America wakes up to AI’s dangerous power

NRC

Should a handful of men be entrusted with the world’s most potent new technology? Five geeks so famous that they can be identified by their first names—Dario, Demis, Elon, Mark and Sam—exercise almost godlike command over the artificial-intelligence models that will shape the future. The Trump administration has stood aside even as those models have gained jaw-dropping capabilities, convinced that unfettered competition between private firms is the best way to ensure America wins the AI race against China.

Until now. Suddenly, America’s free-wheeling treatment of AI looks as if it is coming to an end. The reason is that the models’ dizzying progress also poses a threat to America’s own national security, unnerving members of the Trump administration previously more inclined to worry about overregulation. At the same time, growing resentment among American voters is turning AI into a political lightning-rod. A laissez-faire approach is no longer politically tenable or strategically wise.

De redactie van NRC selecteert de beste artikelen uit The Economist voor een breder perspectief op internationale politiek en economie.

The watershed was Anthropic’s announcement of Claude Mythos on April 7th. The model-maker’s latest creation is so startlingly good at finding software vulnerabilities that, in the wrong hands, it would threaten critical infrastructure, from banks to hospitals. AI models increasingly pose other risks, too, from biosecurity hazards to industrial-scale scamming.

Anthropic’s boss, Dario Amodei, wisely thought Mythos too dangerous for general release. Instead he has reserved it for use by around 50 big firms, in computing, software and finance, so that they can boost their own defences. America’s treasury secretary, Scott Bessent, was so unnerved that he summoned the biggest banks for urgent talks.

It was not the first time the administration had acted. Only weeks ago the Pentagon stepped in after Mr Amodei refused to allow Anthropic’s model to be used in fully autonomous weapons or for mass domestic surveillance. Then, too, the Trump administration was alarmed—because of the power a single firm wielded over a technology central to national security.

A backlash among voters will add to the pressure on the administration to intervene. Opinion polls are leading ever more politicians to think that AI will be one of the big issues in elections in 2028. Americans are far more sceptical of AI than people in other countries. Seven out of ten think AI will hurt job opportunities, a sharp rise from a year ago (and well before they have good evidence). Grassroots opposition to data centres is surging, even though AI has little or nothing to do with rising electricity prices. In a sign of the times, the house of Sam Altman, the head of OpenAI, has been attacked twice in recent days.

History suggests that, with a technology as world-changing as AI, a Mythos moment was inevitable. From John D. Rockefeller to Henry Ford, America’s great industrial innovations were led by a small number of men who grew immensely powerful. Eventually, 20th-century governments stepped in to tame over-powerful industries, from the trust-busting that broke up Standard Oil to the creation of the Federal Reserve and the breakup of AT&T. Those times were at least as polarised and febrile as today’s are. And our calculations suggest that the AI gods are not yet any more dominant than their historical predecessors were.

But history also suggests that controlling AI will be fraught. That is partly because the stakes if things go wrong are so high. It is also because AI is evolving at warp speed.

The trade-offs are acute. Economic growth will benefit from rapidly diffusing AI’s benefits, but the potential backlash could easily lead to overregulation. Doing nothing could leave America vulnerable to malevolent AI-induced chaos, but regulatory overkill would ensure that China wins the AI race. That makes this a perilous moment.

Time is short. Two years ago, during the Biden administration, discussions about regulation were largely about AI’s potential risks. Today its capabilities are already alarmingly powerful and growing more so with every release. The pace of innovation means that debates over the proper role of government, which played out over years, even decades, in the past, now need to be resolved in months.

And the technical hurdles to a more interventionist approach are daunting. Tools of government control, such as nationalisation, are ineffective because talented engineers can move freely between companies and computing power is a commodity. Worse, the leading model-builders are only months ahead of their open-source competitors, including those in China. Sooner or later the capabilities of their models will be available to all.

Even so the Mythos moment could be when a workable scheme to control AI starts to take shape. Trusted users would get early access to the most powerful new models: OpenAI is following Anthropic by rolling out its latest tool to a limited group of vetted cyber-security professionals. Before allowing these models to be broadly commercialised, the government could demand certification from industry-led bodies that have tested them for different uses.

Beware geeks with gifts

This idea has advantages for the big model-builders and the government alike. It avoids the lengthy process of creating a new regulator. By allowing only a few premium users, it enables the model-makers to charge higher prices and limit the use of scarce computing power. Meanwhile, the government can restrict who can use the most powerful models, reducing the risk that China can copy them and catch up faster.

But it also suffers from grave problems. Limited release will reduce competition and increase the clout of entrenched AI companies. It will slow the diffusion of AI's benefits and create a two-tier system within America's economy, disadvantaging the many firms that are repeatedly deprived of privileged early access to powerful new models. What if making AI defences takes a long time or is impossible? What about open-source models? How can you insist that they also follow these rules?

A regulatory system built on these foundations could prove unjust. Insiders could secure themselves against frontier threats; outsiders would have to hope for the best. The opportunities for lobbying and outsize profits would be immense. That would test the honesty and competence of the most openly corrupt administration of America's modern political era. And a fix that concentrates power and wealth yet further among the handful of AI gods risks aggravating the very political backlash that is starting to worry Washington.

Moreover, the Mythos approach can be only half the solution. AI safety cannot be secured nationally. Eventually it will demand international co-operation, starting with China. The new focus on cyber-security also needs to be matched by urgent thinking about the economic and social effects of AI. Dealing with the disruption to jobs and designing an AI-adapted tax system that favours labour are huge problems for which no one yet has good answers. This needs to change. The Mythos moment is a wake-up call for AI safety. It demands hard thinking in other areas, too.