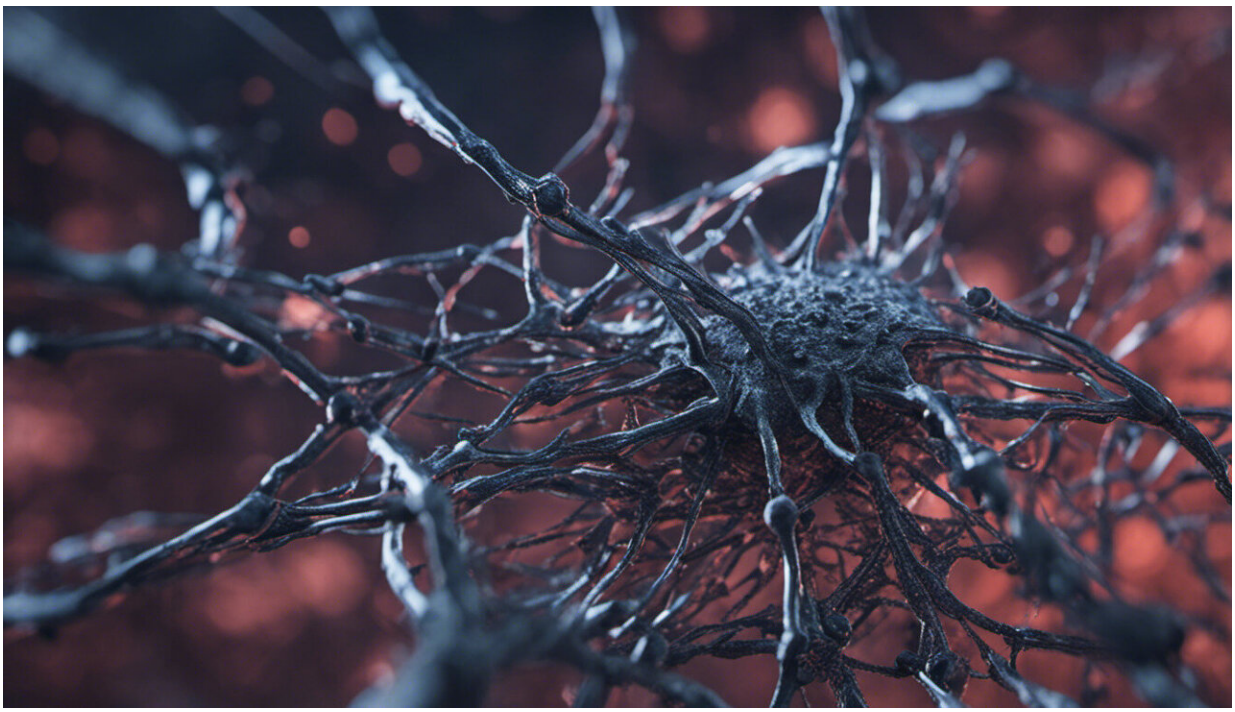
AI developers often ignore safety in pursuit of breakthrough—so how do we regulate them without blocking progress?

March 18 2021, by The Anh Han, Luís Moniz Pereira and Tom Lenaerts



Credit: AI-generated image (disclaimer)

Ever since artificial intelligence (AI) made the transition from theory to reality, research and development centers across the world have been rushing to come up with the next big AI breakthrough.

This competition is sometimes called the "AI race." In practice, though, there are hundreds of "AI races" heading towards different objectives. Some research centers are racing to produce digital marketing AI, for example, while others are racing to pair AI with military hardware. Some races are between private companies and others are between countries.

Because AI researchers are competing to win their chosen race, they may overlook safety concerns in order to get ahead of their rivals. But safety enforcement via regulations is undeveloped, and reluctance to regulate AI may actually be justified: it may stifle innovation, reducing the benefits that AI could deliver to humanity.

Our recent research, carried out alongside our colleague Francisco C. Santos, sought to determine which AI races should be regulated for safety reasons, and which should be left unregulated to avoid stifling innovation. We did this using a game theory simulation.

AI supremacy

The regulation of AI must consider the harms and the benefits of the technology. Harms that regulation might seek to legislate against include the potential for AI to discriminate against disadvantaged communities and the development of autonomous weapons. But the benefits of AI, like better cancer diagnosis and smart climate modeling, might not exist if AI regulation were too heavy-handed. Sensible AI regulation would maximize its benefits and mitigate its harms.

But with the US competing with China and Russia to achieve "AI supremacy"—a clear technological advantage over rivals—regulations have thus far taken a back seat. This, according to the UN, has thrust us into "unacceptable moral territory".

AI researchers and governance bodies, such as the EU, have called for urgent regulations to prevent the development of unethical AI. Yet the EU's white paper on the issue has acknowledged that it's difficult for governance bodies to know which AI race will end with unethical AI, and which will end with beneficial AI.

Looking ahead

We wanted to know which AI races should be prioritized for regulation, so our team created a theoretical model to simulate hypothetical AI races. We then ran this simulation in hundreds of iterations, tweaking variables to predict how real-world AI races might pan out.

Our model includes a number of virtual agents, representing competitors in an AI race—like different technology companies, for example. Each agent was randomly assigned a behavior, mimicking how these competitors would behave in a real AI race. For example, some agents carefully consider all data and AI pitfalls, but others take undue risks by skipping these tests.

The model itself was based on evolutionary game theory, which has been used in the past to understand how behaviors evolve on the scale of societies, people, or even our genes. The model assumes that winners in a particular game—in our case an AI race—take all the benefits, as biologists argue happens in evolution.

By introducing regulations into our simulation—sanctioning unsafe behavior and rewarding safe behavior—we could then observe which regulations were successful in maximizing benefits, and which ended up stifling innovation.

Governance lessons

The variable we found to be particularly important was the "length" of the race—the time our simulated races took to reach their objective (a functional AI product). When AI races reached their objective quickly, we found that competitors who we'd coded to always overlook safety precautions always won.

In these quick AI races, or "AI sprints," the competitive advantage is gained by being speedy, and those who pause to consider safety and ethics always lose out. It would make sense to regulate these AI sprints, so that the AI products they conclude with are safe and ethical.

On the other hand, our simulation found that long-term AI projects, or "AI marathons," require regulations less urgently. That's because the winners of AI marathons weren't always those who overlooked safety. Plus, we found that regulating AI marathons prevented them from reaching their potential. This looked like stifling over-regulation—the sort that could actually work against society's interests.

Given these findings, it'll be important for regulators to establish how long different AI races are likely to last, applying different regulations based on their expected timescales. Our findings suggest that one rule for all AI races—from sprints to marathons—will lead to some outcomes that are far from ideal.

It's not too late to put together smart, flexible regulations to avoid unethical and dangerous AI while supporting AI that could benefit humanity. But such regulations may be urgent: our simulation suggests that those AI races that are due to end the soonest will be the most important to regulate.

Provided by The Conversation