Vulnerability Analysis for AI Crisis Preparedness

Arnav Saxena
saxen196@gmail.com

Bailey Hirota
baileyhirota@icloud.com

Mentor: Isaak Mengesha
isi.mengesha@gmail.com

Abstract

AI safety efforts are heavily focused on prevention, yet complex failures will still occur in real-world conditions. This paper analyzes a set of AI-induced crisis scenarios to identify the institutional and coordination capacities required once harm begins to unfold. Using structured failure-mode decomposition across six crisis response mission areas, we map stressors, failure points, and response needs across scenarios. We find that recurring gaps, particularly in detection, cross-actor coordination, and containment, systematically limit the ability to contain high-impact failures.

1 Introduction

Frontier AI governance has largely emphasized prevention: improving model alignment, restricting deployment, and mitigating foreseeable risks. This focus is understandable, and from an expected-value perspective oriented toward long time horizons, it reflects a rational prioritization of existential risk reduction. Yet this framing rests on an assumption that deserves scrutiny: that sufficiently robust preventive measures can forestall serious failures, and that the primary governance challenge is one of design rather than response. The emergence of AI-enabled cyber capabilities illustrates why this assumption warrants closer examination. When Anthropic disclosed that its Mythos Preview model had autonomously identified thousands of previously unknown vulnerabilities across every major operating system and web browser [1], it exemplified a class of AI-induced failures that resist clean jurisdictional assignment or advance warning—not because safeguards failed, but because capabilities outpaced existing defensive and institutional infrastructure.

The unpredictability of such failures is not an edge case but a structural feature of complex socio-technical systems, and it is the animating premise of this paper. We argue that mitigating catastrophic AI-induced failures—and extracting institutional learning from them—is itself instrumentally valuable for long-run existential risk reduction, in addition to reducing near-term harms. The question this paper addresses is therefore not only how to prevent AI crises, but what happens after failure begins.

2 Background & Setup

This project adopts a crisis-response perspective on AI risk, focusing on system behavior once failures are already underway. We define an emergency in line with the UK Civil Contingencies Act (2004) as an event or situation that threatens serious damage to human welfare, the environment, or national security [8]. We further define catastrophic incidents following FEMA as those that produce extraordinary levels of disruption and rapidly exceed the response capacity of existing institutions [3].

Our analysis builds on prior work identifying a coordination gap in frontier AI governance: existing policies focus heavily on prevention while under investing in the institutional capacity required to respond when failures occur [7]. This gap is structural, arising from the distributed nature of responsibility across actors such as governments, AI labs, and infrastructure providers.

To examine this problem, we apply a structured failure-mode decomposition framework adapted from emergency preparedness methodology. Each scenario is analyzed against six crisis response mission areas: detection and situational awareness, attribution and diagnosis, escalation decisions, cross-actor coordination, containment and mitigation, and recovery and remediation. For each mission area, we identify what an adequate response would require, which actors would need to provide it, and why existing capacity is likely to fall short. This structure allows gaps identified within individual scenarios to be compared across scenarios, surfacing which deficits are specific to particular failure types and which are structural.

3 Main Argument

We analyze AI crisis scenarios through a structured failure-mode decomposition framework adapted from FEMA National Planning Scenarios [4] to illustrate these failure dynamics concretely. For each, we construct a case for plausibility by grounding the scenario in documented deployment trends, historical near-misses, and the pace of AI integration into critical systems, before analyzing where and why response would break down. Notably, we focus on each scenario's cross-actor coordination: which actors would need to act in concert, what authority or information each lacks, what incentive structures work against timely escalation, and why the institutional arrangements that contained analogous historical failures would prove inadequate here. Below, we present the scenarios in summary form.

Scenario: AI-Exacerbated Power Grid Failure (B. Hirota)

Between 2025 and 2030, three converging trends may create conditions for a novel class of grid failure: rapid AI deployment in utility operations, high vendor concentration in predictive maintenance software, and significant reduction in experienced field personnel as manual inspection workflows are retired.

In recent years, AI deployment has shifted utilities from scheduled physical inspections toward continuous algorithmic risk scoring, which decreases inspection times and increases overall inspec-

tion efficiency. However, this also means that equipment maintenance is increasingly mediated by vendor-supplied models rather than direct operator judgment. Vendor concentration amplifies this dependency: utilities largely rely on off-the-shelf software for their monitoring needs, and the top predictive maintenance platforms collectively serve the majority of the market, meaning model updates can propagate across dozens of utilities in a single cloud deployment cycle. Meanwhile, workforce reduction has independently narrowed the human capacity available for physical inspection—field crews have been reduced by 60–66% relative to the pre-AI baseline in some documented cases, and training pipelines cannot replenish that capacity quickly.

A plausible failure pathway emerges from the interaction of these conditions. In this scenario, a vendor deploys an updated transformer monitoring algorithm containing a systematic bias; for instance, under-scoring fault risk for a specific equipment cohort, across tens or hundreds of client utilities simultaneously. Because the algorithm under-scores risk, affected transformers do not receive timely inspection and degrade undetected over months. Because operators are increasingly trained to interpret dashboard outputs rather than conduct independent inspections, this pattern may go unrecognized until equipment begins to fail. If an external stress event, such as a prolonged heat wave or an unexpected load surge, were to trigger a series of failures in this already-degraded equipment, containment would face compounding constraints. Vendor concentration means failures could be correlated across regions simultaneously, and this, in combination with labor shortages, could eliminate the ability of unaffected utilities to send spare crews and equipment to affected ones—the form of mutual aid that helped resolve the 2003 Northeast Blackout, which left over 50 million people without power and caused an estimated $4–10 billion in damages. Meanwhile, large power transformers are custom-built and take 18 months to 4 years to manufacture and deliver, making rapid equipment replacement infeasible even once the problem is identified.

What this scenario illustrates is a potential coordination vacuum across the key actors involved. The vendor has visibility into its own model but not into cross-utility failure patterns. Individual utilities can observe their own equipment but cannot inspect proprietary vendor algorithms. Regulators have authority over neither in real time. By the time failures are recognized as causally connected, likely requiring forensic analysis across multiple utilities simultaneously, significant harm may already have occurred, and no pre-established mechanism exists to trigger a coordinated system-level response across all three.

Scenario: AI-Driven Fragility in Credit Underwriting (A. Saxena)

Between 2023 and 2026, the UK's non-bank specialist lending sector (NBSLS) grew to supply nearly 30% of SME financing, most of this growth was driven by AI-powered underwriting. The non-bank lenders operate differently to traditional bank lending (which relies on human judgement complemented by credit scoring models). Non-bank lenders increasingly depend on third-party machine learning systems that take in large amounts of data (default, macroeconomic, borrower characteristics etc.). The Bank of England's FPC has identified this shift as a key area of concern, noting that 'a scenario in which AI models are increasingly deployed in an autonomous manner

could potentially pose significant additional risks to financial stability [2]. The FPC specifically warns that 'common weaknesses in widely used models could cause many firms to misestimate certain risks and so misprice and misallocate credit as a result' [2]. This scenario examines how such common weaknesses could manifest as a systemic crisis.

Three structural mechanisms drive this fragility, each compared closely to the 2010 Flash Crash. This is the closest historical analogue with algorithmic homogeneity causing the system collapse. The LSE Business Review observed in 2017, the Flash Crash was the 'first market crash in the era of algorithms', revealing that a 'single large sell order in the futures market triggered a cascade of algorithmic selling that spread across asset classes' [6]. In credit markets, the same logic applies but with key differences. In the Flash Crash, human traders could recognise that prices had detached, making a move from there. In credit markets, no 'value lender' exists because there is a common consensus that lenders cannot easily override models. The Flash Crash demonstrated how algorithmic homogeneity can produce systemic collapse. The question is whether credit markets could develop a response capacity before such an event can occur.

First, functional homogeneity. An estimated 44% of model-based credit decisions rely on the top three AI underwriting providers [2]. They each use similar model architecture, similar training data, and similar optimization objectives. The FPC's 2025 report explicitly identifies this as a market vulnerability; "The potential for market concentration in AI-related services, including vendor-provided models, is a further challenge" [1]. The report notes that 'around half of the respondents to the 2024 AI Survey report having only a 'partial understanding' of the AI technologies they use' [2]. During the Flash Crash, algorithmic homogeneity across trading firms caused correlated selling, detaching prices from fundamentals rapidly. In credit markets, the same dynamic would cause simultaneous credit withdrawal across dozens of lenders. The problem here, is that its not as visible.

Second, structural adaptation lag. Unlike human lending, AI models cannot incorporate new information until it has been labeled. The FPC recognizes this as a macrofinancial vulnerability: "In the event that large numbers of firms rely on the same open-source model components or data libraries, a significant unknown error or bias could cause many firms to miesestimate certain risks" [2]. This creates a procyclical lock-in, wherein models continue to predict high default probability for solvent borrowers (tightening credit when it should expand). AI introduces a new dimension to the lessons learned from the Great Recession (2008), the speed and simultaneity of model updates means that mispricing can propagate across the entire sector within hours.

Third, accountability fragmentation. Think about the stakeholders. The lender asserts "we used the model". The provider says "we provided a tool". The regulator, herein, lacks authority over the provider. The Treasury Committee's 2012 inquiry into macroprudential tools revealed longstanding concerns about the FPC's powwers. They noted that the "potential for recommendations to be made on such a wide scope increases the policy uncertainty for firms" and that "it is difficult to envisage circumstances where FPC recommendations will not be enacted" [5]. For AI underwriting, the gap is even wider. The FCA cannot compel third-party model providers to

explain nor adjust their outputs. The Financial Services and Market Act 2023 established a new regulatory regime for critical third parties, explicitly in response to the FPC's 2021 recommendation that "additional policy measures were likely needed to mitigate the financial stability risks stemming from concentration in the provision of services to UK firms" [9]. However, this regime is not yet fully operational, and remains untested for AI failure modes.

A plausible failure pathway emerges from the interaction of these mechanisms. Within 48 hours, third-party underwriting models across dozens of non-bank lenders rescore their portfolios against the new data and systematically flag SME borrowers as high-risk. Credit lines are withdrawn/reduced. SMEs, unable to access working capital, begin defaulting. This feeds back into the models as new default data, reinforcing the initial signal. Meanwhile, banks that provide warehousing financing to non-bank lenders observe rising defaults and withdraw their own credit lines, amplifying the original contraction. The FPC has noted that such correlated behaviour is of concern: "A reliance on a small number of providers for a given service could generate systemic risks in the event of disruptions to them, especially if it is not feasible to migrate rapidly to alternative providers" [2].

The regulator lacks authority to compel model overrides, suspend automated underwriting, or mandate rapid retraining. Unlike the Flash Crash, where circuit breakers were installed after the fact [6], no equivalent mechanism exists for credit markets. A credit freeze produces lasting real-economy damage, a characteristic not seen in equity markets.

This scenario illustrates a coordination vacuum across a set of actors: non-bank lenders, third-party AI providers, the FCA, the Bank of England, and the HM Treasury. The FPC has acknowledged that "public-private sector collaboration will be important for helping address AI-related risks" [2], but no equivalent framework exists for credit underwriting failures. No single actor has both visibility and authority. The critical third-party regime from the Financial Services and Markets Act (2023) is a step toward closing the gap but is not yet operational [9].

4 Implications & Future Directions

These analyses demonstrate the value of structured scenario decomposition as a diagnostic method for AI crisis preparedness. Across the examined domains, the framework consistently reveals mismatches between actors and visibility that affects the ability to safety "shut-down" the problematic scenarios. This is not a coincidence of poorly designed regulatory regimes, but it is a predictable consequence of deploying AI systems whose operational logic crosses institutional boundaries drawn before such systems existed.

The scenarios also reveal that this vacuum manifests with different time signatures across domains. Grid failure accumulates slowly over months; however, algorithmic credit contagion can compress a systemic event into hours, potentially outpacing any regulatory response. This variation is crucial because adequate response capacity can look vastly different across domains and scenarios.

What this method cannot yet resolve is whether the coordination vacuum is primarily juris-

dictional, in which no actor has the relevant authority, or incentive-based, meaning an actor has authority but bears the full cost of intervention while diffuse parties capture the benefit. Disentangling these requires further empirical study of how institutional actors have actually behaved in AI-adjacent near-misses, rather than how their formal mandates assign responsibility. That remains an open question this analysis cannot answer, but one the pre-mortem framework helps to precisely locate.

5 Conclusion

This paper has argued that AI governance has disproportionately focused on prevention at the expense of response capacity. Using structured failure-mode decomposition of distinct scenarios, we have identified recurring gaps that would systematically limit the ability to contain high-impact failures once underway.

Two findings merit emphasis. First, the coordination vacuum we identify is not a design flaw but a structural consequence of deploying AI systems. It speaks to the difficulties in managing operations across institutional boundaries. Grid operators cannot inspect proprietary vendor algorithms; the FCA cannot compel third-party model providers. No actor has end-to-end visibility or authority. Second, this vacuum manifests in different time-frames across sectors. Grid failure accumulates over months, while credit contagion can compress a failure-mode into days. The underlying deficit, however, is the same. Pre-established mechanisms for detection, escalation, and coordinated response are absent.

Prevention work should continue, but response capacity deserves comparable (if not equal) attention. Closing these gaps requires more than incremental adjustments to existing frameworks. We offer the point of view that new institutional infrastructure designed for AI failure-modes is a starting point.

References

[1] Anthropic. Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing, 2026. Accessed: April 2026.

[2] Bank of England. Financial stability in focus: Artificial intelligence in the financial system. Technical report, Bank of England, April 2025. Financial Policy Committee report.

[3] Federal Emergency Management Agency. Catastrophic incident annex to the national response framework, 2008.

[4] Federal Interagency Community. National planning scenarios: Created for use in national, federal, state, and local homeland security preparedness activities. Technical report, U.S. Department of Homeland Security, April 2005. Version 20.2 Draft.

[5] House of Commons Treasury Committee. Macroprudential tools: Written evidence. UK Parliament publications, June 2012. Submissions from ICAEW, ABI, Genworth, BSA, CML and others.

[6] London School of Economics Business Review. Flash crash: The first market crash in the era of algorithms and automated trading. *LSE Business Review*, June 2017. Accessed: May 2026.

[7] Isaak Mengesha. The coordination gap in frontier ai safety policies. 2026.

[8] UK Parliament. Civil contingencies act 2004, 2004. Section 1 definition of emergency.

[9] UK Parliament. Financial services and markets act 2023, 2023. c. 29.