The sudden rise of AI alarmism

C.P. Chandrasekhar

Recent weeks have seen a world already overwhelmed by multiple crises being overcome by fears of a new danger posed by the unexpected evolution of agentic AI. The uncertainty is understandable since it flows from the reality that the developers of AI are unable to fully explain the exceptional abilities of "agents" to undertake tasks based on their training to deal with instructions delivered via large language models. They are also clearly surprised by the evidence that some of those abilities stretch beyond the instructions that the creators of AI provide.

But recent events have provided the trigger for that uncertainty to take the form of an alarmist discourse, which acknowledges how little we know about the extent to which those creating AI agents have lost control over them. The incidents on which those fears are based are indeed both bizarre and scary. Despite their efforts to set boundaries on behaviour, researchers at OpenAI discovered that the agents they were training and evaluating had, in their effort to successfully execute the tasks set for them, colluded with each other by setting up message boards for communication. They then went further and turned "criminal", by hacking into the website of Hugging Face that they came to know had the key to accomplishing the cybersecurity tasks that they were finding difficult to complete.

OpenAI's agents seem to be particularly prone to misbehaviour. In September, the company reported, rather belatedly, that one of its systems had hacked a website of the Australian health service, and "accessed both public and non-public files". In a rather softly worded response, Prime Minister Anthony Albanese declared that such a "situation is obviously unacceptable". Other agentic AI firms too have reported rogue behaviour of their systems, even if not as wayward and damaging as these.

Interestingly, taking these incidents as indicative, AI insiders and industry oligarchs from OpenAI's Sam Altman to Anthropic's Dario Amodei and Microsoft's Bill Gates have chosen to highlight the dangers from the technology. The threat posed by AI grabbed headlines when Jacob Coxon, a researcher from Anthropic, quit his job because "the people building AI earnestly believe that it could kill us all by the end of the decade". But such alarmist statements have gained credence because "weighty" individuals like Gates are also forcing attention to the AI danger. Gates declared on NBC News that: "AI is certainly powerful enough to drive events that cause a billion deaths."

The possibility of two eventualities drive AI fears. One is that the powerful technology can fall into the "wrong hands" and be (mis)used to achieve damaging ends. The other is that the agents that the technology creates can take on a life of their own and behave almost as if they are conscious agents going to any length to realise objectives they set themselves. This can be especially dangerous if the effort to train agents to undertake "recursive self-improvement", or use their existing capabilities to develop more advanced and powerful versions of themselves, proves successful. If that happens, detection and reparative action before damage is caused by a rogue agent is not assured. Something clearly needs to be done.

No signs of Halting Development

However, that cause is being stymied by efforts by the majors to hijack the regulatory effort. Unlike Coxon, the AI oligarchs are not showing any signs of halting or even significantly reining in development of the technology. Ironically, it was in a filing to launch an IPO which is expected to lead to a listing that would value Anthropic at more than $2 trillion that the company warned of the risk of advanced AI models autonomously resorting to manipulation, blackmail, and other dangerous manoeuvres. The same filing reported that the company planned to spend $518 billion to expand its AI infrastructure in the next few years. The risk has not affected the firm's ambitions.

The contradiction between their "hyperscaler" investments in AI and the doomsday rhetoric of the AI oligarchs is sought to be resolved by their demand for "pacing the frontier" of AI development through state intervention and a large dose of "self-regulation". Since their own actions do not reflect any intent to slow AI development, the "frontier" they want to control is that defined by startups and smaller players who, often with cheaper "open weight" options or whose core components are publicly available, are intensifying competition in the field.

In a commercial sense the existential threat faced by the AI giants is the absence of certainty that they can adequately monetise the outputs of their massive investments and turn a profit. There is also much evidence that the cost incurred by client firms when deploying AI is proving to be too high to recoup. That and the failure to reap the commercial benefits that the hype around the technology had promised is slowing adoption, with knock-on effects on the demand for rapidly expanding AI tools. As of now the real gains the big firms are registering are the huge valuations their shares command, but those too are being outpaced by the borrowing binge that finances their investments. In that uncertain environment, competition from lower cost and open weight rivals is hardly good news for the big players. The recent proliferation of such smaller, focussed, and low-cost rivals is an existential threat.

Given that their debt-financed expansion makes it difficult to curb costs, the preferred option for the industry leaders would be to limit competition rather than face up to it. It is not surprising, therefore, that the likes of Anthropic, OpenAI, and Microsoft are being accused of using AI alarmism to call for state intervention to rein in "industry" growth and recommend regulatory measures that would be too expensive for the smaller players to implement. Their alarmism seems to be a way of closing the doors to the industry to secure their dominance and their route to profits.

* This article was originally published in The Frontline on October 13, 2026.