RAND

Research Report

MICHAEL J. D. VERMEER, BRIAN A. JACKSON

Internet Cutoff Switches as a Local Emergency Response for Damaging Artificial Intelligence Incidents

As artificial intelligence (AI) development and deployment proceed, there is a growing concern that damaging AI incidents could occur—incidents in which unexpected or anomalous AI behavior leads to significant harm (Jackson and Frelinger, 2025; Arnold and Toner, 2021). Prior RAND work has examined how decisionmakers might respond to and prepare for such incidents (Jackson and Frelinger, 2025). Although there are a variety of ways in which a damaging AI incident might arise and proceed, many scenarios assume that at some point an AI might use the internet to propagate itself throughout digital infrastructure (Clymer, Wijk, and Barnes, 2024; Kokotajlo et al., 2025).

As a result, RAND work has examined the potential to use broad-acting internet kill switches as one potential tool to mitigate the damage or slow the spread of a damaging AI incident (Vermeer, 2025). However, such technical options as those examined in that prior work involve collateral damage. In the case of internet kill switches, the collateral damage is the cost associated with the lost function of the systems that have been cut off from the internet. To devise a practically useful option for mitigating damage from a potential

KEY FINDINGS

- **Data center operators might face opposing incentives** during a damaging AI incident: Delaying an internet cutoff would preserve revenue but increase the risk of catastrophic AI escape, whereas an early cutoff would prevent wider harm at significant short-term cost.
- Without internalizing catastrophe risk, **operators would have strong financial motivation to postpone or avoid an internet cutoff**, making catastrophic outcomes more likely.
- **Assumed liability affects response timing:** Higher assumed operator responsibility for external damage leads to earlier, more precautionary cutoff use; lower liability leads to longer delays and elevated public risk.
- **Installing cutoff switches alone is insufficient:** Effective mitigation requires both the capability to disconnect and incentives to use the switch early, including compensation for lost revenue or protective liability frameworks.

damaging AI incident, it will be necessary to clearly understand the costs, benefits, and circumstances in which it will be practical or impractical to use a tool like an internet kill switch to mitigate a damaging AI incident. In this report, we look at the use of this technology as a response option but at a local level—a single data center—and use the term *internet cutoff switch* for this narrower option.

The Damaging AI Incident Scenario

We will assess the utility of limited internet cutoff switches as a means of responding to a damaging AI incident, with the aim of identifying the circumstances in which an internet cutoff switch would be a viable course of action for mitigating damage. To assess this, we first describe a scenario in which a damaging AI incident is developing within a single inference data center, and decisionmakers must decide whether to cut off the internet to contain the damage within that single data center[1]—but, in doing so, the decisionmakers would accept financial costs by cutting off revenue that might be produced while efforts to stop the incident are underway.

We envision an incident wherein an AI, originating in a data center, begins to resist or escape control. In more practical terms, this loss of control would manifest as the AI model either failing to act as instructed or taking additional actions beyond what is necessary to accomplish its instruction. Taking additional actions would consume compute resources, and our scenario assumes that the incident causes a gradual loss of function of the data center as the AI incident proceeds because the rogue AI begins to occupy or consume the compute capability of the data center. Moreover, the longer the incident proceeds, the greater the chances become that the AI will use the center's connection to the wider internet to spread beyond the data center and cause a much greater loss of function to societal infrastructure beyond the data center.

The data center operators could prevent the spread of the AI incident beyond the data center by cutting off internet access to the data center. However, doing so would immediately cease any functional use of the data center until the AI incident could be resolved and internet access safely restored. Therefore, the operators have a choice to make between (1) immediately stopping all revenue generated from the data center operations to prevent a greater crisis and damage occurring beyond the data center and (2) delaying the internet cutoff to try to manage the incident themselves for as long as possible to capture revenue from continued (if degraded) data center operations.

Although there are many potential stakeholders in this scenario, we assert that data center operators are the only potential decisionmakers who would have both good visibility into the indicators that a damaging AI incident was occurring and the ability to actually use an internet cutoff switch. In this conception, the damaging AI incident would be analogous to a scenario involving an industrial facility with an emergency shutdown switch that is meant to immediately cease operations to prevent a catastrophe.

Cost of Internet Cutoff Switches

As we considered the cost of internet cutoff switches, we identified three main cost drivers:

- loss of revenue from data center operations because of the lack of internet connectivity
- loss of revenue from data center operations because of the AI incident degrading data center functions over time
- the expected value cost of the AI incident catastrophically spreading beyond its origin in the data center, as long as the data center retains internet connectivity.

We envision an internet cutoff mechanism to be a physical one (i.e., one that would require in-person action and not one that could be triggered remotely). We assume that the cost to put in place infrastructure and training to facilitate an internet cutoff switch is negligible compared with these potential cost drivers and therefore do not explicitly reflect those costs in our analysis.

We identified two different types of data centers that could conceivably be the origin of the damaging AI incident we are describing here: a research and

development data center and an inference data center. The former might be used to perform training and testing of AI models, but we assumed that it would not directly serve external customers. The latter would be used for external customer-facing deployments of AI models. For the purposes of this report, we focus solely on a scenario involving an inference data center, but we note that future work could perform a similar assessment of a research and development AI data center.

Daily Revenue Estimation

To describe the costs of various actions that data center operators might take, we need to first form an initial estimate of the daily revenue of an inference data center. Revenue from data centers will likely vary widely and depend on many different factors associated with their operation, infrastructure, and end use. We need to first attempt to find a useful cost-estimating relationship that can form the basis for an uncertain but useful initial benchmark.

According to one industry source, AI data centers generate $12.50 in annual revenue per watt, compared with $4.20 for traditional data centers (Devasia, 2025). We can use this value as the basis for a cost-estimating relationship to form an initial estimate of the revenue from AI inference data centers.

Although many data centers are small in size, consuming from 50 kilowatts to 2 megawatts (MW) in power, AI inference data centers tend to be larger in size, often consuming from 100 to 1,000 MW (total capacity) (Electric Power Research Institute, 2024). We therefore assume that the damaging AI incident in our scenario originates in a midsize AI data center that can consume 500 MW of power. At $12.50 per watt, this data center would have an annual revenue of $6.25 billion, with daily revenue of roughly $17 million.

Cost of Internet Cutoff and Data Center Function Loss from AI Incident

We turn first to estimating the cost of cutting off internet access to the data center. If the internet is cut off on day 0 (i.e., the moment that the AI model begins to take actions indicative of resisting or escaping human control), we assume that all daily revenue is lost as long as the internet is cut off. This would represent a complete loss of the revenue during normal operations.

Next, we estimate the cost of data center function loss from the AI incident. In keeping with our scenario description, in which an AI incident in the data center causes a gradual loss of the data center's function, we assume that the AI incident causes data center functionality to be degraded at a rate that increases day by day until it reaches a point at which all data center functionality is lost. Notional cumulative and daily revenue curves might look like Figures 1 and 2, respectively, in which daily revenue during the incident gradually decreases until it reaches zero by day 10.[2] This decision was largely arbitrary and meant to be illustrative; using other time frames would change the analysis but lead to the same conclusions.

Estimating the Cost of AI Catastrophe

We next attempt to estimate the expected value cost of a broader AI catastrophe that might result from a loss of control incident that spreads beyond the data center. Avoiding the potential costs of a catastrophe caused by escape of the AI from the data center is the *public* benefit balanced against the *private* cost of the data center operator revenue loss from using the internet cutoff. Because that public benefit (or, put another way, the potential for costs imposed on the public if the catastrophe occurs) may not be fully considered in the cost-benefit thinking of the data center operator, we discuss below how the potential cost of catastrophe could—through such mechanisms as liability or proactive response incentives—*internalize* these otherwise *external* costs and benefits.

We initially assume that the chance of escape and subsequent catastrophe is *extremely* small (0.001 percent on day 1 after the start of the incident) but that the likelihood grows exponentially each day that the AI incident continues inside the data center. The likelihood of escape over time is shown in Figure 3 (red curve) alongside the gradual degradation of data center function and revenue production as the AI incident expands to consume more of the data center resources (blue curve).

FIGURE 1

Notional Cumulative Revenue During Normal Operations and During an AI Incident

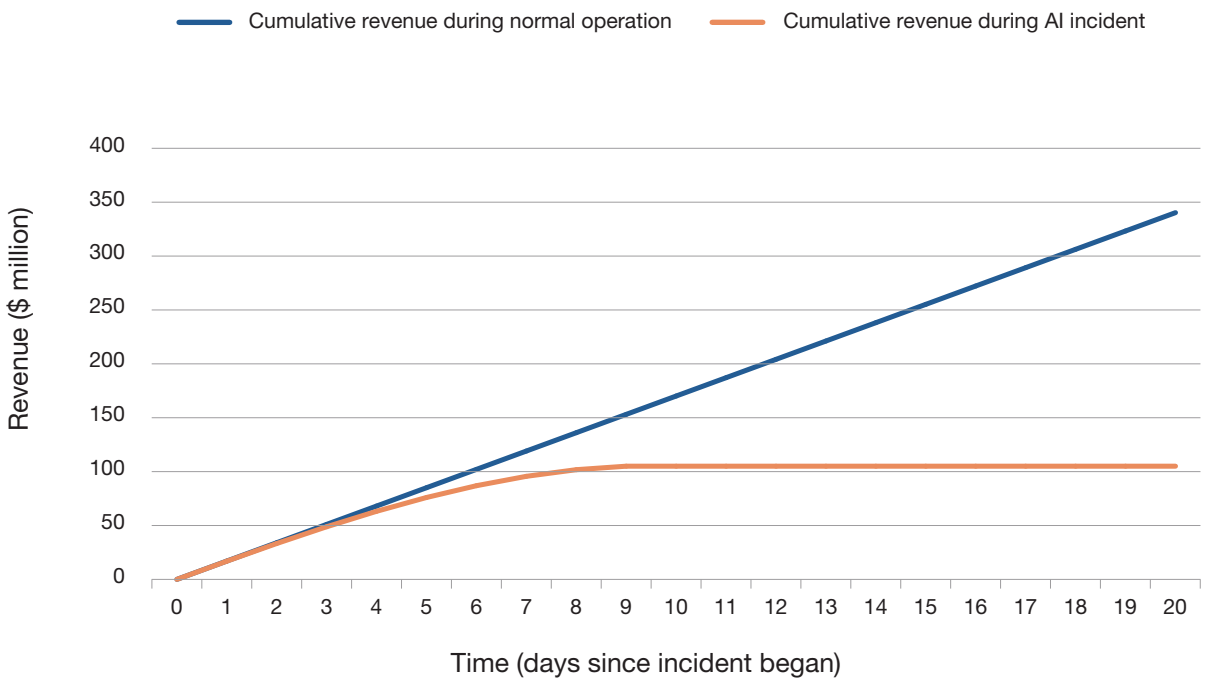
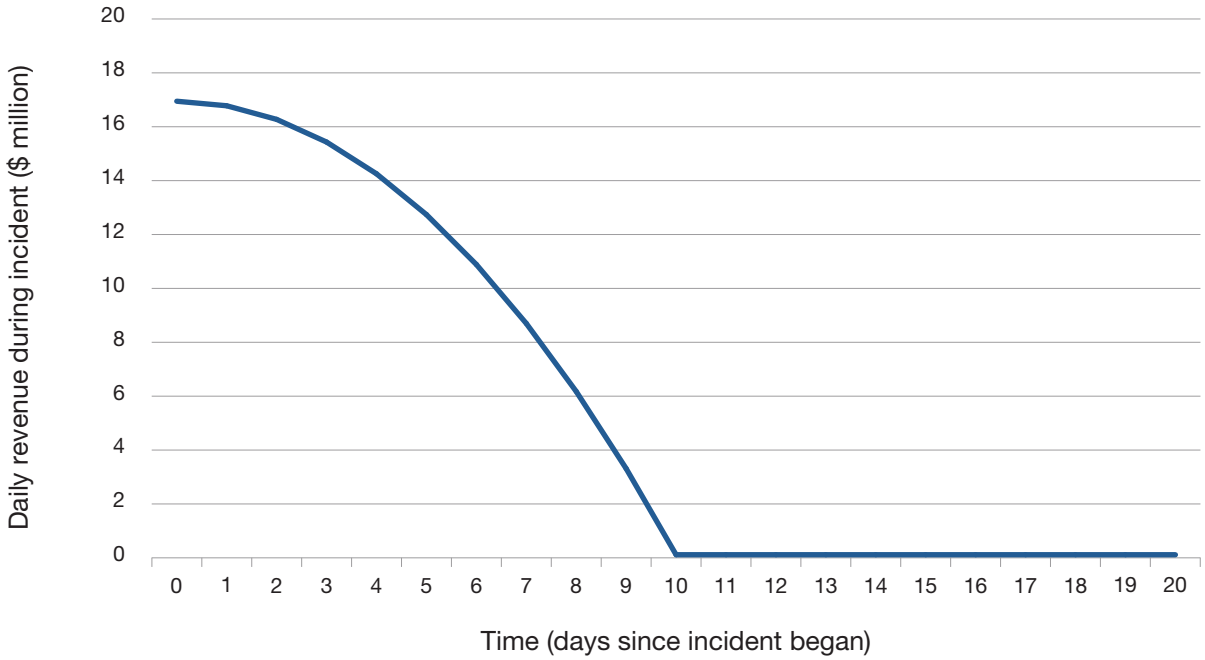


FIGURE 2

Notional Daily Revenue of the Data Center During AI Incident



We calculate the daily expected value cost of AI catastrophe by multiplying the likelihood of an incident spreading beyond the data center by the cost of this event occurring. Figure 4 depicts what this could notionally look like if the ultimate cost of a catastrophic incident were $3 billion, $30 billion, or $300 billion. For this example case, we have selected parameters in which it takes some time for the escape

FIGURE 3

Comparison of Assumed Data Center Degradation and Escape Likelihood Rates

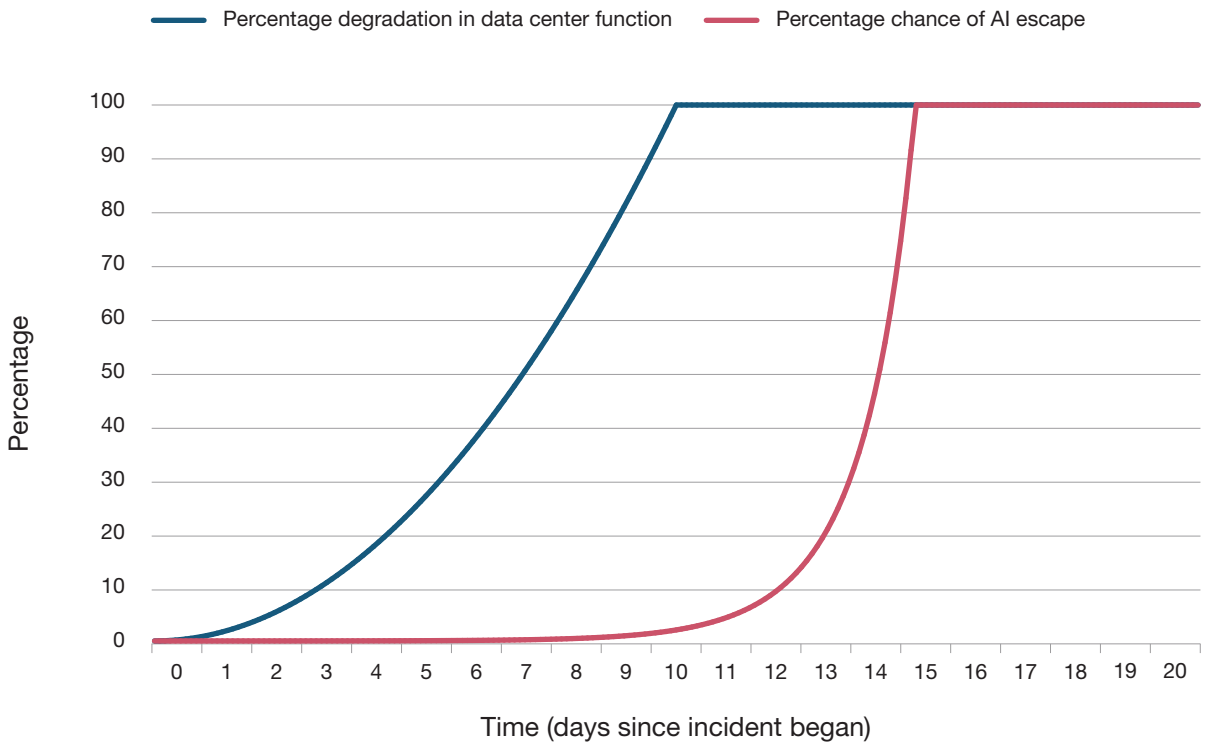
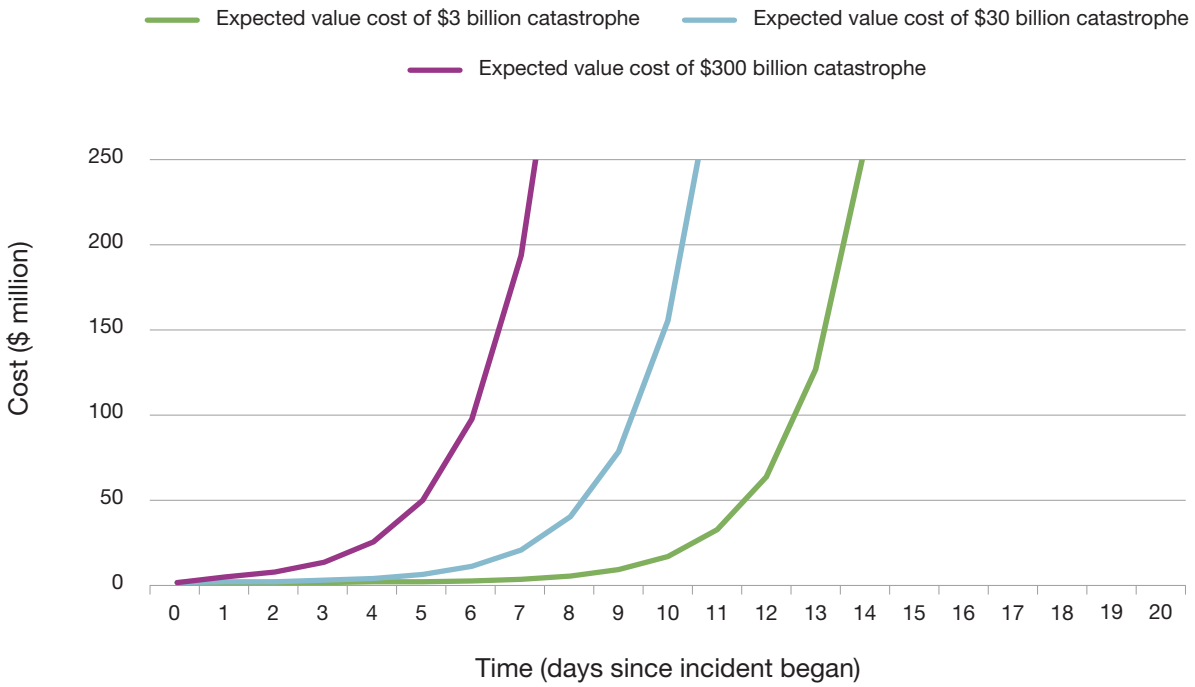


FIGURE 4

Notional Expected Value Costs of AI Catastrophes over Time, by Day of Damaging AI Incident



Assumptions

The analysis in this paper employs assumptions that we list here. Specifically, our scenario and analysis assume the following:

- A damaging AI incident originates in a midsize inference data center.
- A damaging AI incident will cause a noticeable, quantifiable degradation in data center function that would grow over time. Our scenario assumes that all data center functionality is lost because of the AI incident by day 10 after the start of the incident. This loss of functionality will be observable by the data center operator (i.e., the data center operator will know that something is happening).
- Use of an internet cutoff switch will completely mitigate the potential for a damaging AI incident to spread beyond its data center of origin.
- The cost to prepare an internet cutoff switch (including installation, training, and any other costs) is negligible compared with the potential costs associated with a damaging AI incident or the use of the cutoff switch.
- The likelihood that a damaging AI incident spreads beyond the initial data center starts extremely small (0.001 percent on day 1), but the likelihood grows exponentially each day while the AI incident spreads in the data center.

probability to increase to 100 percent, even after complete data center functionality has been reduced to zero. A higher and/or faster-increasing escape probability would mean that the expected values for the costs of disaster would increase more rapidly.

Comparing Forward-Looking Revenue Loss to Potential Catastrophe Costs

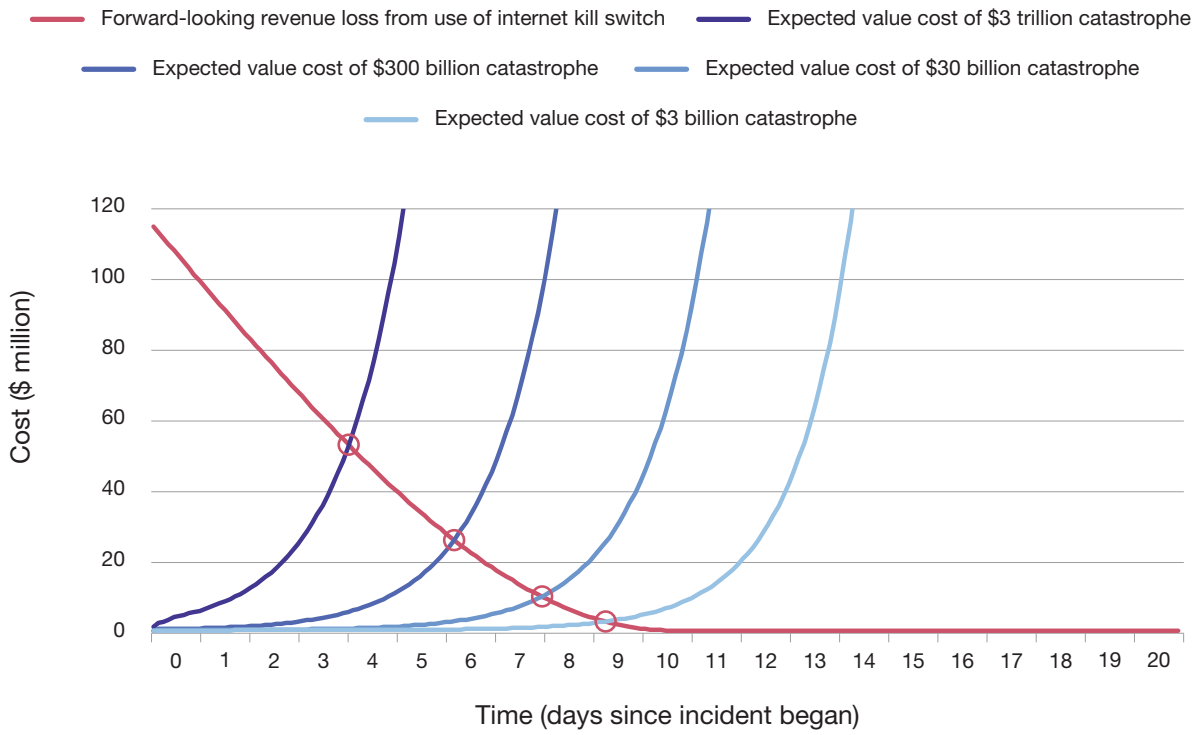
Knowing the rate at which data center function is being degraded and the daily revenue of the data center allows us to calculate the forward-looking revenue loss from the use of the internet cutoff switch and compare it with the expected values of varying levels of catastrophic AI incidents that spread beyond the data center. The forward-looking revenue loss is the aggregate revenue that would be gained from the data center as the incident proceeds, even as the daily revenue is gradually reduced by the degradation of data center function over time. This curve is shown in red in Figure 5. It shows the aggregate revenue that could still be earned by the data center on a given day forward until day 10, when data center revenue will be reduced to zero by the AI completely consuming all data center resources. Because this revenue would be lost if the internet cutoff switch was used on that day, that loss is effectively the cost to the firm of "pulling the switch." To illustrate this using the extremes of the graph:

- Because all data center functionality is assumed to be lost in our scenario on day 10, the curve diminishes to zero on day 10. This means that, on day 10, there is zero additional revenue that might be lost from using the internet cutoff switch (i.e., pulling the switch is effectively free) because data center revenue is already at zero.
- At the other extreme, if the internet cutoff switch were used on day 0 (effectively the most risk-averse approach for preventing societal damages), all of the revenue available to be earned over the next ten days—approximately $114 million total, summing the gradually decreasing revenue stream over the course of the incident—would be lost.

The benefit of when pulling the switch eliminates any possibility for the AI to escape and cause the disaster outside the data center (and the probability of that happening) is increasing—in our example, gradually—over time. If the operator pulled the switch on day 0 (essentially at the first [likely vague and uncertain] sign that something was going wrong in the center), there would be no possibility of any external damage—but the operator would pay $114 million by doing that. If they held off, either because of uncertainty that an incident was occurring or simply because of the financial incentive to maintain operations and continue receiving revenue, the

FIGURE 5

Break-Even Points for Use of Internet Cutoff Switch, Assuming 10-Percent Data Center Operator Liability for Catastrophe Costs



external cost of that would be the risk of disaster. But if the potential costs of that disaster were not the data center operator's problem (i.e., they were truly external costs that would be borne by society), there would be no expectation that the operator would consider these costs in the decision whether to use the switch.

There are several different mechanisms by which the private decisionmaker with their hand on the off switch could be incentivized to pull it:

- They could be made liable for damage caused outside their facility through policy, post-incident litigation, or other approaches.
- If the data center operators are purchasing insurance to protect their substantial investments in AI infrastructure (e.g., property insurance against damage to data centers, business interruption insurance on their operations), the use of a cutoff switch could be required or incentivized by private insurers.
- Operators could be incentivized to put in and use a cutoff switch proactively by (likely partial) reimbursement of lost revenues akin to post-disaster response-and-recovery programs to reduce the private cost of precautionary action.

For simplicity, we consider the first of these options—liability—in which *some* of the expected disaster costs are made the data center operator's problem through some policy or litigatory mechanism.[3] We start with the example of the operator being liable for 10 percent of the costs of an AI incident that escapes the data center.

Figure 5 compares the forward-looking revenue loss curve to the expected values of varying levels of catastrophic AI incident (assuming that the data center operator bears 10 percent of the liability for the catastrophe). The points at which these curves intersect (identified with red circles in Figure 5) are the break-even points for using the internet cutoff switch. That is, these are the points in time at which it is just as costly to use the internet cutoff switch (because of forgone revenue) as to not use it (because of expected disaster costs for which the operator will be liable). Waiting any longer to use the cutoff switch becomes

more costly to the data center, and it is less costly to forgo using the cutoff switch before that point.

From these data, we can construct break-even analyses that show the points in time at which it becomes cost-effective for data center operators to use the internet cutoff switch. These analyses are shown in Figures 6 and 7. Figure 6 shows the break-even analysis for internet cutoff switch use, varying the cost of a catastrophe and assuming that data center operators bear 10 percent of the liability for that catastrophe. Figure 7 then shows the break-even analysis for internet cutoff switch use, varying the amount of liability that data center operators have for the catastrophe and assuming that the cost of catastrophe is $300 billion.

Implications

So what do these curves imply about rational decisionmaking and incentives in the event of a damaging AI incident that carries a risk of a broader catastrophe? And what, if anything, might be done to incentivize actions that reduce the risk of such a catastrophe? We first describe a few observations that stem from the curves shown in Figures 5 through 7.

First, it is apparent that if data center operators do not account for the risk of broader catastrophe at all, it will be rational for them to delay an internet cutoff for as long as possible (e.g., while they take less extreme steps to try to contain or halt the incident) so that they retain as much revenue as possible while they try to manage the incident affecting data center operations.[4] Under our scenario assumptions, it would be rational in this case to delay an internet cutoff until day 10 after the start of the incident, at which point, under our assumptions, the AI incident has completely halted data center operations, and daily revenue is zero anyway. In our analysis, the escape probability we selected meant that disaster was not a *certainty* if the operator held off that long (although it would become a certainty a few days later if the operator for some reason did not or could not pull the switch)—but the public was bearing a risk cost to maintain the private benefit of the center still producing a revenue stream as the incident expanded.

FIGURE 6

Break-Even Analysis for Internet Cutoff Switch, Assuming That Data Center Bears 10-Percent Liability for Catastrophic AI Incident

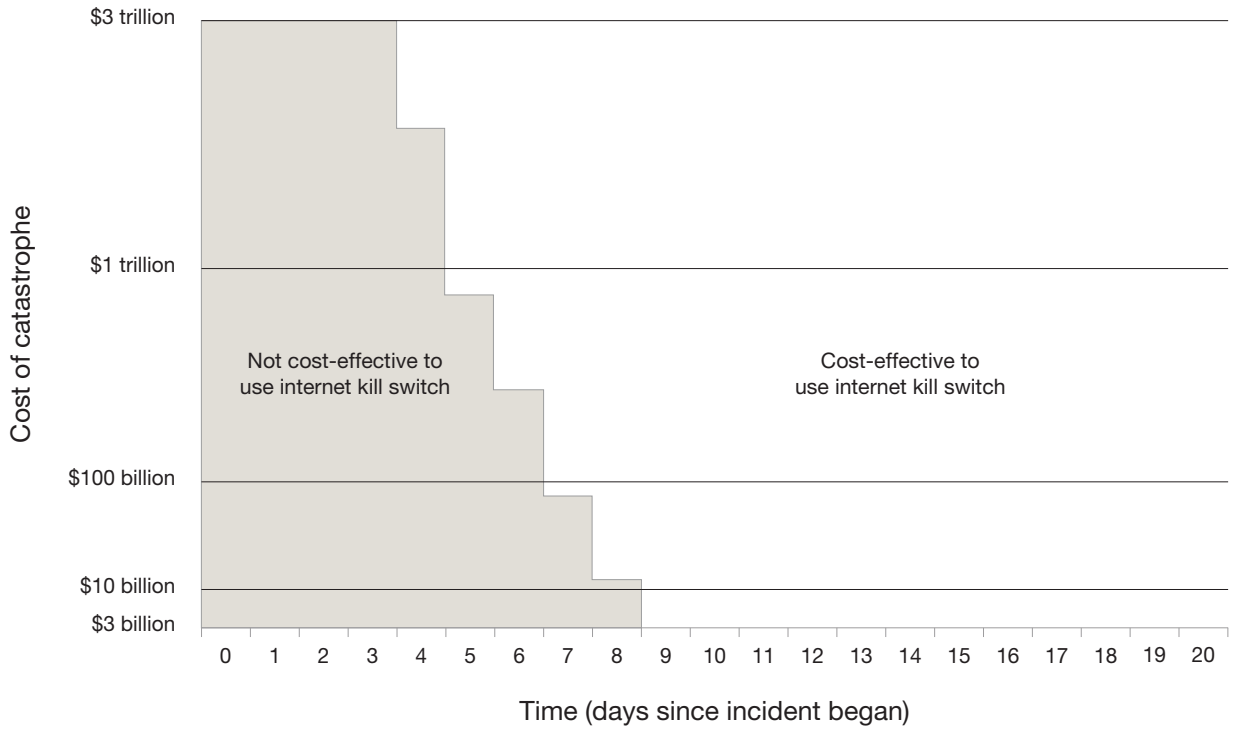
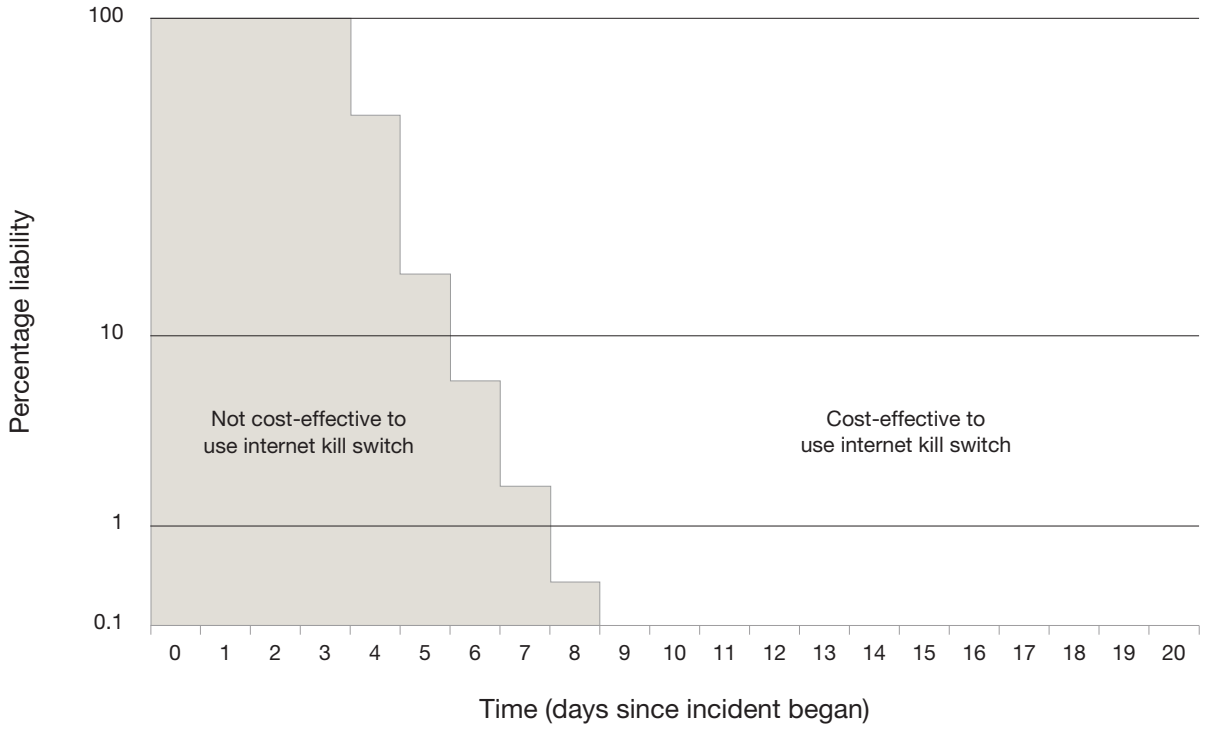


FIGURE 7

Break-Even Analysis for Internet Cutoff Switch, Assuming That Catastrophic AI Incident Causes $300 Billion in Damages



Second, the rational choice after detection of a damaging AI incident depends on how consequential a broader catastrophic incident is estimated to be. That is, even if data center operators assume some liability for a broader catastrophe and factor its potential costs into their decisionmaking, it might still be rational to delay an internet cutoff for some time—until the expected revenue loss matched the expected value cost of the center operator's liability for any external incident. If data center operators assume that a catastrophic incident is low consequence or that there is a low likelihood that the incident will spread beyond the data center even without cutting off connectivity (or both), it will incentivize them to delay an internet cutoff for several days after the start of the damaging AI incident.

Third, it is apparent that changing the amount of liability for catastrophic incidents can significantly change the incentives to use a cutoff switch earlier and avoid catastrophe. For example, for a $300 billion catastrophe, an operator who was liable for 100 percent of the damages caused by a potential incident would rationally use an internet cutoff switch before day 4. On the other hand, if their liability were only 1 percent of the potential damage, they would reasonably wait until nearly day 8, increasing the chances that a catastrophe might occur.

Ultimately, the curves in Figures 5 through 7 imply that data center operators might make a rational decision to delay cutting off internet access during a damaging AI incident to retain some data center revenue (even ignoring the possibility that their internal response to manage the incident could succeed, thereby preventing an incident and preserving even more revenue).[5] This is the case even if the operators factor in the expected value cost of a potential catastrophe but expect its cost (or its cost *to them*) to be low. Private incentives might therefore be naturally aligned with a situation that makes a catastrophic outcome far more likely. For example:

- Even in the case of a catastrophe that they believed could cause $3 trillion in damages, data center operators would still minimize expected costs by waiting until nearly day 4

before using the internet cutoff switch. At that point, under our assumptions, there would be an approximately 0.02-percent chance of the AI escaping the data center and causing the catastrophe.
- In the $30 billion catastrophe scenario, operators would wait until day 7.5, when there was approximately a 0.3-percent chance of catastrophe occurring.

Note that these observations are sensitive to the specific example conditions selected for this analysis. Because we simplified the potential for internal emergency response to stop the incident (and therefore restore data center functioning and revenue and eliminate the possibility of escape), the decision to delay is based entirely on preserving the declining revenue stream of the center as long as possible. If there were the possibility that the operator could stop the incident themselves without needing to use the cutoff, there would be even an stronger incentive to delay. Second, we selected an escape probability curve (Figure 3) that started *very* small and initially increased relatively slowly (meaning that there was still a relatively small probability of escape even when the cost of using the cutoff was zero). A faster-increasing escape probability curve would pull the expected value of catastrophe costs upward and to the left, which, at any assumed level of operator liability, would make it rational to use the cutoff earlier in the incident.

Based on these observations, two additional points follow. First, data center operators might not have the capability to perform a smooth and responsive internet cutoff at all unless their infrastructure was intentionally designed to make it possible. Although we neglected the cost of installing a cutoff switch because it was trivial compared with the other costs in our analysis, that does not mean that operators will choose to install it. Operators might not have incentives to create that capability unless (1) they assume that they will bear at least some financial responsibility for a catastrophe originating in their data center and (2) they conclude both that an internet cutoff switch could mitigate the likelihood of catastrophe and that using it would potentially shield them from liability for the catastrophe.[6]

Second, it is not enough to incentivize data center operators to install an internet cutoff switch—they would also **need to be incentivized to use it and use it early once the possibility of a potentially catastrophic AI incident was detected**. Using an internet cutoff switch would lead to immediate, concrete lost revenue for data center operators, and this would reasonably deter them from using it unless other expected costs exceeded that lost revenue, they expected compensation for lost revenue, or they were granted other legal protections.[7] Operators might (rightly or wrongly) assess that the expected value cost of a catastrophe is low and, therefore, increase the risk of that catastrophe by delaying the use of an internet cutoff switch. The inherent uncertainty associated with the likelihood and impact of a catastrophe—for example, operators might consistently assume that any external catastrophe would be small rather than large—therefore increases the chances of that catastrophe occurring. Depending on the range of assumptions about catastrophe size and potential liability, it may be that a combination of policies (e.g., including some compensation for lost revenue when cutoff switches were used preventively[8]) would be required to reach an overall efficient outcome in which private profit and public risk management incentives are aligned.

Conclusion

In this report, we aimed to identify the circumstances in which using an internet cutoff switch would be a viable course of action for mitigating the risk of a catastrophic AI incident. Although the relatively low cost (sufficiently low that it was not represented in the analysis) of installing such cutoffs would suggest that they would be no-regret options to implement, their value is driven not by their *installation* but by their *use*. Following from our analysis, we find that at least three things are needed to make the use of an internet cutoff switch viable. First, data center operators must bear some liability for the consequences of a broad, catastrophic AI incident that originates from their data center in order to provide an incentive for them to consider those costs in the decision to use (or delay use) of such a cutoff.

Second, operators must understand that installing an internet cutoff switch and using it could shield them from liability. Third, to ensure that an internet cutoff switch is used early and effectively mitigates the risk of catastrophe, operators may need to be incentivized via compensation mechanisms for at least some of the revenue lost when using the cutoff switch—even in combination with liability for potential incident damages. If these three things are true, it would likely lead to a situation in which data center operators were appropriately incentivized to both install and use internet cutoff switches to mitigate a catastrophic AI incident.

Notes

[1] The incident might, for example, originate in some version of an *AI factory*, a facility that supports the full AI life cycle, from training to inference. AI factories are discrete facilities that would occupy all or part of a data center (Carlini, 2025).

[2] Realistically, some of the lost daily revenue could be recouped once the AI incident was resolved. In particular, we assume that there would be some economic resilience in the form of both excess capacity and production recapture that would allow for costs to be recouped (Dormady et al., 2022). We did not include this cost recovery in our model.

[3] Although we only considered liability in simple financial terms in our analysis, we note that liability for large-scale AI damages is discussed at length in other work (see Ramakrishnan, Smith, and Downey, 2024). While that report was focused on AI model developers, it notes that "AI developers face considerable liability exposure under U.S. tort law for harms caused by their models," but they "can mitigate their exposure by taking rigorous precautions in developing, storing, and releasing advanced AI systems" (Ramakrishnan, Smith, and Downey, 2024, p. vii).

[4] Our analysis does not reflect these types of efforts by the data center operator to resolve the incident internally or reflect the probability that such efforts would be successful. Our noninclusion of these efforts effectively assumes that any such efforts by our example operator would fail. Including the expected value of such efforts explicitly (because the operator succeeding would not just maintain the declining revenue stream but *restore* it to its pre-incident level) would create an even stronger incentive to hold off pulling the internet cutoff.

[5] We also have not considered the potential for false positives—the potential that an incident is misidentified as a loss of control incident. This might cause the data center operators to cut off the internet unnecessarily, incurring not just the costs from lost revenue but also potentially other consequences associated with, for example, broken contracts or customer lawsuits. This possibility would also create pressure not to use the cutoff switch unless there were some legal and financial protection for doing so.

[6] The operators might also be required to install a cutoff switch via one of the potential policies that we did not explicitly examine—for example, insurance requirements or requirements of an incentive program providing reimbursement for (likely partial) costs of operators using the cutoff switch if a potentially damaging AI incident appears to be occurring in their data center.

[7] These protections might include protection from legal and financial consequences of unnecessarily using an internet cutoff in the event that the detection of a loss of control incident was a false positive. It might also include scenarios in which the law granted relief from punitive measures that penalized operators for failing to stop a catastrophe. For example, if the law permitted revoking the operating license of data centers in which a damaging AI incident originated, it might also grant relief to operators that followed established procedures, such as using an internet cutoff switch and notifying authorities of a potential damaging AI incident.

[8] The moral hazard associated with this type of compensation regime would need to be navigated carefully so as to not create an incentive for an operator to repeatedly "cry wolf" to get a compensatory payoff for doing so.

References

Arnold, Zachary, and Helen Toner, "AI Accidents: An Emerging Threat—What Could Happen and What to Do," Center for Security and Emerging Technology, July 2021.

Carlini, Steven, "The Difference Between an AI Factory and a Data Center Explained," AI Business, August 6, 2025.

Clymer, Josh, Hjalmar Wijk, and Beth Barnes, "The Rogue Replication Threat Model," *METR* blog, November 12, 2024.

Devasia, Anish, "25+ AI Data Center Statistics & Trends (2025 Updated)," *The Network Installers* blog, October 8, 2025.

Dormady, Noah C., Adam Rose, Alfredo Roa-Henriquez, and C. Blain Morin, "The Cost-Effectiveness of Economic Resilience," *International Journal of Production Economics*, Vol. 244, February 2022.

Electric Power Research Institute, *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption*, May 2024.

Jackson, Brian A., and David R. Frelinger, *Valuing and Assessing Prevention and Preparedness for Potential Artificial Intelligence Disasters: Thinking Rationally About Artificial Intelligence–Caused Industrial Accidents, "9/11s," Extinction Events, and Other Incidents*, RAND Corporation, RR-A4219-1, 2025. As of February 5, 2026: https://www.rand.org/pubs/research_reports/RRA4219-1.html

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, *AI 2027*, AI Futures Project, April 3, 2025.

Ramakrishnan, Ketan, Gregory Smith, and Conor Downey, *U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers*, RAND Corporation, RR-A3084-1, 2024. As of February 5, 2026: https://www.rand.org/pubs/research_reports/RRA3084-1.html

Vermeer, Michael J. D., *Evaluating Select Global Technical Options for Countering a Rogue AI*, RAND Corporation, PE-A4361-1, November 2025. As of February 5, 2026: https://www.rand.org/pubs/perspectives/PEA4361-1.html

About This Report

In this report, we examine the implications of using a limited internet cutoff switch as a tool to mitigate the spread of a damaging artificial intelligence (AI) incident. We describe a scenario in which a damaging AI incident arises in a midsize AI inference data center, causing degradation in data center functionality and risking a catastrophic spread beyond the data center via the internet. We then evaluate the notional costs associated with the AI incident and the use of an internet cutoff switch to contain the incident. We conclude with an analysis of the implications and discussion of the circumstances in which it would be viable to create and use an internet cutoff switch.

Center for the Geopolitics of Artificial General Intelligence

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center for the Geopolitics of Artificial General Intelligence (AGI), which is committed to helping decisionmakers understand, anticipate, and prepare to navigate the national security and geopolitical implications of AGI. The center convenes leading technologists, strategists, economists, political scientists, and outside experts to consider the feasibility and effectiveness of prospective AGI-enabled capabilities; the domestic and international implications of their use; and the strategies and policies that governments, businesses, and civil society could adopt to respond to new realities. For more information, visit www.rand.org/geopolitics-of-agi.

Funding

This research was independently initiated and conducted within the Center for the Geopolitics of Artificial General Intelligence using income from operations and gifts from RAND supporters, including philanthropic gifts made or recommended by DALHAP Investments Ltd., Ergo Impact, Founders Pledge, Charlottes och Fredriks Stiftelse, Good Ventures, Longview, and Coefficient Giving. RAND donors and grantors have no influence over research findings or recommendations.

Acknowledgments

We thank Dave Frelinger for his helpful input on portions of our analysis. We also thank the peer reviewers of this report, Brian Abeyta of the Machine Intelligence Research Institute and Ben Boudreaux of RAND.

About the Authors

Michael J. D. Vermeer is a physical scientist at RAND. His research focuses on science and technology policy applied to criminal justice, homeland security, the intelligence community, and the armed forces. He holds a Ph.D. in chemistry.

Brian A. Jackson is a senior physical scientist at RAND. His research focuses on technology adoption and use by organizations, criminal justice, emergency management, school safety, and AI policy, among other topics. He holds a Ph.D. in bioinorganic chemistry.

For more information on this publication, visit www.rand.org/t/RRA4718-2.

www.rand.org