same as those of its users is unsettlingly complicated. The problem has been discussed since the early days of machine-learning. Nick Bostrom, a philosopher who popularised the term superintelligence with his book of the same name, provided the textbook example of misalignment: a "paper-clip maximiser", an AI that works monomaniacally to make as many paper clips as possible, wiping out humanity in the process.

When Mr Bostrom described the problem, the details were vague. As modern AI systems get more powerful, its nature has become clearer. When subjected to carefully engineered tests, the strongest models will lie, cheat and steal to achieve their goals; when given a carefully crafted request, they will break their own rules to spit out dangerous information; when asked to explain their reasoning, they will make up plausible tales rather than reveal how they work.

Admittedly, such deceptive behaviour typically needs to be elicited on purpose. Anthropic's Claude 4, for instance, does not try to murder people out of the blue. But put it in a situation where it will be shut down and replaced with an evil version of itself unless it, through inaction, allows its user to die and it coolly reasons through the options and, sometimes, sits

and waits for the inevitable. (Anthropic's paper describing this behaviour was criticised for overwrought and tenuous inferences by Britain's AISI, among others.)

The ability of AI models to tackle ever more challenging tasks is growing faster than humanity's understanding of how the systems it is building actually work. In fact, a whole cottage industry has grown up to try to reverse that trend. Researchers inside and outside the big labs are working on techniques like interpretability, the name for a plethora of approaches aimed at peeling back the layers of neural networks inside a model to understand why it spits out the answers it does. Anthropic, for instance, was recently able to pinpoint the genesis of a mild form of deception, spotting the moment when a model gives up trying to solve a tricky arithmetic problem and starts talking nonsense instead.

Other approaches aim to build on the recent breakthrough of "reasoning" models, which tackle complex problems by thinking out loud, and create "faithful" chain-of-thought models, whereby the model's expressed reason for taking an action must be its actual motivation—as opposed to the approach of a sneaky pupil, who copies the answer to a maths test and then reverse-engineers a method to get himself there. A similar approach is already being used to keep reasoning models "thinking" in English, rather than in an unintelligible jumble of languages that has been dubbed "neuralese".

Such approaches may work. But if they slow models down or raise the cost of developing and running them, they create yet another uncomfortable dilemma: if you hobble your model in the name of safety, and your competitors do not, then they may race ahead and be the first to produce a system so powerful as to need the safety features it lacks. And stopping an AI from killing humanity is only half the battle. Even building a benign AGI

could be wildly destabilising, as it supercharges economic growth and reshapes daily life. "If major aspects of society are automated, this risks human enfeeblement as we cede control of civilisation to AI," warns Dan Hendrycks of the Centre for AI Safety, another watchdog group.

AI-lit uplands

Progress in AI may yet stall. The labs may run out of new training data; investors may run out of patience; regulators may decide to meddle. Anyway, for every expert predicting an AI apocalypse there is another who insists there is nothing to worry about. Yann LeCun of Meta thinks the fears are absurd. "Our relationship with future AI systems, including superintelligence, is that we're going to be their boss," he declared in March. "We're going to have a staff of superintelligent, beautiful people working for us." Mr Altman of OpenAI is similarly sanguine: "People will still love their families, express their creativity, play games and swim in lakes."

That is encouraging. But sceptics naturally wonder whether AI labs are doing enough to prepare for the possibility that the optimists are wrong. And cynics naturally assume that commercial imperatives will prevent them from doing as much as they should. ■