Multiple red flags are not yet slowing the generative AI train

TECHNOLOGY

John Thornhill

Ever since the ancient Greeks dreamt up the myth of Prometheus, humanity has been arguing about the dual nature of technology. The fire that Prometheus stole from the gods could warm humans, but also burn us. So it is with the widespread deployment of artificial intelligence systems today. The champions of AI have long argued that this general purpose technology will produce an unprecedented surge in productivity and creativity; its critics fear it carries alarming present-day risks and may even pose an existential threat to humanity in future.

The release last year of powerful generative AI models, such as ChatGPT and Dall-E 2 developed by OpenAI, has reignited that smouldering debate. More than 100mn users have already experienced the weird and wondrous things these types of generative models can do: achieve near-human levels of recognition and replication of text and images, co-create computer code and produce fake viral photos of the Pope in a white puffer jacket.

In a recent post, Bill Gates, the co-founder of Microsoft turned philanthropist, said he watched in "awe" last September as OpenAI's model aced an advanced biology exam, predicting the technology could bring enormous benefits to the fields of healthcare and education. A research report from Goldman Sachs, published this week, forecast that the widespread adoption of AI could significantly boost labour productivity and increase global annual gross domestic product by 7 per cent.

But the rapid development and increasingly pervasive use of generative AI systems has also alarmed many. Some of Google's own researchers, such as Timnit Gebru and Margaret Mitchell, were among the first to flag the dangers of the company's generative AI models baking in existing societal biases, but they were later fired. This week, in an open letter posted by the Future of Life Institute, more than 1,100 signatories, including several prominent AI researchers, amplified the alarm. They called for a six-month moratorium on the development of leading-edge models until better governance regimes could be put in place. Uncontrolled, these machines might flood the internet with untruths, automate meaningful jobs and even threaten civilisation. "Such decisions should not be delegated to unelected tech leaders," the letter writers said.

> Unless businesses can prove their models align with humanity's interests, they can expect a backlash

At least three threads need to be unpicked amid the controversy. The first, and easiest to dismiss, is the moral panic that accompanies almost every new technology, whether it is steam trains, electricity, motor cars or computers. Even Benjamin Franklin's invention of the seemingly innocuous lightning rod was initially opposed by Church elders fearing it was interfering with the "artillery of heaven". As a rule, it is better to debate how to use commercially valuable technologies appropriately than to curse their arrival.

The second is how commercial interests tend to coincide with moral stances. OpenAI started out in 2015 as a non-profit research lab, promising to collaborate with outside partners to ensure the safe development of AI. But in 2019 OpenAI switched to a capped for-profit model, enabling it to raise venture capital funding and issue stock options to attract top AI researchers. Since then, it has attracted big investments from Microsoft and become more of a closed, commercial entity. That said, at least some of the criticisms come from rivals with an interest in slowing OpenAI's development.

But the third and most important thread is that many serious AI experts, well-acquainted with the latest breakthroughs, are genuinely concerned about the speed and direction of travel. Their concerns are magnified by the trend among some big tech companies, such as Microsoft, Meta, Google and Amazon, to shrink their ethics teams.

As Gates wrote in his post, market forces alone will not tackle societal inequities. Civil society organisations are mobilising fast and some governments are aiming to set clearer regulation. This week, the UK published pro-innovation draft rules on AI, while the EU is drawing up a stiffer directive on controlling the technology's use in high-risk domains. But for the moment these efforts seem little more than waving a small red flag at an accelerating train.

Unless the companies leading the AI revolution can credibly prove their models are designed to align with humanity's best interests, they can expect a far fiercer public backlash. Expert independent institutions with the power to audit AI companies' algorithms, and restrict their use, should be next on the agenda.

The writer is founder of Sifted, an FT-backed site about European start-ups