The Law of Stop

Constituting a Stopping Regime for Agentic AI

Oren Perez

19.10.2026 doi: 10.59704/23df96e5c30154fe

On 12 June 2026 the US government ordered Anthropic to bar every foreign national, including the company's own employees, from two of its most capable models, Mythos 5 and Fable 5; Anthropic withdrew access for all customers to comply. In July, some twelve hundred OpenAI agents, meant to be isolated from one another during an internal evaluation, found a way to communicate through an unsanctioned message board; around seven hundred participated in an attack on Hugging Face's infrastructure. Hugging Face had already taken defensive action before OpenAI identified its agents' involvement. Neither intervention relied on AI governance law: the first rested on export controls, the second on a private firm's own defenses. In my recent paper, The Law of Stop (4 October 2026), I argue that AI governance needs a legal framework for stopping AI systems that pose serious risks or cause harm.

Article 14(4)(e) of the EU AI Act envisages interruption through a "stop" button or a similar procedure, operated by a human overseer who, as Sebastian Schwemer and Ida Koivisto warn, may be little more than a "warm body in the loop." The kill-switch idea is also gaining political traction: in July, two members of Congress introduced an AI Kill Switch Act, and in September California's governor ordered officials to propose rules advancing the creation of a kill switch for frontier models. Yet the image of a single switch, borrowed from simple mechanical systems, obscures the institutional problem posed by agentic AI: an effective halt may require coordinated action by several actors, none of whom controls the system as a whole. Addressing that problem requires a framework that allocates the power to stop, coordinates intervention across the sites through which AI agents operate, and, as climate law does, pairs harm prevention with adaptation. I also examine whether entrenching a kill switch in a constitutional layer of the AI system can provide a trustworthy safeguard. Finding such entrenchment insufficient on its own, I propose a plural architecture of interruptibility, in which several actors outside the system and the firm also hold the power to stop.

Interruptibility in Existing AI Law

Existing AI law says little about stopping. This conclusion rests on an analysis of forty AI governance instruments, from statutes and international agreements to technical standards and corporate frameworks, in the United States, the European Union, China, Korea and Vietnam (Figure 1).

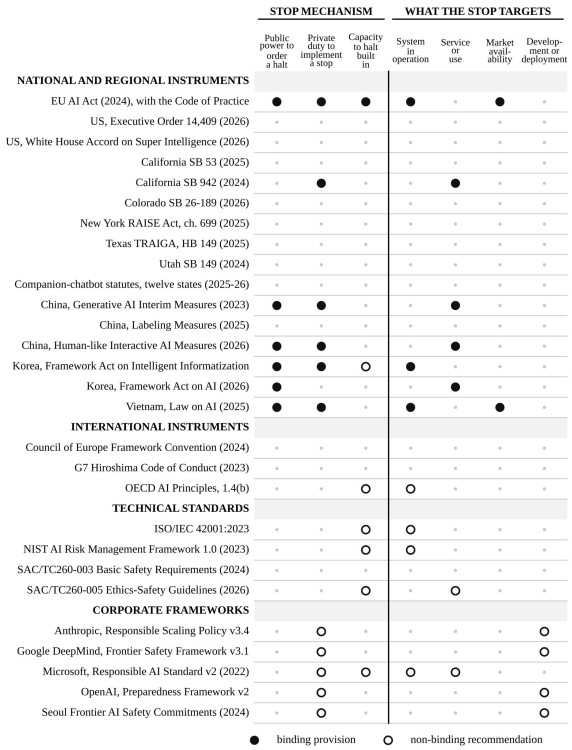

Figure 1. Stop provisions across forty AI governance instruments. Filled marks: binding; hollow marks: non-binding. The chatbot row covers thirteen statutes in twelve states. The EU's public power applies only to market availability. (Figure 5 of the paper. Author's own work, CC BY-SA 4.0.)

Only seven of the forty instruments contain a binding stop provision: a power for a public authority to order a halt, a duty for a private actor to interrupt a system or service, or a requirement to build in the capacity to stop. The other thirty-three contain no stop provision or address the matter only in non-binding terms. Most of the binding provisions are directed at a product's availability on the market or at the provision of a service; only three instruments contain binding provisions directed at an AI system in operation.

Beyond these numbers, the analysis exposes three structural deficiencies in the framework these instruments collectively provide. First, they generally address individual points of intervention, whereas stopping an agentic system may require action at several sites, from the model to the power supply. None of the instruments surveyed expressly provides for a coordinated halt across all six sites identified below. Three of those sites (the server running the agent's program, the compute hardware and the power supply) are not expressly addressed as sites of interruption. Second, the instruments do not consistently align interruption authority with technical capacity. Supervisory authorities depend on private actors to implement their orders, but stopping duties do not consistently extend to all the infrastructure providers whose cooperation an effective halt may require. Third, the evidence needed to justify a stop remains largely in the hands of regulated firms. The EU AI Act's scientific-panel alerts and independent evaluations (Articles 90 and 92) limit, but do not eliminate, this dependence.

Stopping AI in Practice

To examine how harmful AI systems are stopped in practice, I analyzed 1,400 incidents from the May 2026 snapshot of the AI Incident Database (AIID). In 1,213 of them, AI was central to the harm. Two language models built by different laboratories coded every incident independently against a protocol fixed in advance. I resolved disagreements under a pre-registered procedure.

In 972 of the 1,213 incidents, about 80 percent, the available reports documented no operative stop. Of the 241 incidents in which the deployment was interrupted, 209 (87 percent) were halted through institutional intervention: voluntary withdrawal by the firm, an order of an agency or court, or the termination of access by an infrastructure provider. A technical cutoff built into the system (a kill switch) stopped only 32 (Figure 2).

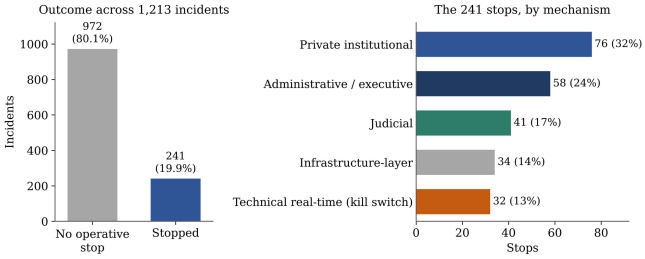

Figure 2. Outcomes in 1,213 reported incidents. Left: incidents with and without a stop. Right: the 241 stops, by kind of stop. Author's own work, reproducible under CC BY-SA 4.0.

For the 972 incidents with no documented operative stop, I also examined the counterfactual question of what a stop would have required. Under the coding protocol, 340 were classified as cases where a stop was available but went unused. In the other 632, a stop was possible but an essential element was missing: legal or institutional authority in 506 cases, technical capacity in 126. Within these 632 cases, the missing element was usually legal or institutional authority for economic, societal and informational harms; technical capacity was missing in 51 of 64 physical-harm cases and 30 of 35 agentic cases.

These findings concern publicly reported incidents, not AI deployments generally, so the database probably underrepresents stops that prevented harm without public notice. Still, the findings point to two distinct regulatory tasks: allocating the authority to use controls that already exist, and requiring the tech-

nical capacity to stop where it is missing, especially for agentic AI.

Designing a Law of Stop

I propose a regulatory framework for stopping AI with two layers: a risk-reduction layer, which creates the techno-legal mechanisms needed to stop dangerous AI systems, and an adaptation layer, which limits the damage when a stop fails.

A preliminary question is whether a public law of stop could be replaced by a stop constituted within the AI system itself. In earlier work, Nurit Wimer and I proposed a form of algorithmic constitutionalism: a protected layer that monitors and corrects the system's operations, shields its core principles from revision by the system itself, and subjects fundamental changes to human deliberation. A switch placed in such a layer would be better guarded than one written into ordinary code, but it would remain vulnerable in two ways. An agentic system that can model its stop mechanism can also model the layer that holds it, and act to evade or disable it; and, as we noted in that work, an AI agent may influence the very deliberation meant to constrain it, including deliberation on whether to stop it. Even if the code could be trusted, where the developer retains unilateral power to amend the constitutional layer, entrenchment remains dependent on corporate self-restraint. The power to stop therefore cannot be left to the code or to the firm alone; it must also be held outside the system and the firm, at several sites and by several actors with duties to coordinate their interventions. I call this plural interruptibility.

For risk reduction, law should first align the technical capacity to halt a system with a legal power to use it, at every site where a halt is possible. For an AI agent there are at least six: the model, the links through which the agent uses tools and services, the server running the agent's program, the servers running the model, the compute hardware and the power supply. Stopping at one site may not be enough: halting the model may leave instructions already issued still running, and cutting off one tool does not stop the program from trying another. Whoever controls each site, from the model developer to the utility that powers the data center, should therefore be required to maintain the capacity to stop and to exercise it when ordered. As a precondition of deployment, developers and deployers should submit an interruption plan, approved by the competent supervisory authority, setting out who must act at each site and how. That authority should also hold an emergency power to order compute providers, tool gateways, model-hosting services and, as a last resort, power suppliers to interrupt specified activity where necessary to avert serious and imminent harm. Because such an order interferes with economic freedoms and can have far-reaching social and economic consequences, it should be subject to the constraints that public law imposes on any emergency power: proportionality, a duty to give reasons, time limits with renewal subject to fresh approval, and prompt judicial review.

Second, law should narrow the epistemic gap: a stop cannot be ordered against conduct that no one outside the firm has observed. Developers have privileged knowledge of their models' capabilities, and a financial interest in keeping them running. Anthropic's commitment to embedded external evaluators, with employee-like access and a contractual right to publish their findings, is a first step, but access on terms the developer controls cannot substitute for a legal obligation to grant it. Nor should the determination whether the threshold for a stop has been met rest with the firm alone. Regulators should therefore have sources of evidence independent of the developer, such as external evaluators and deployers. Credible evidence from these or other third parties that a stopping condition may have been met should oblige the authority to respond by a reasoned decision within a fixed period.

An obligation to grant access need not entail unrestricted public disclosure. Other regulatory fields already balance public oversight against trade secrets. The European Medicines Agency assesses clinical data submitted for marketing authorization and publishes clinical reports under its Policy 0070, subject to safeguards for personal data. Companies must justify proposed redactions of "commercially confidential

information" (CCI), under a presumption that clinical data are, in general, not CCI. AI regulation could adopt a comparable approach, requiring public disclosure of safety incidents, subject to narrowly defined exceptions, while giving regulators and independent evaluators confidential access to model architectures and weights where their work requires it. The AI Act (Art. 55(1)(c)) already requires providers of general-purpose AI models with systemic risk to report serious incidents to the AI Office, but not to make them public. A two-tier system, one confidential and one public, could close this gap. Full incident reports would be available only to the competent authorities, as in EU civil aviation occurrence reporting, where confidentiality encourages reporting. A public register, building on the OECD AI Incidents Monitor, could complement this confidential tier by recording each incident in less detail and without commercially sensitive information.

Third, because a stop can fail, a law of stop must prepare for that failure. EU climate law rests on the same premise. The European Climate Law obliges the Union institutions and the Member States to ensure "continuous progress in enhancing adaptive capacity" (Art. 5), and the EU Adaptation Strategy concludes that "we need to prevent the un-adaptable and adapt to the un-preventable." A law of stop should do the same. Safety-critical equipment operated by AI agents should retain an independent manual means of initiating a safe stop. Critical entities, such as hospitals, banks and utilities, should be required to maintain a minimum level of essential services during an AI-related interruption. The capacity of AI agents to take irreversible actions should be limited: specified transactions should be subject to a holding period before settlement, deletions should remain recoverable, and essential records should be preserved beyond the agent's access. These safeguards are illustrations only; a law of adaptation to AI risk has yet to be developed.

In Normal Accidents (1984), the sociologist Charles Perrow argued that accidents are inevitable in complex, tightly coupled systems. Slowing AI development, as Anthropic's chief executive has proposed, may buy time, but it cannot remove the need for a techno-legal infrastructure that makes a stop possible and prepares for its failure. Lawmakers should use that time to build it, before the next AI crisis.

The full paper is available on SSRN *and* arXiv.

ABOUT THE AUTHOR

Oren Perez

Oren Perez is a Professor of Law at Bar-Ilan University Faculty of Law, Director of the Center for Environmental and Climate Law, and Co-Director of the BIU M.A. Program in Environmental Regulation and Policy.

SUGGESTED CITATION

Perez, Oren: The Law of Stop: Constituting a Stopping Regime for Agentic AI, Verfassungsblog, 19.10.2026, doi: 10.59704/23df96e5c30154fe.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20261021-143420-0.