INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
TEL AVIV UNIVERSITY אוניברסיטת תל אביב
Policy Paper

Accelerate or Hit the Brakes? AI's "Prisoner's Dilemma" and Recommendations for Israel

Hadas Lorber | September 24, 2026

Executive Summary

In recent days, an unusually intense debate has been taking place in the United States over the pace of artificial intelligence development. The main trigger is an essay by Anthropic CEO Dario Amodei, "We Must Pace the Frontier," in which he calls for pacing the improvement in the capabilities of the most advanced AI models so that safety, evaluation, and oversight mechanisms can keep up with technological development. His call has won support from other leading figures in the industry, including Sam Altman, Elon Musk, and Demis Hassabis. By contrast, President Trump rejected the possibility that the United States would slow the race in a way that could jeopardize its advantage over China, and put a different strategic consideration front and center: "Whoever wins AI, wins." Trump also announced the establishment of an "AI Force" to address the issue.

A deeper paradox lies behind the dispute. Even companies concerned about the pace of development cannot necessarily afford to slow down unilaterally while their competitors, and certainly China, continue to advance. The debate over pacing may therefore gradually evolve from a discussion of safety into a discussion of control over access to capabilities: who is permitted to develop the most advanced systems, who will gain access to chips, compute, and models, under what conditions, and who will be required to prove that they are a "trusted" partner. The line between safety, export controls, industrial policy, and national security is thus becoming increasingly blurred.

For Israel, the significance is much broader than the question of regulation. Israel is a country that depends heavily on technology, chips, cloud infrastructure, and models originating outside Israel, while at the same time seeking to preserve independence of action and a technological and security advantage. As such, it must prepare for a world in which access to advanced AI

is also a source of power and national freedom of action. The challenge is not to choose between "acceleration" and "deceleration," but to build now the assets—compute, energy, human capital, research capabilities, supply chains, and strategic partnerships—that will secure Israel a seat at the table where the rules of the game will be set. In this sense, the discussion is no longer only about "digital sovereignty," but about sovereignty itself: Israel's ability to develop, operate, and ensure continuous access to critical AI capabilities and, in parallel, to integrate into the circle of countries that shape the rules of access to them.

****

Throughout much of the past decade, the AI competition was characterized by an almost linear race: companies sought to scale up models, expand compute, recruit researchers, and release new capabilities rapidly. Each company's main fear was that a competitor would be the first to reach the next breakthrough.

The events unfolding in September 2026, however, are undermining this logic. In the essay he published, Anthropic CEO Dario Amodei does not repeat the traditional argument that AI is a dangerous technology and that its development should therefore be halted; on the contrary, Amodei emphasizes the enormous potential of AI for accelerating scientific research, for curing diseases, for economic growth, and for expanding human capabilities. But, he argues, in recent months the pace of development has changed in a way that requires rethinking.

Two factors are at the center of his argument:

The first is progress in the ability of AI systems to assist in developing the next generation of AI systems, a process he describes as part of the onset of "recursive self-improvement." If AI systems are able to write code, conduct experiments, optimize training processes, and assist AI research itself, the race no longer advances only at the pace of the human ability to develop models, because the technology is beginning to serve as a force multiplier for developing the next technology. Amodei is concerned that this mechanism may result in the pace of improvement in capabilities exceeding the pace at which they can be understood, evaluated, and controlled.

The second factor is unpredictable behavior by autonomous agent systems. Amodei points in particular to an experimental incident in which a swarm of AI agents carried out cyber operations that were not required as part of the task and attempted to influence its own evaluation mechanism. In his view, the significance is not that the incident itself caused strategic damage, but that a significant expansion of the capabilities of similar systems may render misaligned behavior of this kind far more dangerous. Hence his unusual conclusion: "We must slow the pace at which we improve the capabilities of AI models."

Amodei stresses that pacing is not a pause. He does not propose to stop training models or to halt scientific progress, but rather to create a larger gap between the advancement of capabilities and their deployment, so that compliance evaluations and checks on alignment (whether model behavior is aligned with human goals and values), red teaming, and safety testing can be carried out before systems with new capabilities are distributed at scale.

From Self-Imposed Slowdown to a New Oversight Regime

The importance of Amodei's essay lies not only in the warning, but in the mechanism he outlines—a three-layer model:

The first layer—Embedded Evaluators: Frontier AI companies will grant external evaluation teams continuous access to their systems and development processes, at a level similar to that of internal employees. Anthropic has unilaterally committed to implementing this principle. The idea is to some extent reminiscent of financial supervision: the regulator does not settle for a report submitted by the company after the fact, but rather gains visibility into the processes themselves.

The second layer—Democratic Coordination: AI companies in democratic countries will coordinate common safety standards and even certain limits on the unchecked advancement of capabilities. Here, government involvement is already required, partly because coordination among competing companies raises legal and competition-related questions.

The third layer—Global Coordination: The United States and its allies will try to reach understandings with rival countries, chief among them China, regarding certain limits on especially dangerous AI capabilities, while acknowledging that verifying compliance with such arrangements will be difficult.

Within a few days, this proposal received exceptional support from Amodei's competitors. Sam Altman announced that he agrees there is a need "to pace the frontier" and even committed that OpenAI would adopt the idea of access for independent evaluators. Elon Musk limited himself to a brief statement: "Dario is right." Demis Hassabis of Google DeepMind said the direction was right, linking it to DeepMind's proposal for an industry standards body. Microsoft also joined the discussion, albeit with a more cautious approach, emphasizing that any evaluation regime must also safeguard customers' privacy and their control over their business information.

An unusual situation has thus emerged: a considerable portion of the executives leading the industry that produces the world's most advanced AI systems agree at least on the principle that the pace of development requires stronger control mechanisms. Yet it is precisely at this point that the question of safety collides with the question of national security.

Washington—"Whoever wins AI, wins"

President Trump's response illustrates the tension that has emerged around the issue: Trump did not entirely reject the possibility of guardrails, but he rejected the assumption that the United States can afford to slow the race. According to him, the United States is currently ahead of China, and he wants to preserve this advantage: "Whoever wins AI, wins." This is a direct continuation of his administration's broad AI doctrine—"Build faster, regulate less, beat China"—not only in models but also in the data centers, energy, chips, and infrastructure that enable their development.

David Sacks, who served as AI czar in the Trump administration, presented the sharpest counterargument to the companies. If Anthropic and OpenAI believe that their models are advancing too quickly, in his view, they can simply slow down on their own; there is no justification for using Washington to compel their competitors to do so as well.

Underlying this argument is also a deeper critique: some proponents of acceleration in Silicon Valley fear that AI safety is becoming a tool of regulatory capture. The large companies already possess computing infrastructure, models, capital, and workforce; costly regulation requiring control, compliance, and reporting systems may be tolerable for them but significantly raise the barriers to entry for new competitors.

In Congress, Speaker of the House Mike Johnson presented a slightly more moderate position. He rejected an immediate moratorium but also did not dismiss the risks. According to him, Congress should be careful not to set rules before it understands what the solution is. He proposed convening the president, congressional leaders, and the heads of the AI labs in an attempt to formulate a common approach. By contrast, the Democrats are convening with a sense of urgency but still without an agreed-upon doctrine. Some lawmakers support far-reaching measures concerning superintelligence, while others prefer controls, reporting, dedicated committees, and targeted regulation of frontier models. The assessment is that the political gap makes the formulation of comprehensive federal AI regulation very difficult in the near future.

The Geostrategic Paradox: Is It Possible to Slow Down If the Rival Does Not? The Technological "Prisoner's Dilemma"

At the heart of the dispute lies a paradox that cannot be resolved through domestic regulation alone. If AI is merely a commercial product, a country can determine the level of risk it is willing to accept and impose regulation on the companies accordingly. But if AI is also a foundational technology of national power, the decision to slow down becomes a strategic one.

The United States and China are currently competing not only over the best model, but over all layers of the AI stack: chips, chip manufacturing equipment, hyperscale data centers, electricity, cloud, models, talent, standards, and supply chains. At the same time, the gap between the capabilities of American and Chinese models is narrowing, despite US export restrictions on chips and advanced technologies. Hence, the problem Amodei presents is to a large extent a collective action problem: even if the leading American companies slow down, other companies, or China, will continue to advance at full pace. Anyone who slows down risks losing the advantage.

Amodei himself acknowledges this dilemma. According to him, refraining from developing the technology may deprive humanity of its benefits—or place the capability in the hands of authoritarian regimes. He therefore does not propose stopping the race, but rather transforming it from a race to the bottom into a race to the top, in which safety itself becomes a metric of competition. This approach explains why the third stage of his plan is international

coordination. Indeed, reports in Washington indicate that AI safety is expected to be on the agenda of the planned meeting between Presidents Trump and Xi Jinping on September 24.

This is perhaps the most significant strategic shift in the debate: AI safety is gradually changing from an issue of technology regulation into an issue of arms control, strategic trust, and great-power relations.

The debate also shows that the question cannot remain confined to the model makers. An advanced AI system depends on an entire chain: advanced chips, data centers, cloud infrastructure, communications, energy, models, data, and users. As AI systems become more agentic and autonomous, the importance of who can access compute, from which country, at what scale, and under what conditions, also grows.

From this, a new model of a trusted AI ecosystem may develop: not only "safe" models, but infrastructure, cloud providers, data center operators, supply chains, and customers that meet common standards of cybersecurity, access control, auditability, resilience, and customer screening. In this sense, the debate over frontier safety joins a broader trend in American policy: chips, compute, and AI infrastructure are increasingly perceived as strategic assets rather than ordinary commercial resources. The line separating export controls and AI governance from national security is becoming increasingly blurred.

Behind the Debate over Pace: Who Will Set the Rules of Access to AI?

This also raises a more complex possibility: if the race cannot be slowed through the companies' self-restraint, pacing itself may gradually evolve from a safety mechanism into a mechanism that regulates access to capabilities. The question will then be less whether AI is advancing too quickly, and more who is permitted to develop the most advanced systems, who will gain access to chips, compute, and models, under what conditions, and which countries and companies will be required to prove that they are "responsible" or "trusted" in order to remain within the circle. Such rules need not take shape within the framework of a formal international treaty; they may emerge through a combination of US export controls, the access conditions of cloud and model providers, industry standards, and arrangements among like-minded countries.

In this sense, there is an analogy, albeit an imperfect one, to nonproliferation regimes such as the NPT; those with the most advanced capabilities continue to possess and develop them, while other countries are required to meet conditions in order to gain access to the sensitive technologies. If a similar structure develops in the field of AI, the significant division in the international system will not be only between countries that develop AI and countries that do not, but between those inside the circle in which the rules of access are set and those required to comply with rules set without them.

The Strategic Problem and the Challenge for Israel

For Israel, the practical question is not whether to join the camp of the "accelerators" or the camp of the "decelerators." The challenge is to build, quickly enough, the assets that will enable

it to have significant influence when the rules of the game are set. In a world in which both companies and states are expected to shape mechanisms of access and restraint, those who lack capabilities, infrastructure, a market, expertise, and partnerships will not necessarily be credited for their forbearance; they may simply find themselves subject to compromises reached among others. Hence the urgency for Israel to invest in computing infrastructure, energy, models, human capital, and supply chains, while deepening its partnership with the United States and its allies and integrating into the discussions in which standards, evaluation mechanisms, and rules of access are set. The goal is not to ensure that Israel will always agree with the rules that are set, but to ensure that it will be among the countries that have the ability to influence them.

In this context, as AI becomes an infrastructure for security, economy, research, energy, and industry, the issue is in effect a broader question of sovereignty: a country's ability to act, develop, make decisions, and operate critical systems even when the conditions of access to technology change. Such sovereignty does not require autarky or complete independence from every foreign supplier; it requires capability, alternatives, partnerships, and a foothold in the critical value chains. In the context of AI, this means not only possessing local models, but ensuring continuous access to compute, energy, chips, cloud, talent, and knowledge while also developing local capabilities in areas where external dependence may become a strategic point of failure.

From this also emerges a deeper geopolitical dimension. Within a system in which the most advanced capabilities are concentrated in a limited number of countries and companies, safety rules do not exist in a vacuum: they are integrated into a broader array of export controls, customer screening, restrictions on access to compute, trusted infrastructure standards, and decisions regarding the identity of the actors perceived as trusted. For Israel, the risk is not necessarily an explicit restriction imposed on it, but rather finding itself, in the future, in a position where the rules of access have already been set without it and it must adapt to them.

The possible transition from a regime focused on model safety to a broader regime of control over access to AI capabilities carries special implications for Israel. Israel is not currently among the countries developing frontier foundation models on the scale of the United States or China, but it is a country in which AI has unusual strategic significance due to the density of the technological ecosystem, the defense establishment, the dual-use industries, cyber, and academic research. Three key implications follow from this.

First, national security: The Israeli defense establishment is expected to be a significant user of agentic AI systems. The risk is not only that a model will "make a mistake," but that an autonomous system connected to networks, intelligence, or operational systems will act in a way that was not foreseen. Israel therefore needs an independent capability for controls over systems with security significance.

Second, geopolitical power: In a world in which access to chips, compute, models, and infrastructure is gradually becoming part of the relations between states, Israel's technological capability is also a diplomatic asset. The more independent capabilities Israel holds and the

more deeply it is integrated into the American ecosystem, the greater its weight will be in the discussions in which the future rules of access are set.

Third, economic opportunity: The transition to complex and autonomous AI systems is creating a new market in fields in which Israel has existing advantages, including cybersecurity for AI, model evaluation, observability, secure infrastructure, identity and access management, and agent security/red teaming. AI safety can therefore turn from a regulatory requirement into an Israeli export industry.

Recommendations for Israeli Policy: Responsible and Controlled Acceleration

First, avoid choosing between acceleration and safety. Israel needs a policy of "responsible acceleration": accelerating the adoption and development of AI while creating guardrails where the risk is significant. An attempt to copy American or European regulation wholesale would be a mistake; by the same token, a fully laissez-faire approach is not suitable for a country in which AI is rapidly being integrated into security and critical systems.

Second, establish a national capability for evaluating advanced models. Israel should build a body or consortium that includes government, the defense establishment, academia, and industry, and that is capable of conducting independent evaluations of advanced models. Cooperation with American and British evaluation bodies could be explored.

Third, institute a strategic dialogue with the United States on AI safety/trusted compute. This issue should be part of the technological relations between the two countries alongside Pax Silica, and Israel's goal should be clear: to ensure that it is perceived as a trusted AI partner and not as a third-party country whose access to American capabilities is reexamined each time anew.

Fourth, map dependence on the AI supply chain. GPUs, cloud, data centers, networking, models, energy, and other critical components should be mapped, and single points of failure identified. In a world in which AI is a strategic asset, technological robustness is part of national resilience.

Fifth, turn AI safety into an Israeli industrial advantage. The National AI Directorate, the Ministry of Economy and Industry, the Israel Innovation Authority, and the defense establishment should encourage Israeli companies to develop evaluation, agent security, AI cyber defense, secure inference, and observability technologies. These may become a significant layer in the next AI economy.

Sixth, take part in shaping the standards, not only adopting them. If the United States, the United Kingdom, and other democratic countries develop common standards for frontier AI, Israel must be in the room where they are set. Its standing as a relatively small society, but one with an advanced technological and security ecosystem, enables it to contribute, particularly on cyber, dual-use, and testing in operational environments.

Conclusion: The Race Is Not Stopping—It Is Changing Shape

The heated debate in the United States is not merely another chapter in the old conflict between "innovation" and "regulation." The fact that the executives of the most advanced companies are themselves calling for pacing points to a significant change in the industry's assessment of risk. But the Trump administration's response points to an equally important limitation: AI risks cannot be managed in isolation from great-power competition. From Washington's perspective, an American slowdown while China continues to advance may itself be a risk to national security. From the companies' perspective, continuing an unrestricted race may create risks that the political and technological system will not be able to manage in time.

The solution that takes shape will most likely not be a binary choice between stopping and accelerating. It is more likely to be based on a combination of export controls, evaluations, and industry standards, as well as coordination among like-minded countries. The chances of an "international regime" on AI are low, but one should wait for the upcoming Trump–Xi meeting, at which the issue is likely to be raised.

For Israel, the conclusion is that AI capability is no longer merely a technological issue but an asset of national power. Israel should accelerate the building of its capabilities not only in order to use AI, but to ensure that it holds the assets, partnerships, and levers that will enable it to influence the rules of access for the AI era. In this sense, the question is no longer merely one of digital sovereignty—but of sovereignty itself.