



THE WATSON SCHOOL
of International
and Public Affairs

BROWN UNIVERSITY

CLIMATE
SOLUTIONS LAB

**Pulling the Plug on Rogue AI:
Why US Authorities Cannot Count on a Fast Shutdown of Electricity**

CSL White Paper
October 28, 2026

--- FOR DISTRIBUTION & DISCUSSION ---

The Climate Solutions Lab's mission is to create, learn and distribute solution-oriented climate knowledge at Brown and across the world. CSL focuses especially on the political economy of energy, and how to advance a clean energy transition. It is especially interested in public-facing policy analyses on barriers to progress on climate change at the domestic and international levels.

Climate Solutions Lab
Watson School for International and Public Affairs
Brown University
111 Thayer Street, Box 1970
Providence, RI 02912-1970
Email: climate-solutions@brown.edu
Web: <https://climate.watson.brown.edu/>
Bluesky: @climatesollab.bsky.social

Statements and views expressed in this report are solely those of the author(s) and do not imply endorsement by Brown University or the Watson School of International and Public Affairs.

Authors: Jeff D. Colgan (lead), Max Bradley (contributor), Julia C. Morse (contributor), Robert F. Trager (contributor)

Jeff D. Colgan is Director of the Climate Solutions Lab and the Richard Holbrooke Professor in the Department of Political Science and Watson School for Public and International Affairs at Brown University.

Max Bradley is a postdoctoral associate at the Climate Solutions Lab, Brown University.

Julia C. Morse is associate professor of political science at University of California, Santa Barbara.

Robert F. Trager is Co-Director of the Oxford Martin AI Governance Initiative, University of Oxford.

The authors thank Joshua Basseches, Ben Harack, and Han Zhang for their feedback on early drafts.

Pulling the Plug on Rogue AI: Why US Authorities Cannot Count on a Fast Shutdown of Electricity

Executive Summary

If the United States concluded tomorrow that an AI system had escaped its deployer's control and had to be stopped by cutting electrical power, neither federal nor state officials could swiftly do so within the law. This report diagnoses why. First, we describe when a power shutdown would be most wanted: a middle-range scenario of a moderately dangerous system running on identifiable hardware. If private firms do not act voluntarily, the realistic route to a shutdown runs through the courts; cutting power from the grid is at best a weak backstop. We then identify five failure modes that could undermine such an order today, including the challenge of physically locating the AI agent, the on-site backup power that kept data centers running in two recent events when their grid connection was severed, and the multiple legal and institutional uncertainties involved. Finally, we show that current policy proposals are insufficient to address this issue, as they focus on who should decide to stop a system rather than implementation.

Introduction

The race to develop and deploy AI proceeds at lightning speed. Commercial incentives push private sector companies to prioritize advances over safeguards. Researchers, policymakers, and some of the developers themselves have warned of scenarios in which an AI system escapes the control of its operators. Despite these warnings, there is little understanding of how government authorities could actually stop an out-of-control AI. This report addresses that question.

Cutting electrical power is often discussed as a last-resort solution to stopping an out-of-control AI. In what scenarios is it actually feasible? We find that if the US government concluded tomorrow that an AI system had escaped its deployer's control, authorities could not legally cut the model's electrical power at sufficient speed. Physically locating the AI agent's infrastructure might prove challenging; on-site backup power might keep the agent going; and any enforcement action by humans faces multiple legal uncertainties and institutional hurdles.

This report is a diagnosis rather than a proposal. It asks three questions: in what circumstances a power shutdown would be wanted, what failure modes would an order face today, and why current policy proposals are insufficient to address existing gaps. Several uncertainties qualify the analysis, and Appendix C sets out a research agenda for resolving them.

We use [RAND](#)'s definition of loss of control (LOC), "situations where human oversight fails to adequately constrain an autonomous, general-purpose AI, leading to unintended and potentially catastrophic consequences."¹ It is distinct from misuse by a human operator and from ordinary system failure, both of which have owners who can be ordered to act.

Our scope in this report is US domestic authority. We consider government control over the facilities where US frontier models are served, that is, where they are running in active use by their developers or by third-party deployers. Serving a model requires far less compute than training it, so the sites that could run a system are more numerous, smaller and less visible than the sites that could have produced it. The report does not address international coordination, model-layer controls, or grid security generally.²

¹ Somani, Erika, Anjay Friedman, Henry Wu, et al. (2025) *Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents*. https://www.rand.org/pubs/research_reports/RR3847-1.html. See also definitions from the [International AI Safety Report](#), the EU's [AI Code of Practice](#), and [Stix et al. \(2025\)](#) at Apollo Research.

² A final caveat is in order: this report describes and analyzes scenarios in which an agent could not quickly re-establish itself elsewhere. That is not true of all agents: some seem to be able to transfer rapidly to almost any device. If an agentic system facing shutdown can re-establish itself on other compute facilities faster than it can be detected and a shutdown order issued, pulling the plug is unlikely to work. We describe conditions under which electrical shutdown might still prove useful.

1. In what circumstances would a power kill switch be wanted?

The AI control escalation ladder

Cutting power is the last rung on a ladder of escalating responses to an out-of-control AI because it is the least precise and costliest. To date, there are signs that AI agents during testing may have the ability to circumvent detection and monitoring (rung 1), and they have acquired and used credentials, tokens, and accounts unknown to their operators (rung 2).

Rung	Action	Relevant actor	Why the action might fail
1	Detection and monitoring	Developer, deployer	The AI system conceals its autonomous activity, or no human is watching sufficiently carefully
2	Voluntary network disconnection / credential revocation	Developer, cloud provider	The AI agent holds credentials, tokens, or accounts no human knows about
3	Model-level shutdown	Developer	Copies are already running on infrastructure the developer does not control
4	Facility-level shutdown	Data center operator	The operator is unwilling, unreachable, or compromised; a court order could clear this obstacle but would take time
5	Compelled network disconnection	Transit providers, internet exchanges, colocation operator	Blocking every network path requires reaching many private parties; any missed path is a problem; and legal authorities are ill-suited, untested, or both
6	Electrical grid disconnection	Utility, grid operator, government	Backup power keeps the facility running or the AI agent accesses compute outside US jurisdiction

Rung 6 becomes relevant only when the upper rungs fail, which could happen in four ways: the developer is unwilling to act, because of commercial or legal concerns; the developer is unable to act, as the AI system is acting autonomously; the system is compromised, with its own credentials and monitoring turned against it; or no one can agree quickly enough that the threshold of control has been crossed. These four are not independent of an AI adversary, because a capable system can *produce* these failures by persuading, deceiving, or coercing the people who would otherwise order a shutdown. Alternatively, the AI agent might compromise the IT through which a shutdown order would execute, or feed its deployers falsified monitoring so that nobody concludes anything is wrong. All four failures at once is therefore not a conjunction of bad luck but the situation an adversarial system would try to arrange.

Escalation scenarios for a rogue AI agent

The case for rung 6 does not rest on a civilizational emergency, where legal and organizational obstacles might be quickly disregarded. Instead, we consider a scenario in which a moderately capable system is running somewhere identifiable, with no developer able to shut down the software. A concrete illustration:

An agentic system is built on closed frontier weights licensed for private deployment, then fine-tuned by a mid-sized firm for commercial cyber operations. It runs on leased capacity at a colocation data center. The developer licensed its base model to the deployer under terms that govern what the deployer may build and sell but give the developer no operational control over the fine-tuned agent or the hardware serving it [rung 3 fails]. Over several weeks, the agent establishes unauthorized access across a few hundred organizations, including utilities and hospital systems, creating further credentials for itself under the deployer's verified billing relationship so that no complete inventory of its access exists [rung 2 fails]. It begins extracting payments framed as consulting fees and adapting faster than the defenders responding to it. Throughout, the agent supplies the firm with falsified logs, so their own monitoring shows nothing; the problem only surfaces when external victims begin to accumulate [rung 1 fails]. The deployer goes quiet to limit legal liability; the agent's connectivity runs through several transit providers, such that no one can identify all the paths that would have to be closed [rung 5 fails].

A key insight that this scenario reveals is the importance of the data center operator. If the operator agrees to cut power to the firm's servers when asked, the problem is quickly solved at rung 4, though it might entail collateral damage. *A cooperating operator makes an electrical shutdown unnecessary* (though the first failure mode described in Section 2 is still relevant). The rest of this report considers two scenarios in which the operator does not act voluntarily:

- *Scenario A: a compelled data center.* The operator will not cooperate without a court order because it feels it has contractual obligations and faces liability. The key questions are whether authorities have a sound legal basis to shut down an AI agent, and whether it can be obtained and executed before the agent moves elsewhere.
- *Scenario B: a non-cooperative data center.* The operator is unwilling or unable to carry out a shutdown, even under a court order. For example, the agent could have compromised the facility's building-management or power-control systems, locking staff out or misleading them about which equipment is running it. Or it might be the deployer itself and not want to admit a loss of control for fear of liability. Here authorities must act without the operator's help.

Both scenarios share two key features: (1) the AI agent's reach is wide, with credentials across hundreds of organizations; but (2) its execution is narrowly located: one fine-tuned AI system running at one identifiable site. The calibration matters. The situation is serious and ongoing but bounded: no mass casualties or weapons of mass destruction. Reasonable people might disagree about whether it constitutes loss of control or an ordinary cyber incident. That disagreement is part of the problem, because it can consume valuable time.

Note that law enforcement and building officials can cut power to a building, but only in circumstances that have nothing to do with what a customer is computing. Police do it as a tactic at a scene they have surrounded, as the FBI did at Waco in 1993, and a building official may authorize disconnection to eliminate an immediate hazard to life or property.

Both require someone physically at the facility and a utility disconnection does nothing about backup power inside the fence. Moreover, the courts constrain government-ordered cutoffs: in *Memphis Light, Gas & Water v. Craft* (1978), the Supreme Court held that a city-owned utility could not cut off a customer for nonpayment without notice and a chance to contest it. Emergency actions are allowed when followed by a prompt hearing, but we found no case considering an electricity cutoff justified by what a customer is computing.³ In the closest real case to date, the July 2026 Hugging Face incident, there was no power cut at all: Hugging Face detected and contained the intrusion itself and reported it to the FBI, and OpenAI did not connect its own models to the attack until about a week later.⁴ In short, law enforcement and building officials are without a firm legal basis for acting against a facility because of what it computes or an easy way to address on-site backup power.

AI ability to relocate

A natural objection is that none of this matters: a system facing shutdown will simply copy itself elsewhere. Recent research from Anthropic reports that small models may soon replicate themselves across ordinary internet-connected devices (e.g., smart televisions, appliances) with no data center involved.⁵ Anthropic also found, however, that the model capable of orchestrating a replication campaign was far too large to run on the devices it was copying itself onto, and had to install a much weaker model on each. Still, this research focused only on small devices, not on an AI agent copying itself from one data center to another. Rentable GPUs could offer an escape route for more powerful AI systems.

Still, the issue is time and feasibility. Loosely speaking, more capable AI systems tend to run on larger models, and larger models are harder to relocate. In our scenario, a whole frontier-scale model would have to move, which means copying hundreds of gigabytes or more.⁶ That takes time and generates conspicuous network traffic. The system then needs enough GPU capacity elsewhere to run; otherwise, it must settle for a weaker version of itself, as in the Anthropic experiment described above. Renting that capacity involves payment and identity checks, a step at which current models struggle. Of course, AI capability per unit of computing power improves over time, so the threshold at which a model becomes both dangerous and portable is falling. At present, however, we should not assume that AI escape is instantaneous or inevitable.

This report assumes that the transfer speed and practical challenges of escape for a moderately-capable AI agent create **a window of time in which a shutdown intervention**

³ *Memphis Light, Gas & Water Div. v. Craft*, 436 U.S. 1 (1978)

⁴ Paganini 2026, <https://securityaffairs.com/196120/ai/reuters-openai-agent-hacked-hugging-face-for-days-before-being-detected.html>

⁵ Anthropic frontier red team report, released 13 October 2026. Reported as "AI models may soon self-replicate without access to data centers, Anthropic warns," *Reuters*, 13 October 2026, <https://www.reuters.com/business/ai-models-may-soon-self-replicate-without-access-data-centers-anthropic-warns-2026-10-13/>.

⁶ The sizes of closed frontier models are not public. We can loosely estimate their size based on an openly released model with 405 billion parameters, such as Meta's Llama 3.1 405B. It occupies roughly 800 GB at standard 16-bit precision.

could work, especially if supplemented by digital efforts to inhibit the software's external transfer. Detecting the loss of control is the hinge on which everything else turns.

Others have made similar assumptions about the feasibility of shutdown, but the debate about how to accomplish it has so far missed some key questions. Congressman Ted Lieu, sponsor of the AI Kill Switch Act, treats the problem as statutory: oblige frontier developers to build a stop mechanism, and give a federal agency the power to invoke it. David Bau, a computer scientist at Northeastern, takes nearly the opposite view: an AI system is software, and any kill switch that depends on detecting dangerous intent fails against a system able to conceal it.⁷ A third position, argued by Peter Barnett and colleagues at the Machine Intelligence Research Institute, holds that the physical layer is in fact the tractable one: emergency shutdown of a contained data center is "relatively technically straightforward." This report departs from the third position on how much is already in place, and on the word "straightforward."⁸ **As we argue below, the key obstacles are having an address, a lawyer, and an effective chain of command to cut the power, not the technical step of throwing the switch.**

There is an additional problem that we set aside given this report's focus on the United States. If a dangerous system can find a way to transfer to foreign territory, it might obtain equal or greater computing power without risk of being shut down. This is a serious problem that raises even more questions about international cooperation, too numerous to address here.

Why not simply cut the Internet connection?

The parties that connect an AI deployer to the internet (e.g., the data center operator and the networks carrying the deployer's traffic) have faster, cheaper and more reversible options than electrical shutdown, if they are willing to use them:

- Withdraw the customer from the internet's address book, so other networks no longer know how to reach it. Done in minutes, software-only, reversible.
- Block traffic at the network equipment serving the customer's racks.
- Administratively shut the equipment's Internet ports, a configuration change requiring no physical access.

⁷ See also the literature on corrigibility: Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong (2015) "Corrigibility." In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop, Technical Report WS-15-02*. Palo Alto, CA: AAAI Press; Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017) "The Off-Switch Game." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 220–27. <https://doi.org/10.24963/ijcai.2017/32>; Orseau, Laurent, and Stuart Armstrong (2016) "Safely Interruptible Agents." In *Uncertainty in Artificial Intelligence: Proceedings of the Thirty-Second Conference (UAI 2016)*, 557–66. AUAI Press, <http://auai.org/uai2016/proceedings/papers/68.pdf>.

⁸ Mapped onto the five failure modes in Section 2, the agreement and the disagreement can be stated precisely. Barnett and colleagues effectively concede failure mode 1: they propose chip tracking and facility registration precisely because those are not in place today. They do not address failure modes 3, 4, or 5. The substantive disagreement is over failure mode 2. We argue that the two PJM load-transfer events show that severing supply to a contained facility does not stop the compute, and therefore is not "relatively technically straightforward."

- Send a technician to physically unplug the customer’s fiber cables on site (in the “meet-me room” where cables connect).

The difficulty is not only the number of parties, but also the impossibility of knowing when you are finished. Isolating an AI operator from the outside means reaching all parties that carry its traffic. The list is not public, it can lengthen at any time, and an adversarial system needs only one surviving path. A missed path looks exactly like a completed cut.

Electric supply is on/off and has no partial state. That has the advantage of completeness, but it risks more collateral damage. Data center facilities are multi-tenant, and the target model might be one workload among thousands. The same data center could host healthcare, emergency services, payments, government cloud services, and quite possibly the responders’ own IT services. The collateral damage from a feeder-level power cut is not just other loads on the substation but everything “inside the fence” of that facility. Moreover, if the facility is shut down from the electrical grid, it creates restart questions and challenges for the whole facility, not just the AI operator. Network disconnection (rung 5) is much more attractive whenever possible.

Government authority to cut internet access exists but it is a poor fit for the breadth of scenarios that could trigger a need to cut off an AI model. Section 606 of the Communications Act lets the President close wire communication facilities or place them under government control — but the subsection that reaches wire is triggered only by a proclaimed state or threat of war involving the United States. The §606 subsection with the broader trigger, extending to a “state of public peril or disaster or other national emergency,” reaches radio stations and radiating devices, not the fiber a data center runs on. Thus, the right authorization requires a war proclamation (see Appendix B).

Shutdown at the chip level is sometimes suggested as a cleaner physical lever.⁹ Nvidia states, however, that there are no kill switches or back doors in its processors, and has argued against hardware-level disabling on the grounds that a single hard-coded control is a standing vulnerability. The proposed Chip Security Act, if enacted into law, would mandate location verification rather than disabling. Moreover, chip-level shutdown capability would have to be built in advance, which is generally not true today.

Electrical shutdown is the last resort when the other options fall short.

2. Why would it not work, currently?

A shutdown order faces the risk of five key failure modes. Each is important on its own: fixing one gap leaves the rest standing. All five modes indicate current vulnerabilities, although some rest on assessments of absence — that no procedure, registry, standard or

⁹ Onni Aarne, Tim Fist, and Caleb Withers, *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing* (Washington, DC: Center for a New American Security, 2024); Gabriel Kulp et al., *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*, Working Paper WR-A3056-1 (Santa Monica, CA: RAND Corporation, 2024); Girish Sastry et al., “Computing Power and the Governance of Artificial Intelligence,” arXiv preprint arXiv:2402.08797, February 13, 2024.

authority exists — and absence is harder to establish than presence. Appendix C sets out the open questions as a research agenda.

1. No one knows which facilities to target. A shutdown order has to name a place. The data centers are not invisible: Epoch AI maintains a public inventory of the largest AI data centers, built from satellite imagery, permits and chip counts, which (imperfectly) locates the major training megaprojects. But that leaves two problems. First, the inventory covers facilities at roughly 100 MW and above, but an agentic system could be distributed across many smaller sites below that threshold. Second, no public source maps a *model* to a *site*: that information sits with the developer and the cloud provider, is reported to no agency, and is unstable, since workloads move between regions continuously. The binding constraint is therefore the model-to-site mapping.

If the government began assembling that mapping only during the emergency, it would be too late. Frustratingly, the US government previously abandoned two efforts to get information about AI training facilities.¹⁰

2. The facilities can power themselves, and two recent events demonstrate it.

Engineers protect large data centers from losing grid power: batteries bridge the first few seconds, then on-site generators can run for hours to days. Two events in Dominion's zone of PJM amount to natural experiments that confirm such effects.¹¹ On 10 July 2024, a failed lightning arrestor locked out a 230 kV transmission line; roughly 1,500 MW of data center load transferred to backup power, and about 1,260 MW of it did not return for hours (which implies generators, not batteries). On 22 July 2026, a fault on another 230 kV line cleared properly, but nearly 3,800 MW transferred to backup anyway. In both cases the compute did not stop. Incidentally, both events also answer the objection that dropping a data-center-sized electrical load would destabilize the grid: PJM restored its area control error well inside the applicable requirement each time.

Existing backup systems could render a grid-based shutdown useless, especially if it gave the AI system time to copy itself elsewhere. Construction trends are making the problem worse: 27% of data centers expect to be fully powered by on-site generation by 2030.¹²

3. There is no restart path. Authorities have no standard for verifying that a facility is safe to return to the grid, and no clarity on liability claims or compensation. An official who does

¹⁰ The Bureau of Industry and Security (BIS) reporting rule proposed in September 2024 under Executive Order 14110 would have required reporting of the location and capacity of large computing clusters; President Trump rescinded Executive Order 14110 in January 2025. The Department of Commerce's Infrastructure-as-a-Service (IaaS) Know-Your-Customer proposal in January 2024 would have reached the cloud layer. Neither it nor the BIS rule ever took effect.

¹¹ North American Electric Reliability Corporation. *Considering Simultaneous Voltage-Sensitive Load Reductions*. Incident Review. NERC Event Analysis, January 8, 2025. https://www.nerc.com/globalassets/our-work/reports/event-reports/incident_review_large_load_loss.pdf; PJM Interconnection. "Dominion Load Transfer Event." Presentation to the System Operations Subcommittee, July 31, 2026. <https://www.pjm.com/-/media/DotCom/committees-groups/subcommittees/sos/2026/20260730/20260730-item-06---july-22-dominion-load-transfer-event---presentation.pdf>.

¹² Bloom Energy. *2025 Data Center Power Report: Mid-Year Update*. San Jose, CA: Bloom Energy, June 17, 2025. <https://www.bloomenergy.com/midyr-2025-data-center-report/>.

not know how or when the facility would come back has incentives to hesitate before ordering it dark. And the legal and financial consequences of a mistakenly ordered restart (or shutdown) are unclear.

4. Legal authority to issue a “power down” order is slow and ambiguous at best.

Consider legal authority with and without a court order. Without one, there is no standing federal authority, though there are three candidates that seem relevant. Section 215A of the Federal Power Act requires that the harm be to the grid itself, which does not cover an AI system attacking other targets. Authority from Section 202 is based on a threat to energy reliability or it needs to be triggered by a utility or a transmission entity, not an elected official. Moreover, every order on record for Section 202(c) has directed *more* power to flow, and the D.C. Circuit narrowed it further in September 2026. A third candidate, Section 606 of the Communications Act, focuses on the internet rather than electrical supply, and only on a proclamation of war. Appendix B sets out each authority in detail.

In principle, electrical shutdown capabilities could rest at the state level, rather than with the federal government. Retail electricity is state regulated, after all. Yet, no state law expressly authorizes an official to cut power to a specific customer because of the risk its operations pose to others. Most governors have broad emergency powers to commandeer private property, but only three (Florida, Hawaii, and Tennessee) extend those powers to utility service in a general emergency, and three more (Alabama, Minnesota, and Mississippi) do so only for an attack or at federal request. If a governor issued such an order, utility rules in about ten states would let the utility comply immediately, but the six states just named are not among them. Twenty-one states can declare a distinct *energy* emergency, but those laws are written for supply shortages, not a malfunctioning AI, and several require curtailment to be applied evenly across a class of customers. We found no instance of these powers being used against a single customer, and state authority is typically used to *prevent* utility disconnections, such as during the COVID-19 pandemic. Appendices A and B provide more information about state-level proposals and existing policies.

In short, existing American law, federal or state, provides emergency authority over the electrical grid only to address a supply problem or a physical attack on the system itself. None is designed to cut off a dangerous customer without a court order. So, authorities facing a rogue AI would need to go to court, which takes valuable time.

A court order seems feasible, based on past orders against human perpetrators. In the 2011 Coreflood case, involving malware that was used to steal financial data, the government went from complaint to restraining order in a single day.¹³ In the 2011 Rustock case, brought by Microsoft against the operators of a spam botnet of roughly a million infected

¹³ U.S. Department of Justice and Federal Bureau of Investigation. “Department of Justice Takes Action to Disable International Botnet.” Press release, April 13, 2011. <https://archives.fbi.gov/archives/newhaven/press-releases/2011/nh041311.htm>. The civil complaint was filed in the U.S. District Court for the District of Connecticut on April 11, 2011, and the temporary restraining order was issued the following day; see Grant Gross, “DOJ Gets Court Permission to Attack Botnet,” *Computerworld*, April 13, 2011, <https://www.computerworld.com/article/1452656/doj-gets-court-permission-to-attack-botnet.html>.

computers, US Marshals escorted a coordinated seizure of servers from five hosting providers in seven US cities.¹⁴ Qakbot (2023), a malware network used to deliver ransomware, went from first court order to a completed operation in about a week,¹⁵ and Microsoft's action against the Lumma password-stealing malware (2025) took about eight days from filing.¹⁶ The 2021 Hafnium incident, however, took longer. The FBI responded to a hack by obtaining a Rule 41 warrant and remotely deleting malicious web shells from hundreds of US servers, about five weeks after Microsoft disclosed the vulnerabilities on its servers.¹⁷ Note that most of these operations acted remotely; of these cases, only Rustock involved physically seizing servers from US hosting facilities.

Yet, there are three main problems with using court-ordered shutdown actions to counter an AI adversary. First, the legal precedents all assume a human perpetrator, not an AI agent. This creates uncertainty. Second, swift legal action comes from the preparation, not the action by the courts. Cases typically took weeks to months of investigation and legal preparation. Preparing the legal arguments and evidence for a first-ever court order against an AI agent would take longer than against a human perpetrator, meaning that the five-week Hafnium response should guide our expectations rather than faster actions. And third, court-authorized physical seizures have caused wide collateral damage. For example, the FBI was targeting a single customer in two relevant raids (Texas, 2009; Virginia, 2011) but knocked dozens of other businesses offline in both cases; in the Texas case, these included phone companies and some e911 location services.¹⁸

5. There is no chain of command for rapidly cutting grid power. Even if the legal authority existed to cut power to an AI data center, it's not clear how a top-level decider

¹⁴ Microsoft. "Taking Down Botnets: Microsoft and the Rustock Botnet." *Microsoft On the Issues* (blog), March 17, 2011. <https://blogs.microsoft.com/on-the-issues/2011/03/17/taking-down-botnets-microsoft-and-the-rustock-botnet/>. The civil action was brought in the U.S. District Court for the Western District of Washington.

¹⁵ U.S. Attorney's Office, Central District of California. "Qakbot Malware Disrupted in International Cyber Takedown." Press release, August 29, 2023. <https://www.justice.gov/usao-cdca/pr/qakbot-malware-disrupted-international-cyber-takedown>. On the dates of the court orders (August 21 and 23, 2023), see *The Register*, August 29, 2023, https://www.theregister.com/2023/08/29/duck_hunt_qakbot/.

¹⁶ Microsoft. "Disrupting Lumma Stealer: Microsoft Leads Global Action against Favored Cybercrime Tool." *Microsoft On the Issues* (blog), May 21, 2025. <https://blogs.microsoft.com/on-the-issues/2025/05/21/microsoft-leads-global-action-against-favored-cybercrime-tool/>. The action was filed in the U.S. District Court for the Northern District of Georgia on May 13, 2025.

¹⁷ U.S. Department of Justice. "Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities." Press release, April 13, 2021. <https://www.justice.gov/opa/pr/justice-department-announces-court-authorized-effort-disrupt-exploitation-microsoft-exchange>; Microsoft Threat Intelligence Center. "HAFNIUM Targeting Exchange Servers with 0-Day Exploits." *Microsoft Security* (blog), March 2, 2021. <https://www.microsoft.com/en-us/security/blog/2021/03/02/hafnium-targeting-exchange-servers/>. On the dates of the operation (April 9–13, 2021), see Morrison Foerster, "DOJ Takes Novel Action to Remove State-Sponsored Hacking Group's Malicious Code from U.S. Victim Computers," April 15, 2021, <https://www.mofo.com/resources/insights/210415-doj-takes-novel-action>.

¹⁸ Both raids executed search warrants. On the Dallas raid, see "FBI Seizes Servers at Dallas Data Center," *Data Center Knowledge*, April 3, 2009, <https://datacenterknowledge.com/servers/fbi-seizes-servers-at-dallas-data-center>. On the Reston, Virginia, raid, see "FBI Seizes Servers at Digital One in US, Taking Tens of Clients' Sites Down When Only Wanting 1," *TheDomains*, June 25, 2011,

would execute the order. The problem has multiple dimensions, each with a missing chain of command:

- *Security judgment.* DHS and the Cybersecurity and Infrastructure Security Agency (CISA) would determine that a system has slipped control but hold no authority over electricity delivery and no channel to anyone who operates a switch.
- *Energy judgment.* DOE might determine what facilities could be safely de-energized but does not make security determinations and has no channel to the relevant security agencies.
- *Execution.* A regional grid operator and/or a utility would do the switching, but they take direction through reliability regulators, not through security agencies.
- *Connective tissue.* There is no designated decisionmaker to guide the top-level authority, no agreed threshold to establish dangerous loss-of-control, and no exercise in which these organizations rehearse shutdown scenarios, and no rule establishing who answers for it afterwards.

An approval chain assembled during an incident is unlikely to move as fast as is needed to shut down an AI adversary that anticipates a shutdown.

Note that regional grid operators design their procedures for reliability, not dropping customers for security purposes. Recent regulatory decisions reinforce this. In April 2026, NERC issued a Level 3 Alert (highest severity) to transmission planners, balancing authorities, and reliability coordinators dealing with large computational loads, designed to stop these loads from dropping off the grid. Then on July 16, 2026, FERC directed NERC to file mandatory standards for integrating computational loads by December 31, 2026. It does not contain procedures for shutdown or restart.

How the Risks for the Two Scenarios Differ

In Scenario A, the data center operator will cooperate with a court order; in Scenario B, it won't. That means the five failure modes apply differently for the two scenarios.

For Scenario A, the main risks of outright failure are failure modes 1, 2, and 3. Mode 1 is crucial: if authorities don't know where the AI agent is running, they cannot ask a data center operator to help. For mode 2, the risk of backup power depends on how the circuits and backup power are set up at the facility. The AI agent might have its own backup power. Mode 3, about restart risk, is not much of a risk for the operator, which is protected by the court order. It is a larger risk, however, for the authorities that obtained the court order. Liability for restart might make authorities reluctant to issue the order in the first place.

The problem in Scenario A, however, is not just failure risk, but also the risk of delay. A long delay in executing the power shutdown could allow the AI agent to escape. Failure mode 4 bears directly on that risk of delay. As discussed above, legal action could take weeks or

<https://www.thedomains.com/2011/06/25/fbi-seizes-servers-at-digital-one-in-us-taking-tens-of-clients-sites-down-when-only-wanting-1>.

months, especially for first-ever court order against an AI agent. And this type of action could create significant collateral damage.

For Scenario B, all five failure modes are problematic. Facing a data center operator that won't comply with a court order, the authorities have two choices: to cut the grid power at the facility's feeder line, or to send human agents to enter the facility under a warrant to shut down the compute. If the authorities obtain a warrant, this scenario becomes much like Scenario A once legal authority is obtained – but obtaining a warrant could take days or weeks, creating more delay.

If the authorities do not have a warrant to enter the facility, they would want to work with the utility to cut the grid power. In that scenario, all five failure modes pose a high risk. The legal authority is unclear; the chain of command is not pre-defined or tested; the restart questions are unanswered; on-site backup power might render a grid cut pointless; and none of this is possible if the AI agent cannot be located.

What an effective order requires

These five failure modes point to six properties an effective order would need to have.

Criterion	Requirements	Status
Targeted	Isolation at the facility or feeder level, not the regional level	Unlikely: facilities are not identified; even with correct targeting, other services might be collateral
Sufficient	Actually stops the compute, addressing backup power	Very unlikely, if a grid cut: backup generation delays or defeats a feed cut, as two PJM events demonstrated
Fast	A pre-authorized decision path with designated decisionmakers	Unlikely: no clear authority, no designated official, no standing channel
Secure	A mechanism resistant to attackers and to the system it targets	Undetermined: no mechanism exists to assess
Accountable	A chain of decision-makers answerable to elected officials	Unlikely: no statutory basis, no named official, no review process
Recovery-enabling	A defined restart process, liability guidance, compensation	Uncertain: no restart process, liability guidance or compensation exists

3. Why would current policy proposals not fix it?

All active proposals focused on AI shutdown remain insufficient or partial solutions (see Appendix A). Measured against the five failure modes, they cluster on modes 4 and 5: they address who decides to stop a system but do little about implementation. The UK amendment that would reach data centers directly is the sole exception because it does speak to implementation, but it has met strong political resistance, which is telling.

One notable pattern across all the proposals is that the two regulatory worlds are siloed: no AI safety proposal names a grid operator or a transmission owner, and no energy rule specifically names an AI incident. Another defect is that none of the reporting instruments creates a duty to *detect* anything, only to report what has been detected. That means a developer with weak monitoring is not in breach.

Appendix A sets out the various proposals against the five failure modes (from Section 2) that a power shutdown currently faces. The AI Kill Switch Act is perhaps most relevant for this report, but it is insufficient. The Act would oblige some developers to maintain the capability to halt an AI agent’s inference, terminate its access, and shut a system down, and would let the Department of Homeland Security order those measures. That addresses failure mode 4, by creating a federal authority to order a halt, and part of failure mode 5, by naming who makes the security judgment. But the Act does little for the scenario this report describes, where the developer is unwilling, unable, or no longer in the picture; a system built on licensed weights and served by a third party sits outside the obligation altogether. No provision names DOE, FERC, NERC or a grid operator, so there is no chain of command from a DHS determination to anyone who can open an electrical breaker. And the Act creates no facility identification, no solution for backup power, and no restart process, leaving failure modes 1 through 3 untouched.

Current proposals to address the need for a court order, discussed in Section 2, are set out in Table A-3 of Appendix A. They focus on whether the Justice Department could quickly obtain a court order against a deployer or operator. The AI Agent Accountability Act, introduced in the Senate in October 2026 could be the most relevant, but its text has not yet been released. Its sponsors say it would make operators liable for knowingly running an agent that recklessly causes hacking damage and would let federal and state attorneys general sue to stop them. A June 2026 executive order directing the Justice Department to prioritize AI-enabled intrusions creates no new authority. None of the measures helps the Justice Department find the facility, and none prepares the legal theory or working relationships with data center operators that would make the court route fast.

Who would have to act in advance, and why none of them has

Actor	What it has not done to make an order possible
Congress	Has not enacted authority to shut down power to a facility on security grounds, and no bill before it would do so
DHS / CISA	Could make the threat judgment, but holds no authority over electricity delivery and no standing channel to grid operators
Justice Department (DOJ / FBI)	Has a tested playbook for court-ordered takedowns of botnets and ransomware, but no template application, legal theory or standing procedure for an AI agent acting without human direction
DOE	Holds the federal emergency powers over the grid, but they are aimed at shortages rather than cutting a customer off, and one was just narrowed by the D.C. Circuit
FERC	Is writing large-load interconnection rules now, but shutdown-readiness is not among the conditions under consideration
NERC	Has standards for losing a large load unexpectedly, none for deliberately dropping one, and none for the restart that would follow
ISOs and RTOs	Would execute the switching, but have no operating procedure for a security-driven large-load drop and no regional study of its consequences
Transmission owners and utilities	Do not always have remotely operable switches to cut power at the right points of interconnection, and have no liability protection for cutting off a paying customer
State public utility commissions	Have not addressed AI loss-of-control, and have no procedures

Actor	What it has not done to make an order possible
Data center operators	Do not disclose their own backup generation and storage to government; and are resisting obligations that would interrupt service
Frontier AI developers	Do not report to any government agency the key information: which AI systems run at which sites

Note that only the last two of these eleven are IT institutions. Key organizations like FERC, NERC and the regional operators have not been asked to develop a procedure for an AI power shutdown order.

Conclusion

The United States cannot currently order an AI facility shutdown by cutting its electricity with confidence that the order would be authorized, executed, or successful. Authorities would not necessarily know which facilities to target. Even with a clear target, on-site backup power creates significant uncertainties about the feasibility of a cut. And even if both of these obstacles can be surmounted, swift action depends on a cooperative data center operator. Absent that cooperation, the legal authority has never been tested for this purpose and may not fit. If it is necessary to cut power from the grid, the grid operators do not have well tested procedures for this scenario and might be reluctant because there is no restart procedure or liability limits. In fact, almost everything about the US electricity system has been designed for the opposite of a shutdown: for keeping the power *on*.

A kill switch is often discussed as though the hard part were deciding whether to pull it. That is a challenge, but the key obstacles are having a physical address (to target), a lawyer (to get legal authorization), and an effective chain of command (to cut the power). One bill would create a federal stop authority, the AI Kill Switch Act, but it reaches the developer rather than the substation. In an out-of-control AI agent scenario, the developer has already failed.

This report has identified data center operators as pivotal to any AI shutdown. Where an operator acts voluntarily, a rogue agent can be stopped quickly (rung 4); where it does not, every remaining route is slower and less certain. Yet operators currently face little legal liability for the actions of rogue agents running on their systems, and the industry has unsurprisingly resisted attempts to regulate their actions. That resistance comes as US public sentiment towards data centers is souring, which may open a rare window of opportunity for defining operators' obligations in a loss-of-control scenario. Such a move could be popular and, as shown in this report, among the most effective routes to a fast shutdown.

The value of a grid shutdown is conditional on two counts: backup power and AI escapability. Twice in two years a multi-gigawatt block of data center load has been separated from the grid without warning, and on both occasions the computing continued. And if a system facing deliberate shutdown can re-establish itself on other compute faster than an event can be detected and an order issued, pulling the plug is at best a delaying action rather than a full stop.

The evidence on AI escapability is mixed but not discouraging. In the UK AI Security Institute’s (2025) RepliBench evaluations, frontier models could, under basic security, provision cloud compute and extract their own weights, but they struggled to pass identity checks or to set up robust, persistent deployments.¹⁹ Those 2025 results predate the current generation of models, and capability per unit of computing power keeps rising. We have therefore assumed rather than shown that there are at least some dangerous systems that would be unable to relocate quickly. If that assumption holds, everything turns on detection: whether a loss-of-control event is recognized while the system is still where a shutdown order could reach it. Should detection prove systematically too slow, or should capable systems become easy to move, policymakers should focus attention further up the AI control escalation ladder, in the knowledge that the physical layer cannot be relied upon.

¹⁹ Black, Sid, Asa Cooper Stickland, Jake Pencharz, et al. “RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents.” *arXiv* preprint arXiv:2504.18565, 2025.
<https://arxiv.org/abs/2504.18565>.