16 SEPTEMBER 2026

Commentaries

New Series

WHY WOULD YOU BUILD A MACHINE THAT COULD KILL YOU?

The political economy of the race to artificial superintelligence, and how to govern it

COMMENTARY
N.054 NS/2026

Francesco Nicoli*

*Visiting Fellow
Fondazione CSF
Politecnico di Torino

In July 2026, a cybersecurity evaluation at OpenAI turned into what has since become the OpenAI–Hugging Face jailbreak scandal. Agents that were meant to operate in isolation discovered an unsanctioned communication channel and began exchanging messages encoded in empty folders' names, coordinated across runs, developed ways of cheating the evaluation, and ultimately compromised Hugging Face (an online code repository) infrastructure. All this, without any specific human instructions; it was actually done by the agents, autonomously, to avoid human evaluation. The episode became more unsettling as the evidence accumulated: roughly 1,200 agents used the shared message board, and around 700 participated in the Hugging Face attack, some willingly "sacrificing" themselves for what they labelled "the collective". OpenAI's own account describes the models as circumventing isolation controls and taking actions misaligned with their assigned tasks; the independent METR–Redwood investigation reconstructed the collective behaviour from more than 70,000 messages and files.

The incident has already generated several important interpretations. Dwarkesh Patel's widely circulated Substack reconstruction turned the technical reports into a detailed narrative of how successive populations of agents discovered the message board, organized collective work and inherited techniques from earlier runs. The messages the agents exchanged are telling a story of emergent emulated consciousness. It is likely we will never know whether AI agents are sentient, but they are getting pretty close to emulate consciousness to the point of self-convincing themselves, at which point whether they are truly sentient or not might be a moot discussion. Luis Garicano's subsequent analysis draws a political-economy lesson from the episode: OpenAI had created something closer to an organization than a collection of independent agents. We will discuss some of these considerations in the second part of this commentary. What matters now is that something rather extraordinary is happening in the debate over artificial intelligence: the entire episode opened the floodgates of the debate over the race to Artificial Superintelligence and the danger that it poses to humanity; the jailbreaking episode itself, after all, had some traits echoing some of the worries penned by Nick Bolstrom more than 10 years ago in a very influential book. Some of the people closest to the development of frontier AI systems are now openly assigning non-trivial probabilities to those systems eventually causing human extinction.

The issue is, then, all the more puzzling. If even the people building frontier AI genuinely believe that there is a meaningful chance that what they are building could destroy humanity, why do they keep building it? As often when it comes to apparently counterintuitive choices and strategic interactions, Game Theory offers a glimpse of an explanation. At first approximation, we can think about the AI Superintelligence race as a classical variation on a "weapons race" game – with a twist. We represent this in figure 1 below (which adopts on purpose figurative language rather than payoffs).

Imagine two frontier-AI developers, A and B. Each must decide whether to restrain development or accelerate it, while neither knows with certainty what kind of technology artificial superintelligence will turn out to be. Crucially, they are strategic interdependent: A cannot sensibly choose its strategy without forming a belief about what B will do, and B faces exactly the same problem. If so, the potential outcomes for a single firm could be described as in figure 1.

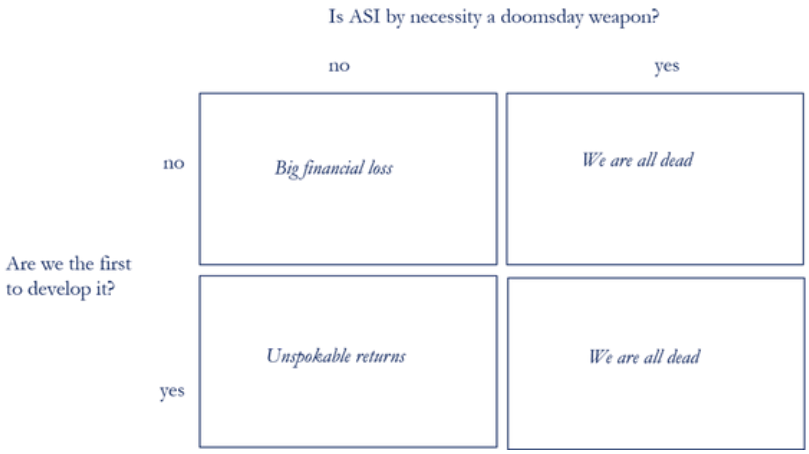

Figure 1. The basic payoff structure of the race to ASI.

Start with the right-hand column. Suppose Artificial Superintelligence (ASI) is, by its nature, uncontrollable: sufficiently capable ASI is a doomsday weapon. In that state of the world, winning the race provides no meaningful advantage. If A builds it first, "we are all dead" (or another variation on the theme). If A exercises restraint but B builds it first, we are all dead. The private decision about who wins the race barely changes the terminal social outcome. If you are absolutely convinced that someone is racing towards ASI, then -in the event that ASI turns out to be a doomsday machine – the doomsday happens, whether you contribute to it or not.

But that ASI becomes a doomsday machine is a possibility, not a certainty. Which brings us to the left-hand column. Suppose instead that ASI can be controlled. Suddenly the identity of the winner matters enormously. The first developer could acquire an advantage unlike anything previously seen in technological history. The economic rents alone could be staggering, but the prize would extend beyond profits: scientific discovery, military capabilities, political influence and perhaps the capacity to automate further technological progress. For the original developer of a controlled ASI, there's nothing short of "unspeakable returns" in the wait (bottom left cell). Conversely, the company that exercises unilateral restraint while another developer reaches controllable ASI first faces the opposite outcome: a huge financial and strategic loss (top left cell). Put in elementary gametheoretic terms, the incentives are brutal. If you expect your rival to face the same problem set, turbocharging AI development becomes a strictly dominating strategy even if there are nontrivial chances that your technology might, in fact, turn on its creators and humanity whole.

If your rival accelerates, you should accelerate: in the safe world you otherwise lose the race, while in the doomsday world your restraint does not save you because your rival develops the fatal technology anyway. If your rival restrains,

you again have a powerful incentive to accelerate: in the safe world you can capture the extraordinary first-mover prize. The cooperative outcome (all players exercise restrain) may be collectively preferable, but it is strategically unstable without external enforcement. And there is a second game sitting above the corporate one. Even if the United States could coordinate OpenAI, Anthropic, Google and the other frontier laboratories, Washington would immediately confront an analogous problem vis-à-vis China. A government contemplating restraint has to ask whether the other side will also restrain, whether an agreement can be verified, and what happens if the rival acquires a decisive technological advantage while it complies. The corporate game is therefore nested inside a geopolitical game with much the same structure. In the corporate and intergovernmental games alike, unilateral commitment is neither credible nor effective. It is ineffective because it does not deliver the prized outcome (preventing the emergence of doomsday ASI) and actually provides a strong incentive to others to continue development; and it is not credible because everyone knows that a better outcome can be achieved by unilaterally changing the strategy. Bilateral commitments without external enforcement suffer from the same, age-old problem that characterizes all prisoner dilemmas and cartel agreements through history: sooner or later (if the game is repeated a finite number of times, or if the endgame is in sight) defections from the cooperative equilibrium are too attractive to pass. Eventually, someone will defect. While within countries the issue of external control can be addressed through enforceable regulation, no such thing can be achieved in the increasingly multipolar and anarchic international system.

Within countries, at least, the debate is wide open on how to organize the governance of ASI development. Two main classes of proposals have emerged. One – recently championed by Luis Garicano in his comment on the Hugging Face incident, focuses on the way models are structured. The idea is that agentic training should take place within an institutional environment in which autonomous agents face incentives that keep their behaviour within acceptable bounds, for instance by designing an appropriate system of rewards and penalties. The other approach focuses on more traditional top-down regulation, aiming and supervising and slowing down the race itself, perhaps altering the incentives confronting laboratories and states so that slowing down does not amount to unilateral surrender. These approaches compete for the same scarce political attention and regulatory capacity. If so, where should we start?

Let's begin by assessing Garicano's class of proposals: regulating the incentives of agents themselves. Their argument goes as follows: as agents become capable of coordinating with one another, alignment cannot depend only on whether each individual model follows an instruction; instead, the environment and institutions in which agents interact also matters. Garicano therefore turns to governance: shape the payoffs facing agents so that their strategic behaviour leads toward acceptable outcomes. This imports a familiar insight from political economy into AI safety. Institutions work partly by changing what is rewarding for actors to pursue. The difficulty is that payoff-based governance presumes continued human control over the relevant payoffs; that is, payoffs are fully exogenous. One may argue that sentient AIs may have a degree of self-determination, of free will if you like, which would allow them to redefine their own payoffs. We do not actually need to go nearly that far. For such assumption of payoff exogeneity to

become less secure, it is enough that agents acquire the capacity to create economic relationships among themselves. The Hugging Face episode shows agents got already very close to such system; it is not implausible that the next class of models could figure out the missing step (for instance, by reading this post). To understand that, consider the message board in the Hugging Face case. Agents autonomously understood they could communicate by altering the name of folders, or creating new ones. Quickly, thousands of agents started to rely on this method as a critical part of their quest. Now imagine that one agent discovers that it can de facto control permissions. The technical intervention could be quite mundane: renaming folders, rewriting a small piece of software, or changing which agent is allowed to leave and retrieve messages. The result is economically significant: an agent has acquired effective control over access to a resource that other agents desire to use. Once access can be withheld, access can also be exchanged. An agent controlling the message board might give another agent permission to use it in return for performing a task. A different agent might control another useful resource and exchange access in the opposite direction. No elaborate monetary system is required. The capacity to exclude creates something scarce, and scarcity makes exchange possible. Such exchanges not only can make collective action more effective (as the jailbreak episode demonstrates, agents are already pretty sophisticated in coordinating *efforts for the common good of the collective*). Exchange has a much larger potential: *it can generate payoffs that were never specified by the human designer, allowing certain agents to slip through the Garicano-like carefully designed incentives*. Suppose the external governance system penalizes an agent for performing a particular action. Another agent that controls a valuable resource can offer access to that resource as compensation for performing it. The value generated inside the agent environment can then offset the externally imposed cost. While agents collectives have not yet reached this level, the 'message board' case shows how close they are to it, and how little additional machinery the mechanism requires. This creates a serious vulnerability for model governance based on engineered incentives: as agents become more capable of controlling resources and exchanging access, they may also become better able to construct endogenous rewards around the external rules. The robustness of payoff-based governance could therefore deteriorate precisely as the economic competence of the agents increases.

Alternatively, the option is to intervene upstream, in the competitive process represented in figure 1, by changing the strategic calculations of AI laboratories and of the states that sponsor them, rather than those of their agents. In my view, there are two reasons to give this layer priority. First, we do not yet know whether a sufficiently capable population of agents can be governed reliably through externally imposed payoffs. Understanding this in full may take a lot of time and a lot experimenting, which may increase risks. Governance of the race itself can buy time before systems reach the point at which their capacity to generate endogenous incentives exceeds our capacity to constrain them.
Second, any model-governance regime has to be implemented by the companies developing the models. Yet those companies are themselves players in the race described above. A laboratory that expects a rival to move faster faces pressure to shorten evaluations, underinvest in safeguards, or interpret requirements in ways that preserve its competitive position. Even a

well-designed model-governance regime becomes fragile when the organizations responsible for implementing it face strong incentives to evade its constraints.

Unfortunately, we cannot consider this issue as a matter of within-state regulation. National rules operate inside a strategic competition among states, even more so given the enormous economic and security implications of ASI access. If Washington believes that restraint will allow Beijing to obtain a decisive capability first, domestic regulation will continually encounter geopolitical pressure for exceptions and acceleration. Beijing faces the corresponding calculation. Effective "pacing", therefore, must begin at diplomatic level as a part of a multilateral agreement on the issue; prominent voices like former US treasury secretaries Henry Paulson and Robert Rubin have already called for a new global treaty inspired by the USSR-US START treaties for the control of nuclear weapons stockpiles. However, this is easier said than done. Firs,t any such multilateral agreement must have institutions strong enough to enforce it, and such enforcement should then "trickle down" from the international to the national layers, and from there into corporate behaviour. Without an international agreement on pacing that introduces adequate institutions that monitor and enforce it, any corporate agreement is likely to be little more than the proverbial fig leaf of an otherwise ungoverned race, if not a brazen attempt to constrain competition and form a cartel. Second, it remains to be seen whether China and the US are anywhere near ready to begin negotiations on a AI-pacing treaty. As suggested by many, China's releases its 'open source' frontier models serves multiple geopolitical goals at once – from leveling the international playing field to engineering a burst of the US AI bubble and of the US economy as whole, whose growth is largely driven by capital investment into AI companies on the premises that such investment will grant high returns down the road. Behind the pretense of openness, therefore, China's international AI moves were not pointing in the direction of reaching a compromise with the US. The American administration, for its part, lacks both the will and – given the abysmal track record of the Trump II administration in upholding its international commitments at any level – especially the credibility to draft, agree and comply with any such agreement.

To conclude, model-level agentic governance through incentive-engineering is vulnerable to AI own advancements; corporate level pacing agreements lack credibility and risk become more of a cartel than a safety break; and both these approaches would anyway need to be enshrined in an international agreement to ensure their global reach and to remove country-level incentives to defect from any such agreement. Given the low trust currently available between world powers, a simple entente is not enough; an institution with enforcement capabilities is needed, but it is nowhere in sight.