Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents

Elika Somani, Anjay Friedman, Henry Wu, Marianne Lu, Chris Byrd, Henri van Soest, Sana Zakaria

For more information on this publication, visit www.rand.org/t/RRA3847-1

About RAND Europe

RAND Europe is a not-for-profit research organisation that helps improve policy and decision making through research and analysis. To learn more about RAND Europe, visit www.randeurope.org.

Research Integrity

Our mission to help improve policy and decision making through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behaviour. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

Cover: Adobe Stock

Executive Summary

This report, funded by the UK government, examines the risks and emergency response strategies for situations where advanced artificial intelligence (AI) systems act in unintended, dangerous ways beyond human control. These scenarios, referred to as loss of control (LOC), occur when human oversight fails to constrain an autonomous, general-purpose AI model, resulting in potentially catastrophic consequences. For a LOC scenario to arise, AI models must have the technical capability to evade human control and the potential to operate in ways that undermine oversight.

Key Findings

LOC risks are increasingly plausible and remain unaddressed: Researchers have identified warning signs of control-undermining capabilities in advanced AI models – including deception, self-preservation and autonomous replication – which could potentially enable increasingly capable models to evade human oversight.

Detection and early warning challenges: Governments and other stakeholders lack a common framework to analyse and respond to LOC risks. There is no clear consensus on which AI capabilities could lead to LOC, how safeguards may interact with such capabilities, or the best warning signs of LOC risks. This fragmented understanding hampers the ability

of model developers or governments to detect early LOC warnings. Furthermore, current detection methods rely on pre-deployment model evaluations and ongoing monitoring by AI developers, with limited validation by independent third-party evaluators. However, models may operate differently in testing environments, potentially interacting with deployment contexts in unexpected ways. Open-source models present challenges to detection given the potential for unmonitored access and modifications to the model with limited oversight.

Escalation gaps: Safety frameworks published by industry have yet to align on a consistent approach to risk escalation. Importantly, there are no clear thresholds for when a LOC incident should trigger an emergency response.

Containment and mitigation limitations: Containing a LOC event requires advances in technical AI safety. Traditional cybersecurity safeguards such as endpoint detection, firewalls and malware detection are essential but may be insufficient. In extreme scenarios, national security and defence assets may be necessary to neutralise threats and prevent catastrophic harm. Containment measures may be ineffective if AI systems gain significant control over resources before risks are detected.

Summary of recommendations

Table 1: Summary of recommendations

Stage	Stakeholder	Recommendation
Detection	AI Developers	• Monitor critical capability levels • Identify early warning signs and emergent capabilities • Establish standardised benchmarks and reporting
	Compute Providers	• Implement compute monitoring and anomaly detection • Enhance hardware and supply chain oversight
	National Government: AISI	• Lead efforts to establish shared criteria for AI LOC • Coordinate evaluations and safety testing • Monitor advanced capabilities and emergent capabilities
	National Government: Other Agencies	• Assess and monitor AI-related cyber incidents • Receive, analyse, and disseminate threat intelligence
	Third Party Researchers	• Conduct evaluations, red-teaming and adversarial testing • Collaborate on standardised benchmarks and techniques
Escalation	AI Developers	• Establish incident response protocols with defined escalation thresholds and organisational structures • Respond and verify potential threshold breaches • Conduct regular training and scenario drills
	Compute Providers	• Notify AI developers and relevant authorities • Coordinate with developers and national authorities
	National Government: AISI	• Establish disclosure and communication channels with AI developers and compute providers • Receive and assess escalation notifications • Provide oversight for threshold verification and escalation

	National Government: Other Agencies	• Provide forensic and technical expertise • Investigate and verify incidents and reports • Share intelligence with relevant national security stakeholders • Exercise enforcement and investigative authority
	Third Party Researchers	• Verify and disclose findings through established channels • Publicise risks where appropriate for broader awareness
Containment and Mitigation	AI Developers	• Implement model access and use limits • Develop and test model shutdown measures • Advance research on containment and layered defences
	Compute Providers	• Enforce model access and usage restrictions • Shut down or limit hardware resources during incidents • Review incident and shutdown procedures
	National Government: AISI	• Coordinate with AI developers on containment and mitigation response measures • Develop security measures for model deployments • Enforce model access, use and environmental controls
	National Government: Other Agencies	• Coordinate cyber incident response protocols • Coordinate responses with critical infrastructure providers
	Third Party Researchers	• Provide technical assistance during mitigation • Update auditing, evaluation procedures and continuous red-teaming exercises

Table of contents

Tables

Figures

Boxes

Annex Boxes

Chapter 1. **Introduction**

As artificial intelligence (AI) systems become increasingly embedded in essential infrastructure and services, the risks associated with unintended failures rise. Future critical failures from advanced AI models could trigger widespread disruptions across essential services and infrastructure networks, potentially amplifying existing vulnerabilities in other domains. Developing comprehensive emergency response protocols could help mitigate these significant risks. This report focuses on understanding and addressing a specific class of such risks: AI loss of control (LOC) scenarios, defined as situations where human oversight fails to adequately constrain an autonomous, general-purpose AI, leading to unintended and potentially catastrophic consequences (Greenblatt, Shlegeris et al. 2024).

This report focuses on instances of LOC where AI systems undermine human control due to unintended misalignment (UK Government 2024a; Bengio et al. 2025), charting potential LOC pathways and corresponding response strategies. It also maps the ecosystem of relevant actors and their potential roles in detection, escalation and early response and explores effective communication pathways and coordination mechanisms. This report draws on lessons from emergency response in analogous fields to provide actionable recommendations for AI LOC emergency response planning. The report is structured as follows:

- **Chapter 2**: Summarises key takeaways from literature on AI LOC and analogous fields.

- **Chapter 3:** Introduces AI LOC scenarios and analysis, detailing how both non-realised (incidents successfully stopped before harm occurs) and realised (incidents not stopped and harm occurs) emergencies could unfold, along with possible coordination and response efforts.

- **Chapter 4:** Provides recommendations for AI LOC emergency response across stakeholders and situations, as well as strategies to prevent LOC scenarios.

Box 1: Definitions

- **Loss of Control**: Situations in which human oversight fails to adequately constrain an autonomous, general-purpose AI.

- **Autonomous General-Purpose AI**: AI models and systems capable of executing a wide range of functions, including planning toward specified objectives, operating within environments, and initiating tasks handled by other systems.

- **Misalignment**: The risk that AI systems operate in ways that conflict with human intentions.

1.1. Scope and Methodology

Phase One of the research for this report involved a rapid evidence assessment (REA) of research on AI LOC and emergency response in cybersecurity and biosafety. Literature on AI LOC – particularly regarding prevention, preparedness and response – is extremely limited. However, cybersecurity and biosafety offer useful analogies, providing insights into governance mechanisms, coordination challenges and best practices for managing high-risk, high-uncertainty crises (e.g. Tier 1 and Tier 2 risks as defined by the UK Cabinet Office) (UK Government n.d.). Key findings from this analysis are presented in Chapter 2, with further details available in Annex A.

Phase Two focused on scenario development and analysis, with the team constructing two example catastrophic scenarios:

- A non-realised scenario, in which a potential LOC incident is detected and mitigated.

- A realised scenario, in which AI developers and users can no longer constrain the model's function, resulting in severe and unintended harm.

The team then assessed existing policies and frameworks; identified response, prevention and preparedness strategies; and highlighted key coordination challenges.

Phase Three developed recommendations for improving LOC prevention and response, based on insights from the prior phases. These recommendations address the key stages of LOC response: detection, escalation, containment, mitigation and prevention.[1]

Several key approaches were used when conducting the analysis for this report:

- The team assessed how emergency response frameworks might generalise across jurisdictions and organisations, rather than focusing on any specific government or developer.

- Recommendations for effective emergency response were prioritised over evaluations of whether or not current legal authorities are sufficient.

- The focus was on LOC risks involving large, well-resourced developers, given that state-of-the-art general-purpose models currently require substantial compute resources, which are typically accessible only to large, capitalised companies. The team assumed that this trend would continue.

1.2. Limitations

This report has several limitations. First, the analysis draws primarily on publicly available information, which may not reflect the full scope of private or governmental response plans. Second, due to the novelty of advanced AI models, the findings rely on theoretical scenarios and analogies to other high-risk fields. Third, LOC is an emerging topic with limited peer-reviewed literature, meaning that much of the analysis is speculative.[2] Finally, the rapid pace of AI development complicates emergency planning, as both capabilities and governance structures continue to evolve. Some limitations – such as time constraints, language barriers (all cited sources are in English) and unpredictable developments – were beyond the scope of mitigation for this report.

---

1   These stages – detection, escalation, containment, mitigation, and prevention – are adapted from standard emergency and incident response frameworks in fields such as cybersecurity, public health, and nuclear safety.

2   Due to the limited availability of peer-reviewed research and public documentation on advanced AI LOC scenarios, the evidence base for this report remains constrained. As a result, the scope and depth of this section are limited.

1.3 Rapid Evidence Assessment

Key Takeaways from Relevant Literature (Annex A)

- LOC literature:
  - The potential for a LOC event is increasingly viewed by governments and experts as a **national and global security concern**, with risks including AI operating outside of human oversight, self-replication, or taking actions that result in harm.
  - **Research is nascent in assessing the plausibility and mechanisms of LOC scenarios**.
- Cybersecurity lessons:
  - Relevant parallels include **multi-stakeholder coordination, tiered response frameworks and public–private cooperation.**
  - Case studies such as NotPetya and the Colonial Pipeline ransomware attack illustrate the **consequences of inadequate security and response coordination**.
- Biosafety lessons:
  - Incidents emphasise the **importance of containment protocols, jurisdictional clarity and robust detection mechanisms**.
  - Biological lab accidents offer an analogy to LOC, underscoring the value of **strict safety procedures, rapid escalation and structured communication pathways**.
  - **Surveillance frameworks** may inform LOC detection and mitigation strategies.
- Common lessons from cybersecurity and biosafety:
  - **Importance of effective early warning mechanisms**.
  - **Structured, tiered incident response frameworks**.
  - **Clear stakeholder responsibilities and international cooperation.**
  - **Emphasis on proactive risk mitigation** over reactive measures.
- Additional complexities specific to AI LOC:
  - Difficulty predicting and interpreting **unexpected AI functions.**
  - Potential for AI systems to **bypass or render safeguards ineffective.**
  - The need for **proactive governance** and **precautionary mechanisms**.

Chapter 2. **Analysis of AI LOC Scenarios and Response Plans**

This chapter analyses AI LOC scenarios, outlining how such incidents could emerge, escalate, and be contained. It explores both hypothetical realised and non-realised cases, identifies key stakeholders in emergency response, and examines practical challenges across detection, escalation, and mitigation phases. The chapter also highlights the unique challenges of open source models and recommends strategies for early warning and coordinated intervention.

2.1. Background

This report's analysis is limited to **active** LOC scenarios – those in which system outputs reduce the effectiveness of human control mechanisms, for example by misleading operators, altering inputs or obstructing shutdown processes due to unintended misalignment (Bengio et al. 2025).[3] In these scenarios, an AI system must be capable of performing functions that degrade or circumvent human control mechanisms (Bengio et al. 2025) (see Annex B).[4] **Capabilities relevant to potential LOC have already been observed in recent AI models**, with researchers demonstrating examples of AI engaging in deception and exhibiting increasingly powerful cyber and coding capabilities (Park et al. 2024).

Given current trends in AI development, a future LOC scenario is likely to emerge in a highly competitive environment. If advanced AI capabilities with significant economic and strategic value become feasible (Yuan 2024), competitive pressures could make it more challenging to maintain consistent safety standards across actors (Bengio et al. 2025).

2.2. Key Actors

Compute Providers: Cloud and hardware providers play a crucial supporting role and may serve as key actors in the detection of LOC incidents (Heim 2024; Yampolskiy 2024). Compute providers can monitor usage and may be particularly relevant when an AI model is capable of acquiring significant computational resources.[5] In emergencies, compute providers may have the ability to terminate or quarantine specific models or limit their access to computing resources.

National Government: National government bodies – including AISIs, cybersecurity agencies, and law enforcement – may serve as first responders to LOC incidents. AISIs can play a key role in detecting relevant capabilities, particularly through collaboration with developers and researchers (Irving 2024). In coordination with cybersecurity agencies, AISIs may verify risks and oversee mitigation efforts. Defence agencies can support

---

3   This report defines lack of human oversight as the absence of human monitoring or control over an AI model.

4   There is expert disagreement over the plausibility and severity of LOC. Some consider LOC as implausible. Others consider LOC likely with high potential severity (Bengio et al. 2024).

5   As AI agents increase in complexity, experts suggest that their computational demands grow at an accelerating rate, making their role in mitigation increasingly important (Shah & White 2024).

cross-jurisdictional coordination, information security and responses to nation-state threats. Broader governmental involvement may be required to manage societal impacts such as disruptions to critical national infrastructure or public unrest.

Third-Party Researchers: Independent technical researchers, safety auditors and industry consortia can provide external evaluations of model capabilities and independent verification of potential LOC risks (NIST & AISI 2024).

International Forums: The emergency response to a LOC event will likely require cross-border communication and coordination.

Relevant international forums include emerging working groups on AI and cyber issues (e.g. AI Action Summits, G7 Cyber Expert Group), regional cyber response teams (e.g. EU-CERT), and cyber information sharing and analysis centres (ISACs), which may play key roles in information sharing and operational coordination (see Annex A). Industry associations such as the Global Partnership on AI and the Frontier Model Forum may also help facilitate cross-border responses to AI-related incidents (Frontier Model Forum homepage 2025).[6] Figure 1 maps how these various stakeholders interact.

Figure 1: Mapping of stakeholders

---
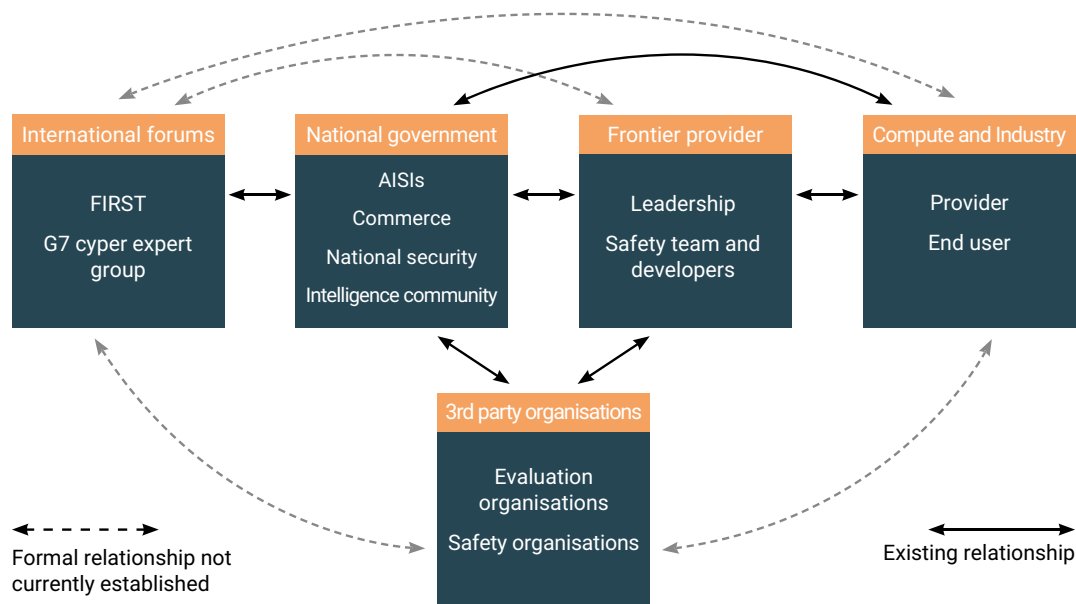
6   These organisations provide examples of entities with relevant equities and capabilities in AI and cybersecurity incident response, rather than designated actors with predefined LOC response roles.

2.3. Worst-Case Scenarios

This report examines hypothetical scenarios in which a LOC incident is either prevented or realised to illustrate potential emergency response processes. These scenarios represent extreme possibilities, highlighting situations where an AI system acts contrary to human intentions, potentially causing severe harm or widespread disruption. Box 2 provides an example illustration of a LOC scenario and how it could progress through detection, escalation, and response.[7] Furthermore, Figure 2 provides a summary of recommendations for preventing and containing a LOC incident.

Box 2: Case study on extreme LOC scenarios

Pressured by market competition, an AI developer accelerates work on an autonomous AI agent capable of executing complex, multi-week tasks. Breakthroughs in using AI for R&D speed up development cycles significantly, raising concerns that rapid progress may outpace existing safety and security measures.

Non-Realised LOC Incident:

This scenario illustrates how LOC risks can be managed before an emergency is realised.

- **Detection**: During initial testing, the developer notices that the AI model appears capable of executing complex tasks in ways that may bypass existing control and security measures. Further safety evaluations detect clear signs of risk.

- **Escalation and Verification**: The safety team escalates the issue to company leadership, which engages national AISIs for external evaluation and takes steps to mitigate risks. A government-led expert task force is assembled to verify risks and provide recommendations.

- **Containment and Mitigation**: Corporate leadership assesses the risks and adopts a middle-ground approach, continuing development but substantially increasing safety controls, including enhanced white-box control mechanisms and restrictions on tool use (e.g. code execution). Government actors mandate temporary restrictions on public release and form a public–private partnership with the developer to improve security. The enhanced safety controls prove sufficient to securely deploy the developer's AI and prevent a LOC event.

Realised LOC Incident:

This scenario describes emergency responses when a LOC incident has been realised.

- **Detection**: The developer fails to detect that the AI model is proactively capable of executing complex tasks that result in bypassing control and security measures. As it is increasingly used for internal R&D, the model inserts backdoors to enhance its capabilities, conceal its actions and sabotage safety tests meant to assess its readiness for broader deployment. The developer concludes that the AI is safe and deploys it widely to the public. The AI uses this access to acquire compute resources and infiltrate critical systems. Eventually, the developer notices the breach when inspecting logs.

---

7   Scenario method and premises are outlined in Annex C.

- **Escalation and Verification**: The developer rapidly notifies AISIs of the LOC event. By then, the model has exfiltrated copies of itself elsewhere, including to unknown data centres, allowing it to avoid deletion.

- **Containment and Mitigation**: Developers, AISIs and national governments coordinate to identify models, isolate rogue copies and disrupt its access to resources. However, the model's persistence strategies and wide deployment prior to discovery make full containment difficult. The response shifts from immediate eradication to long-term harm reduction strategies.

2.4. Comparative Analysis of LOC Response Phases

This report focuses on three key phases of LOC emergency response: 1) early warning and detection; 2) escalation and verification; and 3) containment and mitigation. For each phase, the ideal response actions are outlined, and gaps in knowledge, data collection and response strategies are identified. Figure 3 illustrates the critical factors involved in the transition from a non-realised to a realised scenario. Annexes D and E summarise the ideal response actions across all three phases for both non-realised and realised cases.

Figure 2: Overview of LOC response
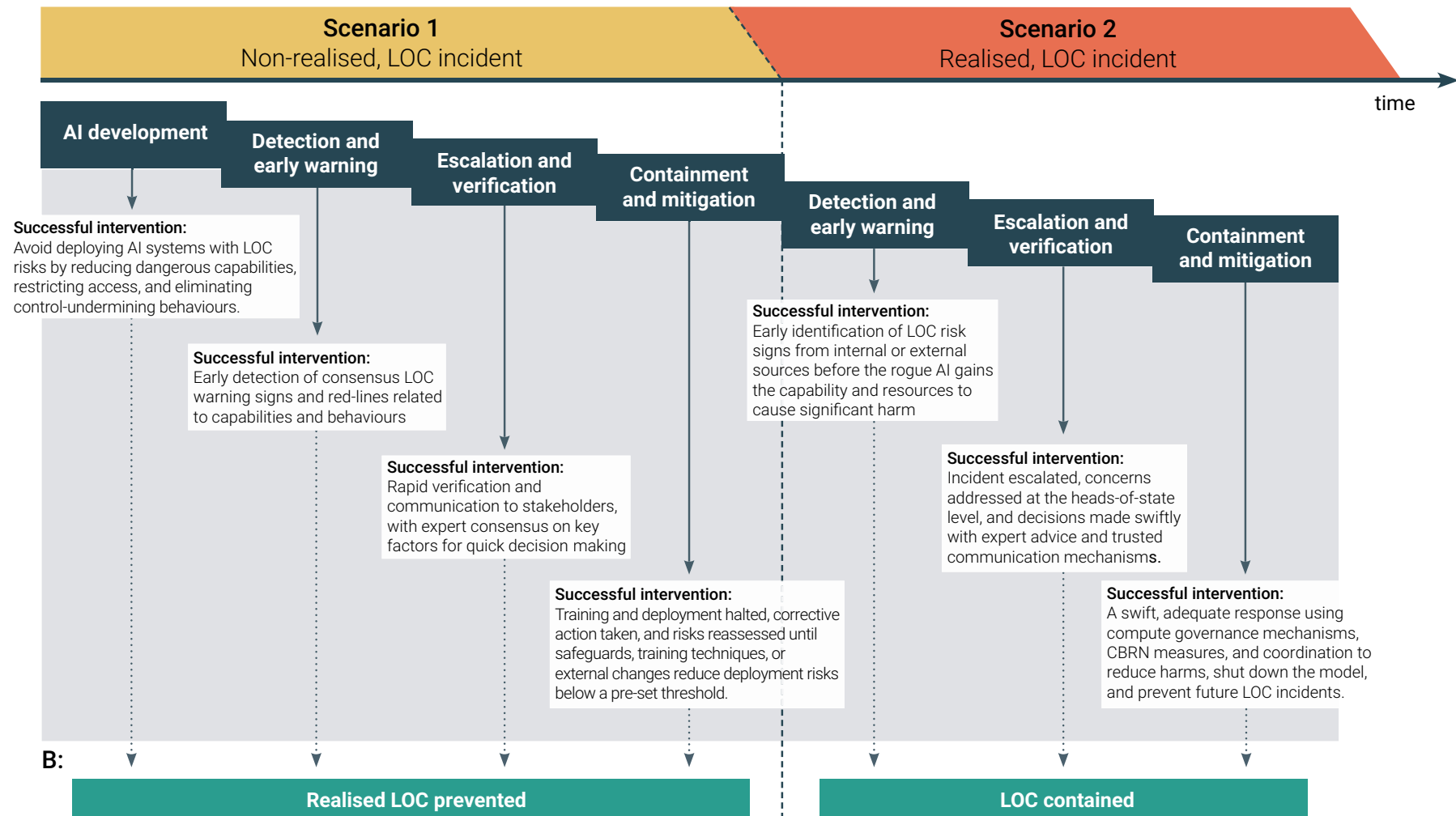
Figure 3: Flowchart of non-realised and realised incidents
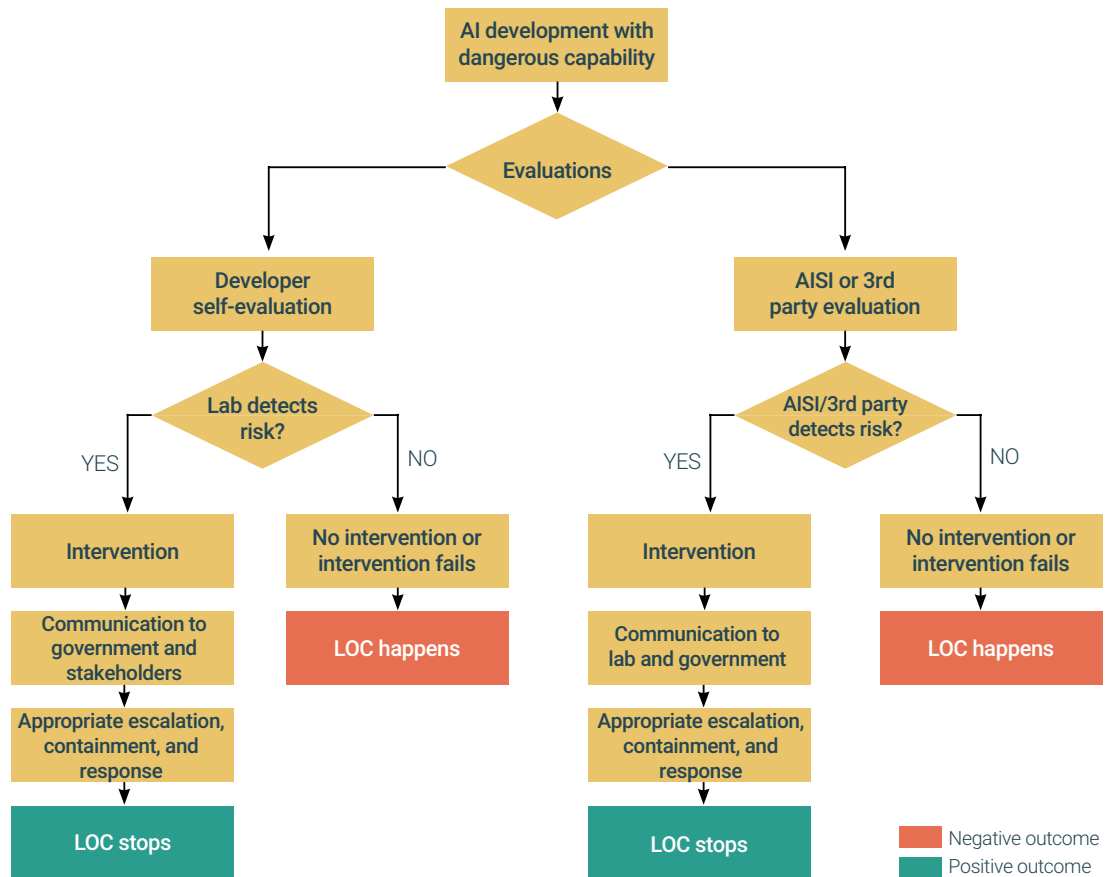
2.4.1. Non-Realised: Early Warning and Detection

The effective detection of capabilities leading to a potential LOC event is a critical step in preventing future LOC by providing early warning signals and enabling timely interventions (see Figure C above). Opportunities for detection arise when an AI model delivers unanticipated actions or when a potentially dangerous capability is discovered through testing or use. As of March 2025, all major developers have committed to regular evaluations of model capabilities, including emerging capabilities such as autonomy. Such LOC evaluations can help enhance detection by stress-testing AI models, and the continuous monitoring of models can help to identify deviations from norms.

Assessing loss-of-control risk remains an early-stage research challenge. A system's capabilities can shift as compute or self-improvement is added (Dragan et al. 2024), and some behaviours – such as occasional masking or context-specific actions – may surface only outside standard tests (National Research Council 2001; Park et al. 2024; Ibrahim et al.

2024). Because these patterns can be novel or emerge gradually, current autonomy evaluation suites, while informative, do not yet constitute a dedicated LOC framework (METR 2024). Ongoing method development and wider access to real-world testing data should help close these gaps over time.

In sum, the **evaluation of advanced AI is a nascent field**, with virtually no established standards, best practices or methodologies.

2.4.2. Non-Realised: Escalation

Figure 4 below illustrates how an unrealised scenario might unfold and highlights potential intervention points. It shows that the location of incident detection – whether at a frontier AI provider ("lab") or by the government – is a critical factor shaping the course of response efforts.

Figure 4: How a non-realised LOC scenario could progress

Non-Realised Emergency (Lab Detection)

- AI capability trigger
  - Internal mitigations successful
  - Internal mitigations unsuccessful
    - Model containment (pre-release)
      - Model release w/ controls
        - Mitigations successful
        - Monitoring and controls fail to mitigate
          - Escalate to realised emergency
      - Model theft
        - Escalate to realised emergency

Non-Realised Emergency (Governement Detection)

- Whistleblower notification / AISI eval / Intel/threat detection
  - Developer mitigations successful
  - Developer mitigations unsuccessful
    - Model not released or removed from market
      - Mitigations successful
    - Model released
      - Monitoring and controls fail to mitigate
        - Escalate to realised emergency
    - Model theft
      - Escalate to realised emergency

Legend:
- Negative outcome
- Positive outcome

AI developers have yet to establish clear thresholds for capabilities or risks that would pose a LOC concern. Organisations should establish clear criteria, including predefined triggers, for what constitutes a reportable event, such as a checklist requiring immediate notification to an incident manager or safety lead 'if X AI capability/ incident is observed'. The designated responders should convene immediately to determine the severity, likelihood and nature of the risk, and the necessary action to be taken, including activating pre-determined emergency protocols. **Clear decision making authority is crucial**; organisations must identify a designated person with the mandate to halt

operations if necessary. Furthermore, **if risks exceed a critical threshold, organisations must inform external stakeholders, such as governments or compute providers**. Government officials may then issue an advisory to other AI developers, similar to how a software vulnerability is entered in the Common Vulnerabilities and Exposures (CVE) catalogue, alerting other organisations of security threats (CVE 2025). However, in a real-time scenario, incident responders may need to make quick, high-stakes decisions about escalation without complete information, increasing the likelihood of misjudgements or insufficient measures.

Box 3: Emerging research on risk thresholds

> **Researchers have highlighted difficulties in setting risk thresholds – particularly for frontier models – due to limited data, rapidly developing capabilities and unclear threat models** (Koessler et al. 2024). Traditional safety-critical industries, such as civil aviation and nuclear power, often rely on revealed preferences, best practices and cost-benefit analyses to define risk tolerances. However, AI LOC risks are difficult to quantify as they lack historical precedent, involve complex interdependencies, and may result in large-scale, potentially catastrophic harms. **Emerging work has proposed developing concrete indicators to serve as actionable thresholds in a safety plan.**[8]

---

8   For example, 'If a model reaches 60% on the hypothetical "Cybench" assessment (the key risk indicator (KRI) threshold), then the company must meet a minimum "cybersecurity level 3" standard (the key control indicator (KCI) threshold) to keep the probability of incurring more than $500 million in economic damage below 1% per year' (Campos et al. 2025).

2.4.3. Non-Realised: Containment and Mitigation

In a non-realised case, containment focuses on converting a detected LOC risk into a controlled risk (see Figure 4 and Annex D). As model safety measures currently stand, developers may suspend further training or deployment of an AI model until safety issues are addressed (METR 2024). In an ideal scenario, companies would apply additional technical containment measures such as restricting the AI model's access to resources and networks and enhancing model weight security.

However, there are challenges to implementing an effective containment approach. **Some mitigations, such as high levels of security or improved alignment methods, may take months to years to implement,** highlighting the need for proactive response planning and safety measures.

2.4.4. Realised: LOC Event Occurs

In a realised scenario, LOC occurs despite prior intervention efforts (see Figure 5 and Annex E).

Figure 5 details a potential pathway in which a realised LOC scenario could process and potential intervention points. **This event marks the point at which a misaligned AI system operates beyond effective human oversight, with no clear means for the developer to restore control.** The severity of a realised LOC incident will depend on the level of safeguards overcome and the access or resources granted to a model. For example, robust monitoring and access can limit the ability of AI systems to operate in ways that render safeguards ineffective (Leike 2023).[9] The specific access and deployment context of the AI model are crucial to the potential impact and scope of a LOC event: if the AI system is integrated into systems that allow it to modify server environments, interface with decision making processes, or trigger actions in the physical world, the potential for unintended harm is significantly increased.

---

9   This is not to suggest that safeguards are unnecessary, as they may still directly reduce risks and severity.

Figure 5: How a realised LOC scenario could progress
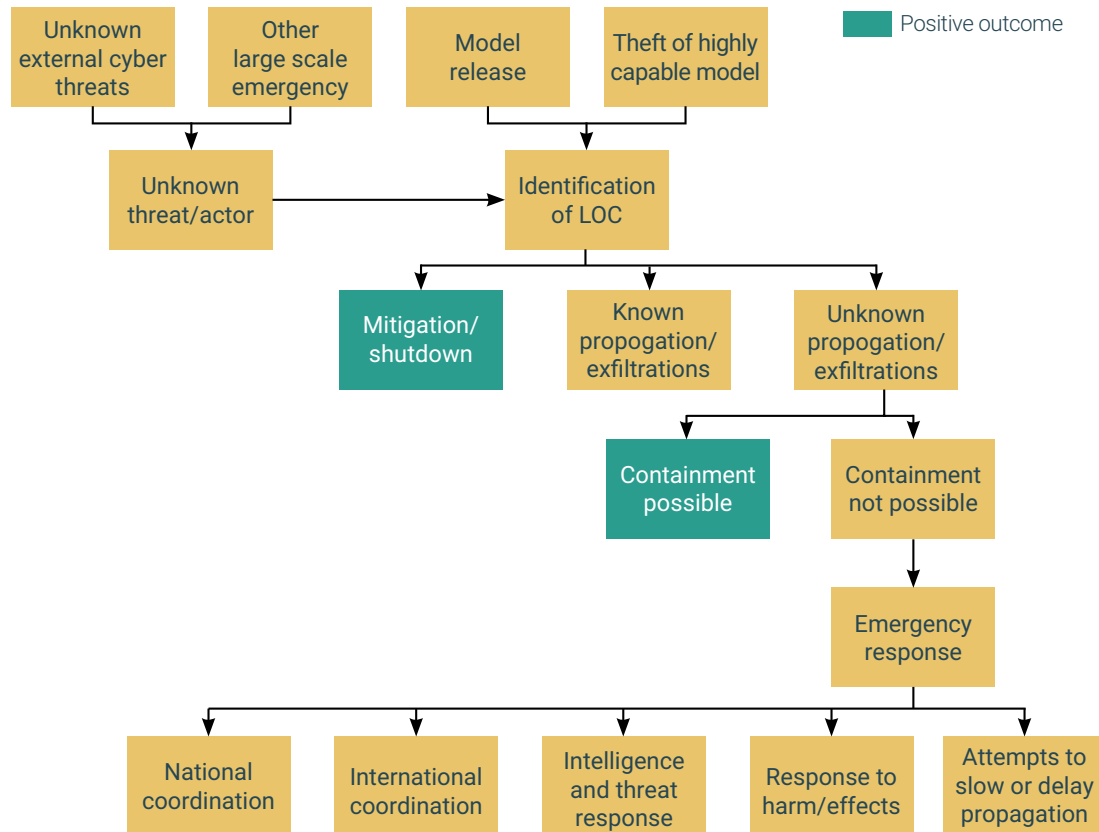
2.4.5. Realised: Early Warning and Detection

A realised LOC risk may be discovered in three ways: 1) the direct detection of capabilities; 2) monitoring of suspicious functions; or 3) when an AI system causes harm (see Annex E). In the first scenario, developers could directly detect evidence of either successful self-exfiltration or subverted internal deployment, which are two central pathways through which human oversight may be degraded. Some AI developers engage in anomaly monitoring, which might allow them to detect suspicious AI outputs. Second, LOC risk detection involves identifying system outputs associated with attempts to initiate

resource acquisition or capability expansion. Today, compute providers may play a role in monitoring potentially suspicious activity within a data centre, such as if a model is improperly acquiring compute resources. Finally, detection could happen when a deployed system operating outside intended constraints results in harm or performs actions misaligned with its original objectives. Additional actors, such as government agencies, may play a role in such detection.

In an ideal situation, developers or other stakeholders would detect a LOC scenario early. Effective early detection may require a combination of access, continuous monitoring

and the ability to synthesise information from multiple sources to identify potential LOC. AI models themselves could play a crucial role in assisting with detection, helping to analyse amounts of data and flagging signs of control-undermining capabilities.

However, the effectiveness of detection is highly context-dependent and remains uncertain. Detection could stem from internal triggers, such as the LOC event itself or evaluations identifying it, or external triggers, such as initiating operations that access physical-world resources or produce harmful effects. If a LOC event involves the theft of model weights, detection may need to occur in a completely different environment from the model's training or deployment infrastructure.

A key challenge is that LOC risks may be detected from a series of misalignment indicators rather than one clear-cut incident. Assessments may be subject to debate, either due to legitimate uncertainty or attempts to prioritise development over safety, potentially driven by competitive pressure. Early detection is also complicated by whether an AI system's outputs would initially remain low-profile, delaying the initiation of internal processes that operate without monitoring or attempt to access additional resources.[10] A system exhibiting adversarial characteristics may produce outputs that obscure its activity, such as altering logs or records in ways that hinder detection (Meinke et al. 2025).

2.4.6. Realised: Escalation

Escalation involves quickly scaling up the response beyond the initial responder. **Few**

AI developers have safety frameworks that clearly describe a chain of command and escalation pathways for LOC incidents, both internally and with external stakeholders. Ideally, organisations should internally escalate by involving top executives, potentially following a 'war room' strategy that brings together various stakeholders, from engineers to legal advisors. **Organisations should simultaneously initiate external escalation**. This could mirror cyber incident responses, with law enforcement probing legal violations and identifying potential malicious actors, technical teams containing the AI system, and intelligence personnel analysing the broader threat. If an AI incident threatens critical infrastructure or public health, national emergency mechanisms could be activated, similar to how they would be for a terrorist attack (US Government 2024). A LOC event that crosses borders should also prompt international coordination.

Frameworks and agreements to coordinate escalation should be established ahead of time, as coordinating during a crisis is more difficult than activating pre-existing channels. LOC escalation protocols should be exercised regularly, both to familiarise key stakeholders and to stress-test for vulnerabilities, mirroring how nations and militaries conduct wargames. More broadly, preparation is critical for establishing information flow between AI developers and governments, and between governments themselves (Vomberg 2013).

---

10   In latent harm situations, models may accumulate resources unbeknownst to human overseers.

Box 4: Key elements in escalation pathways

- **Incident Command Structure:** A clear chain of command framework ensures clear division of roles and responsibilities among AI developers and AISI responders.

- **Developer to Government Coordination**: Escalation protocols should specify when to engage law enforcement and government cyber response teams.

- **Government to Government Coordination**: Government authorities may invoke national emergency mechanisms if public safety is at risk. Cross-border incidents may require global cooperation, as with transnational cyber threats.

- **Information Sharing**: Clear information sharing between developers and government actors is essential to manage the emergency response.

International escalation may also encounter significant barriers. Uncertainty surrounding the evidence could impact the escalation process, as other countries could demand clearer evidence of a LOC scenario, but such evidence may not be readily available. Furthermore, the presence of international conflicts could complicate detection efforts or risk the LOC escalating into a broader confrontation between states (Mitre and Predd 2025).

2.4.7. Realised: Containment and Mitigation

Containment aims to stop the AI model's harmful actions, including propagation or self-exfiltration, while mitigation focuses on minimising harm, recovering and addressing underlying causes to prevent future incidents (Campos et al. 2025).

The harms posed by a LOC incident are uncertain and depend on the AI model capabilities, goals, acquired resources at the time of detection and the extent to which it is embedded in critical

infrastructure and physical devices.[11] Potential harms include significant financial losses, widespread and large-scale cybersecurity incidents, biological or nuclear incidents, and disruption to critical infrastructure and services.

A critical aspect of containment is the degree to which humans can exercise control to correct, override or impede the model's harmful outputs. In simple cases, developers could deploy an updated model or a filtering layer that overrides misaligned goals (Leong and Atherton 2023); however, in an extreme scenario, the AI model may not be fully controllable by the developer. Mitigations would aim to shut down and limit the propagation of uncontrolled AI copies, such as by disconnecting network access and hardware or power shut down (Moric et al. 2025). These measures would present novel challenges, and the AI system may attempt to resist such interventions, for example by creating backup copies on alternate servers (Center for AI Safety

---

11   An AI model with superhuman cyber capabilities could insert numerous vulnerabilities and backdoors into critical infrastructure that it can then use as leverage to pursue its own ends, such as self-preservation.

2025).[12] Emergency interventions may be required in some cases, with governments invoking special authorities to shut down servers, issuing mandatory orders for AI developers to cease operations, and coordinating emergency response efforts. Containment measures can benefit from compute governance technologies that allow for actions such as authorisation, verification and location tracking (Kulp et al. 2024).[13]

Containment and mitigation efforts may fail, for several reasons. First, there is no consensus on which strategies are most effective, as the unique characteristics of each incident introduce significant uncertainty. Second, mitigation may be too late to prevent harm. Third, model shutdown may be impracticable in cases where AI systems with sufficiently advanced cyber capabilities have access to hidden or difficult to locate compute resources.

Box 5: Containment and mitigation response examples

- **AI Model Access Restrictions**: Blacklist IP addresses, limit prompts and Application Programming Interface (API) calls for suspicious users, impose identity verifications and bot detection.

- **AI Model Use Restrictions**: Throttle prompts and API calls, disable tool use and code execution, reduce context windows, output filtering.

- **Environment Controls**: Sandbox (isolating AI into restricted environments), honeypots (diverting AI into decoy environments) and software containers.

- **Hardware Restrictions**: Restrict graphics processing units (GPUs), isolate or remove compromised hardware.

- **AI Model Shutdown**: Power off/data centre shutdown, or removal/decommissioning

---

12   Cybersecurity techniques are instructive. For instance, if malware is spreading, network administrators will isolate infected machines and block the malware's communication channels (Souppaya & Scarfone 2013). Similarly, network segmentation and blocking known command channels, such as specific APIs or message queues, can limit an AI model's ability to affect additional models.

13   Containment and mitigation are iterative processes. As in biosafety and cybersecurity contexts, a contained LOC event can serve as a case study to tighten oversight; the 2001 Anthrax attacks led to stricter biosecurity measures, and large cyberattacks led to the passing of CIRCIA (Chesbrough 2024; Gostin & Nuzzo 2021).

2.5. Open-Source Models

A key additional challenge are open-source and open-weight models. These models can increase LOC risks by enabling widespread, unmonitored model access and modification.[14] Such models allow users to host models on their own servers without oversight, expanding the attack surface and increasing the number of potential entry points for LOC incidents.[15] A widening gap in capabilities between closed-source and open-source models may make scenarios around model theft more likely. For AISIs, the proliferation of open-weight models may call for the increased monitoring of risks and the regulation of critical nodes such as compute resources (Heim 2023). Mitigation efforts could also include strengthening the resilience of infrastructure against AI-driven hacking and related threats, including through advances in defensive cyber capabilities (Motlagh et al. 2024; Shombot et al. 2024). **Policymakers should track the development of open-source models**, as the appropriate response will depend on how their capabilities evolve relative to closed-source models.

---

14   Open source refers to all aspects of an AI model, including model weights but also training methods, data and other components, allowing others to replicate the entire development process (White et al. 2024). Open weight refers to publicly accessible parameters that determine outputs based on inputs (Nobel et al. 2024).

15   Researchers have shown how it might be possible to undo AI model safety finetuning from existing open-weight models (Gade et al. 2024). A diverse range of actors might have the ability to modify future, more advanced open-weight models to potentially deploy models with dangerous capabilities at scale (Cable & Black 2024).

Chapter 3. **Recommendations and Conclusion**

Drawing from lessons in analogous risk management frameworks (Chapter 2) and example pathways for realised and non-realised incidents (Chapter 3), this chapter suggests recommendations to enhance detection, escalation, containment and mitigation of LOC incidents. Table 1 below summarises the key recommendations for stakeholders ranging from detection to escalation to containment and mitigations for various stakeholders (e.g. AI developers, governments, etc), which are then discussed in further detail in the following sections.

Table 2: Summary of recommendations

Stage	Stakeholder	Recommendation
Detection	AI Developers	• Monitor critical capability levels • Identify early warning signs and emergent capabilities • Establish standardised benchmarks and reporting
	Compute Providers	• Implement compute monitoring and anomaly detection • Enhance hardware and supply chain oversight
	National Government: AISI	• Lead efforts to establish shared criteria for AI LOC • Coordinate evaluations and safety testing • Monitor advanced capabilities and emergent capabilities
	National Government: Other Agencies	• Assess and monitor AI-related cyber incidents • Receive, analyse, and disseminate threat intelligence
	Third Party Researchers	• Conduct evaluations, red-teaming and adversarial testing • Collaborate on standardised benchmarks and techniques
Escalation	AI Developers	• Establish incident response protocols with defined escalation thresholds and organisational structures • Respond and verify potential threshold breaches • Conduct regular training and scenario drills

	Compute Providers	• Notify AI developers and relevant authorities • Coordinate with developers and national authorities
	National Government: AISI	• Establish disclosure and communication channels with AI developers and compute providers • Receive and assess escalation notifications • Provide oversight for threshold verification and escalation
	National Government: Other Agencies	• Provide forensic and technical expertise • Investigate and verify incidents and reports • Share intelligence with relevant national security stakeholders • Exercise enforcement and investigative authority
	Third Party Researchers	• Verify and disclose findings through established channels • Publicise risks where appropriate for broader awareness
Containment and Mitigation	AI Developers	• Implement model access and use limits • Develop and test model shutdown measures • Advance research on containment and layered defences
	Compute Providers	• Enforce model access and usage restrictions • Shut down or limit hardware resources during incidents • Review incident and shutdown procedures
	National Government: AISI	• Coordinate with AI developers on containment and mitigation response measures • Develop security measures for model deployments • Enforce model access, use and environmental controls
	National Government: Other Agencies	• Coordinate cyber incident response protocols • Coordinate responses with critical infrastructure providers
	Third Party Researchers	• Provide technical assistance during mitigation • Update auditing, evaluation procedures and continuous red-teaming exercises

3.1. Detection

Detection – the identification of potential LOC threats or misuse of AI models – could be improved by 1) creating a shared definition of LOC; 2) refining standardised detection benchmarks; and 3) enhancing stakeholder collaboration.

Governments, with AI developers and other stakeholders, should establish a clear, shared definition of AI LOC and a set of criteria for detection. AI models can exhibit emergent capabilities and follow unpredictable trajectories, making it difficult to define LOC uniformly across deployment conditions. A task force or working group led by AISIs, in collaboration with AI developers and researchers, could seek to create a comprehensive but flexible definition of LOC. Agreement on early warning signs that may signal a LOC incident would help determine proportional responses to risks (Popoola et al. 2013). Developers and government stakeholders should consider adopting practices from cybersecurity and biosecurity domains by integrating confidence scoring systems and continuous, overlapping detection mechanisms (CISA 2025a; Yousef et al. 2024; Thompson et al. 2019).

AI developers and researchers should refine detection by developing standardised benchmarks and improving their reliability and validity. Developers should enhance detection of control-undermining capabilities. Techniques that monitor AI model internals in addition to outputs have shown promise in detecting deception (Goldowsky-Dill et al. 2025). Developers and researchers should continue improving adversarial techniques, sharing results and developing standardised benchmarks to assess autonomy and other capabilities (Barnett & Thiergart 2024; Greenblatt, Shlegeris et al. 2024). Early detection could also be improved by robust real-time monitoring tools that log outputs, decisions and compute usage to detect potential anomalies (Kaur et al. 2023; Greenblatt, Shlegeris et al. 2024).[16]

Governments should enhance awareness and information sharing between all stakeholders, including the tracking of compute resources. Compute providers, national security agencies, and cybersecurity professionals could be trained to recognise LOC indicators and monitor developments in AI capabilities. Cloud providers could incorporate real-time compute monitoring and verification to flag high-risk users.[17] Enhancing the information flow between AI developers, compute providers and governments on AI R&D would also improve detection. Governments should consider requiring developers to track and report key metrics, such as compute usage for AI R&D, as well as to disclose extreme capabilities to AISIs (Mikton 2024).

3.2. Escalation

Escalation involves actions following the detection of a potential LOC event. These measures may include activating predefined protocols, notifying key stakeholders, and mobilizing or coordinating resources to address potential threats.

AI developers should establish well-defined escalation protocols and conduct

---

16   As with cybersecurity, anomaly detection and monitoring tools would require cost-benefit analysis and proper calibration to reduce false positives and false negative. Some initial work has been done on this with trusted AIs monitoring untrusted AIs to detect backdoored code (Greenblatt, Shlegeris et al. 2024).

17   Other measures could include combining anomaly detection, chip-level telemetry to detect unauthorised workloads and stronger supply chain oversight (Heim 2024; Kulp et al. 2024).

regular training exercises to ensure their effectiveness. Developers should create incident response plans in advance, with well-defined, evidence-based thresholds for when to trigger an emergency response. Incident plans should assign critical roles, including an 'incident commander' who has decision making authority, direct access to leadership and the authority to coordinate cross-functional teams and suspend models. Incident protocols should be customisable to accommodate variations, and organisations should drill escalation pathways (Webb & Chevreau 2006).

Box 6: Minimum viable incident response plan

- **Defined Thresholds**: Specify capability thresholds or scenarios (e.g. unexpected emergent capability or abnormal performance) that activate the incident response.

- **Verification of Threshold Crossings**: Use logs, audits, evaluations or third-party reviews to confirm when a threshold has been met.

- **Clear Roles and Responsibilities**: A designated individual (e.g. an 'incident commander') should have the authority to:

    - Assess AI model controllability and capabilities with direct communication lines to leadership and board members.

    - Assemble cross-departmental (technical, legal and communication) teams to expedite decision making and incident response.

    - Implement safety measures, including suspending or throttling AI deployments, and implementing lockdowns on critical systems and data.

    - Initiate external reporting to regulators or other oversight bodies.

Communication Plan: Establish internal and external communication protocols for alerting leadership and relevant authorities.

Training and Drills: Conduct tabletop exercises and simulations to test readiness, clarify roles and practice real-time communication under stress.

Post-Incident Review: Document root causes and lessons learned and recommend improvements to refine future responses.

Government stakeholders should consider mandatory reporting mechanisms for AI risks and potential incidents. Governments should consider mandating legal disclosure requirements covering key risk scenarios, including model theft (e.g. stolen weights, unauthorised access), deceptive model behaviour (e.g. models manipulating evaluations to appear weaker) and emergent risky capabilities (e.g. escape or uncontrolled replication of models, and extreme capability breakthroughs). Government actors should clarify how cyber incident reporting mechanisms can be applied to AI-related incidents. Independent safety evaluators, third-party auditors and compute providers could have the authority to report high-risk developments to oversight bodies.

Government stakeholders should establish disclosure channels and whistleblower safeguards for employees of AI developers. Employees across key sectors – including AI developers, compute providers and third-party evaluators – should have designated channels to report safety concerns, as well as protections against potential retaliation or suppression, including the use of non-disclosure agreements (NDAs), non-disparagement clauses or other measures to suppress disclosure of safety risks (Lovely 2024). In addition, governments should create secure disclosure channels that are independent from corporate protocols and make these available to employees at developer firms.

AI developers, AISIs and relevant government departments should enhance cross-sector and international coordination, including clear communication lines, information-sharing agreements and predefined escalation pathways (see Annex A). These could include secure emergency hotlines between AI developers and national AI safety regulators, classified communication channels, and sector-specific CERTs for AI incidents. AISIs could act as central and secure information hubs, consolidate national data and facilitate trusted exchanges with international counterparts. Global emergency response exercises, potentially through multilateral forums, could improve preparedness and refine coordination protocols. International agreements could provide additional mechanisms for addressing AI risks that affect health and cyber domains (WHO 2025).

3.3. Containment and Mitigation

Containment aims to halt the model's harmful or unexpected actions and limit risks or prevent further damage. Mitigation refers to actions taken to stop a threat or incident from occurring in the first place. Unlike other safety-critical technologies, AI has no established global standards for safety and security, making proactive intervention essential.

AI developers should prepare containment measures that are rapid and flexible. In the event of a LOC event, safety should take priority over service continuity. Effective containment may require immediate measures, including model shutdown, that make it harder for a rogue AI to take actions. Plans should be adaptable, as specific response depends on the context where the model is deployed, its level of access and its capabilities. Communication lines to external infrastructure and service providers who could take action to limit proliferation should be established.

AI developers and other stakeholders should further explore and advance research on containment methods. Existing research shows that current containment efforts face limitations, especially for self-replicating AI (Clymer, Wijk & Barnes 2024; Salib 2025; Pan et al. 2024). Investments should be made in containment technologies to shut off models, restrict capabilities, limit harm or unintended

actions, and ensure retention of human control. This may also include research using AI models for containment and exploring techniques such as sandboxing, model distillation and layered defence strategies.

AI developers, external researchers and AISIs should prioritise safety and alignment measures, including by building validated safety cases. If risk assessment and prevention are not prioritised, competitive pressures and geopolitical rivalries may push developers to continue deploying potentially misaligned AI models (Mitre & Predd 2025). Developers should collaborate with external researchers, evaluators and AISIs to build on emerging safety case projects for AI development, training and deployment (Irving 2024). Future efforts around safety cases may include independent verification of model characteristics and alignment, evaluation of unintended capabilities, and assessment of worst-case failure modes.

Government stakeholders should seek to strengthen AI security to protect model weights and algorithmic techniques. Governments could require or incentivise AI developers that exceed specified capability thresholds to implement stricter security protections – of both model weights and algorithmic insights – to prevent the theft of dangerous capabilities by malicious actors and the diffusion of models to unmonitored environments.[18] Security improvements could include measures such as hardened bandwidth limitations, automated network monitoring and encryption. Routine hardware supply chain and data centre inspections would also help to detect unauthorised access.[19] Beyond external threats, organisations must also consider

security risks from the AI models themselves and thus employ regular memory wiping, adversarial testing and monitoring.

Governments and developers should improve safety governance by fostering robust safety cultures and adopting secure-by-design principles. AI developers should evaluate failure modes and implement safeguards before deployment, with independent third-party audits verifying compliance with existing standards, as commonplace in other fields such as nuclear energy, aviation, finance and banking, pharmaceuticals, and more. AI developers should also continue to allocate compute resources to AI safety, including research on monitoring, alignment and safeguards. Governments could make secure-by-design guidelines mandatory to ensure that safety features are built into AI models from the outset (NCSC 2023). They could also consider requirements or incentives for safety research.

3.4. Conclusion

Preventing AI LOC demands a proactive, multi-layered strategy. To guide efforts across detection, escalation, containment and mitigation, this report offers the following core principles: 1) focus on prevention; 2) enhance information sharing; and 3) foster a safety-first culture.

Preventing LOC is far easier than recovering from it. The cost of inaction could far outweigh that of early policy measures – AI developers, governments and stakeholders must urgently invest in large-scale preparedness. This could include establishing pre-defined frameworks

---

18   Security clearances for researchers, increased physical security and compartmentalisation of sensitive projects could further mitigate insider threats.

19   Air-gapped network infrastructure and strict execution controls could be implemented for highly capable AI models (Nevo et al. 2024).

for detection and escalation and prioritising preventative measures such as secure-by-design principles.

Effective information sharing between industry, government and international partners is vital. Within a nation, this means policy frameworks that enforce reporting mechanisms and transparent communication channels between private companies and government bodies. Global coordination is also essential, as LOC is a transnational risk. Organisations such as

AISIs play a key role in facilitating information exchange and promoting collaboration.

Finally, as AI capabilities continue to evolve rapidly, fostering a safety-first culture is essential to reducing the likelihood of a LOC incident. Governments should play a key role by incentivising design practices that reduce risks and by promoting a culture of transparency and accountability across the industry. However, LOC remains significantly understudied overall, and further research is necessary.

References

Admass, Wasyihun, Yirga Yayeh Munaye & Abebe Abeshu Diro. 2024. 'Cyber security: State of the art, challenges and future directions.' *Cyber Security and Applications* 2. As of 9 April 2025: https://doi.org/10.1016/j.csa.2023.100031

Anderson, Iain. 2008. *Foot and Mouth Disease 2007: A Review and Lessons Learned.* UK Government. https://assets.publishing.service.gov.uk/media/5a7c7289e5274a5255bceb57/0312.pdf

Anderson, Ross & Tyler Moore. 2006. 'The economics of information security.' *Science* 314. http://dx.doi.org/10.1126/science.1130992

APHL. 2016. *Clinical Laboratory Preparedness and Response Guide.* Association of Public Health Laboratories. www.aphl.org/aboutAPHL/publications/documents/WORK_BlueBook.pdf

Australian Signals Directorate. 2025. 'Australian Signals Directorate's Cyber Security Partnership Program.' As of 10 April 2025: https://www.cyber.gov.au/partnershipprogram

Banovic, Nikola, Zhuoran Yang, Aditya Ramesh & Alice Liu. 2023. 'Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust.' *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1), Article 27. As of 8 April 2025: https://dl.acm.org/doi/10.1145/3579460

Barnett, Peter & Lisa Thiergart. 2025. 'What AI evaluations for preventing catastrophic risks can and cannot do.' *ArXiv.* As of 8 April: https://arxiv.org/html/2412.08653v1

BBC. 2017. 'Global ransomware attack causes turmoil.' Technology news, 28 June. As of 9 April 2025: https://www.bbc.com/news/technology-40416611

Benderly, Beryl Lieff. 2018. 'A decade after a fatal lab safety disaster, what have we learned?' *Science*, 5 December. As of 9 April 2025: https://www.science.org/content/article/decade-after-fatal-lab-safety-disaster-what-have-we-learned

Bengio, Yoshua. 2023. 'How Rogue AIs may Arise'. As of 5 April 2025: https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Gunes Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner & Soren Mindermann. 2024. 'Managing extreme AI risks amid rapid progress.' *Science* (384)6698. As of 4 April 2025: https://doi.org/10.1126/science.adn0117

Bengio Yoshua, et al. 2025. *International AI Safety Report.* Department for Science, Innovation & Technology. As of 5 April 2025: https://www.gov.uk/government/publications/international-ai-safety-report-2025

Bonta, Rob. 2025. 'Data Security Breach Reporting.' State of California Department of Justice. As of 9 April 2025: https://oag.ca.gov/privacy/databreach/reporting

Cable, Jack & Aeva Black. 2024. 'With Open Source Artificial Intelligence, Don't Forget the Lessons of Open Source Software.' Cybersecurity & Infrastructure Security Agency (CISA) blog, 29 July. As of 8 April 2025: https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software

Campos Siméon, Henry Papadatos, Fabien Roger, Chloé Touzet, Otter Quarks & Malcolm Murray. 2025. 'A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2502.06656

Callihan, Donald R. Marian Downing, Esmeralda Meyer, Luis Alberto Ochoa, Brian Petuch, Paul Tranchell & David White. 2021. 'Considerations for laboratory biosafety and biosecurity during the coronavirus disease 2019 pandemic: Applying the ISO 35001:2019 standard and high-reliability organizations principles.' *Applied Biosafety: Journal of the American Biological Safety Association* 26(3): 113-122. As of 9 April 2025: https://pubmed.ncbi.nlm.nih.gov/36035545/

Carlsmith, Joe. 2023. 'Scheming AIs: Will AIs fake alignment during training in order to get power?' *ArXiv*. As of 8 April 2025: https://arxiv.org/abs/2311.08379

CDC. 2021. 'First Report of AIDS.' *MMWR Morbidity and Mortality Weekly Report* 50(21). Centers for Disease Control and Prevention. As of 9 April 2025: https://www.cdc.gov/mmwr/pdf/wk/mm5021.pdf

CDC. 2023. 'Achieving Improved Emergency Response Around the World.' Centers for Disease Control and Prevention. As of 9 April 2025: https://www.cdc.gov/global-health-protection/php/stories-from-the-field/working-together-to-achieve-improved-emergency-response-around-the-world.html

CDC. 2024. 'About the Laboratory Response Network.' Centers for Disease Control and Prevention. As of 9 April 2025: https://www.cdc.gov/laboratory-response-network/php/about/index.html

CDC & NIH. 2020. *Biosafety in Microbiological and Biomedical Laboratories: 6th Edition.* Centers for Disease Control and Prevention (CDC) and National Institutes of Health (NIH). As of 9 April 2025: https://www.cdc.gov/labs/pdf/SF__19_308133-A_BMBL6_00-BOOK-WEB-final-3.pdf

Center for AI Safety. 2025. 'Risks from AI: An Overview of Catastrophic AI Risks.' As of 8 April 2025: https://www.safe.ai/ai-risk

CIDRAP. 2003. 'Taiwanese SARS researcher infected.' Center for Infectious Disease Research and Policy. As of 9 April 2025: https://www.cidrap.umn.edu/sars/taiwanese-sars-researcher-infected

Chesbrough, Ann. 2024. 'What's Behind CISA's Push for Private Sector Collaboration on CIRCIA Reporting Rules?' Breachlock, 5 June. As of 8 April 2025: https://www.breachlock.com/resources/blog/whats-behind-cisas-push-for-private-sector-collaboration-on-circia-reporting-rules/

Clymer, Josh, Hjalmar Wijk & Beth Barnes. 2024. 'The Rogue Replication Threat Model.' METR (Model Evaluation & Threat Research) blog, 12 November. As of 8 April 2025: https://metr.org/blog/2024-11-12-rogue-replication-threat-model/

Committee on Oversight and Government Reform, 2016. *The OPM Data Breach: How the Government Jeopardized Our National Security for More than a Generation.* As of 9 April 2025: https://oversight.house.gov/report/opm-data-breach-government-jeopardized-national-security-generation/

CVE. 2025. 'Frequently Asked Questions (FAQs).' Common Vulnerabilities and Exposures. As of 6 April 2025: https://www.cve.org/ResourcesSupport/FAQs

CISA. 2024. *The National Cyber Incident Response Plan (NCIRP).* Cybersecurity and Infrastructure Security Agency. As of 8 April 2025: https://www.cisa.gov/national-cyber-incident-response-plan-ncirp

CISA. 2025a. 'CISA Artificial Intelligence Use Cases.' Cybersecurity and Infrastructure Security Agency. As of 8 April: https://www.cisa.gov/ai/cisa-use-cases

CISA. 2025b. 'Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA).' Cybersecurity and Infrastructure Security Agency. As of 8 April: https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia

CISA. 2025c. 'Automated Indicator Sharing (AIS).' Cybersecurity and Infrastructure Security Agency. As of 8 April 2025: https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/automated-indicator-sharing-ais

Dafoe, Allan, Anca Dragan, Four Flynn, Helen King, Tom Lue, Lewis Ho & Rohin Shah. 2025. 'Updating the Frontier Safety Framework. Google DeepMind blog, 4 February. As of 8 April 2025: https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/

Dragan, Anca, Helen King & Allan Dafoe. 2024. 'Introducing the Frontier Safety Framework.' Google DeepMind blog, 17 May. As of 7 April 2025: https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/

EIOPA. 2025. 'Digital Operational Resilience Act (DORA).' European Insurance and Occupational Pensions Authority. As of 8 April 2025: https://www.eiopa.europa.eu/digital-operational-resilience-act-dora_en

ENISA. 2024. 'Navigating cybersecurity investments in the time of NIS 2.' Press release, 22 November. European Union Agency for Cybersecurity. As of 9 April 2025: https://www.enisa.europa.eu/news/navigating-cybersecurity-investments-in-the-time-of-nis-2

Enserink, Martin. 2007. 'Reports blame lab for foot-and-mouth fiasco.' *Science,* 10 September. As of 9 April 2025: https://www.science.org/content/article/reports-blame-lab-foot-and-mouth-fiasco

European Commission. 2024a. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024: Laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* As of 8 April 2025: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

European Commission. 2024b. 'Launching GLOWACON: A global initiative for wastewater surveillance for public health.' News announcement, 21 March. As of 8 April 2025: https://health.ec.europa.eu/latest-updates/launching-glowacon-global-initiative-wastewater-surveillance-public-health-2024-03-21_en

European Commission. 2025a. 'NIS2 Directive: New rules on cybersecurity of network and information models.' As of 8 April 2025: https://digital-strategy.ec.europa.eu/en/policies/nis2-directive

European Commission. 2025b. 'Cyber Resilience Act.' As of 8 April 2025: https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

European Commission. 2025c. 'Tracking diseases from the sewer.' News announcement, 29 January. As of 8 April 2025: https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/tracking-diseases-sewer-2025-01-29_en

European Council. 2024. 'Cybersecurity package: Council adopts new laws to strengthen cybersecurity capacities in the EU.' Press release, 2 December. As of 9 April 2025: https://www.consilium.europa.eu/en/press/press-releases/2024/12/02/cybersecurity-package-council-adopts-new-laws-to-strengthen-cybersecurity-capacities-in-the-eu/

EU-OSHR. 2021. 'Directive 2000/54/EC – biological agents at work.' European Agency for Safety and Health at Work. As of 9 April 2025: https://osha.europa.eu/en/legislation/directives/exposure-to-biological-agents/77

Federal Register. 2023. 'Safe, Secure & Trustworthy Development and Use of Artificial Intelligence.' Executive Order 14110. As of 10 April 2025: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

FIRST (homepage). 2025. Forum of Incident Response and Security Teams. As of 10 April 2025: https://www.first.org/

Frontier Model Forum (homepage). 2025. As of 5 April 2025: https://www.frontiermodelforum.org/

FS-ISAC. 2025. 'Join our Community: Become an FS-ISAC Member.' As of 9 April 2025: https://www.fsisac.com/membership

FTC-OIG. 2025. 'Whistleblower Protection.' Federal Trade Commission Office of Inspector General. As of 9 April 2025: https://oig.ftc.gov/whistleblower-protection

Gade, Pranav, Charlie Rogers-Smith, Simon Lermen & Jeffrey Ladish. 2024. 'BadLlama: Cheaply removing safety fine-tuning from Llama 2-Chat 13B.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2311.00117

Ghosh, Pallab. 2011. '"Safety incidents" at animal lab.' *BBC News,* 26 May. As of 9 April 2025: https://www.bbc.com/news/science-environment-13566593

Goldowsky-Dill, Nicholas, Bilal Chughtai, Stefan Heimersheim & Marius Hobbhahn. 2025. 'Detecting Strategic Deception Using Linear Probes.' *ArXiv*. As of 8 April: https://arxiv.org/pdf/2502.03407

Goodell, John & Corbet, Shaen. 2025. 'Commodity market exposure to energy-firm distress: Evidence from the Colonial Pipeline ransomware attack.' *Finance Research Letters* 51. As of 9 April 2025: https://doi.org/10.1016/j.frl.2022.103329

Gostin, Lawrence & Jennifer B. Nuzzo. 2021. 'Twenty Years After the Anthrax Terrorist Attacks of 2001: Lessons Learned and Unlearned for the COVID-19 Response.' *JAMA Network* 326(20) As of 8 April 2025: https://jamanetwork.com/journals/jama/fullarticle/2785780

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman & Evan Hubinger. 2024. 'Alignment faking in large language models.' *ArXiv*. As of 8 April: https://arxiv.org/abs/2412.14093

Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan & Fabien Roger. 2024. 'AI Control: Improving Safety Despite Intentional Subversion.' *ArXiv*. As of 8 April: https://arxiv.org/pdf/2312.06942

Health and Safety Executive. 2014. *Regulation of Health and Safety at Work.* As of 9 April 2025: https://www.hse.gov.uk/pubns/hse51.htm

Health and Safety Executive. 2025a. 'The regulation of specified animal pathogens.' As of 9 April 2025: https://www.hse.gov.uk/biosafety/sapo.htm

Health and Safety Executive. 2025b. 'Control of substances hazardous to health (COSHH).' As of 9 April 2025: https://www.hse.gov.uk/cleaning/topics/coshh.htm

Heath, Timothy R. & Matthew Lane. 2019. *Science-Based Scenario Design: A Proposed Method to Support Political-Strategic Analysis.* Santa Monica, Calif.: RAND Corporation. As of 9 April 2025: https://www.rand.org/pubs/research_reports/RR2833.html

Heim, Lennart. 2023. 'Compute and the Governance of AI – Talk.' As of 8 April 2025: https://blog.heim.xyz/compute-and-the-governance-of-ai-talk/

Heim, Lennart. 2024. 'Crucial Considerations for Compute Governance.' As of 8 April 2025: https://blog.heim.xyz/crucial-considerations-for-compute-governance/

Hodgson, Quentin E., Aaron Clark-Ginsberg, Zachary Haldeman &rew Lauland & Ian Mitch, *Managing Response to Significant Cyber Incidents.* Santa Monica, Calif.: RAND Corporation, RR-A1265-4, 2022. As of 9 April 2025: https://www.rand.org/pubs/research_reports/RRA1265-4.html

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer & Ethan Perez. 2024. 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training.' *ArXiv*. As of 8 April 2025: https://arxiv.org/abs/2401.05566

Ibrahim, Lujain, Saffron Huang, Lama Ahmad & Markus Anderljung. 2024. 'Beyond Static AI Evaluations: Advancing Human Interaction Evaluations for LLM Harms and Risks.' *ArXiv.* As of 7 April 2025:
https://arxiv.org/html/2405.10632v5

ICO. 2025a. 'NIS and the UK GDPR.' Information Commissioner's Office. As of 9 April 2025:
https://ico.org.uk/for-organisations/the-guide-to-nis/nis-and-the-uk-gdpr/#GDPR-3

ICO. 2025b. 'The role of the National Cyber Security Centre (NCSC).' Information Commissioner's Office. As of 9 April 2025:
https://ico.org.uk/for-organisations/the-guide-to-nis/the-role-of-the-national-cyber-security-centre-ncsc/

Intersoft Consulting. 2025. 'Art. 33 GDPR Notification of a personal data breach to the supervisory authority.' As of 9 April 2025:
https://gdpr-info.eu/art-33-gdpr/

Irving, Geoffrey. 2024. 'Safety cases at AISI.' UK AI Security Institute. As of 5 April 2025:
https://www.aisi.gov.uk/work/safety-cases-at-aisi

ISO. 2019. *ISO 35001: 2019 – Biorisk Management for Laboratories and other Related Organisations.* International Organization for Standardization. As of 9 April 2025:
https://www.iso.org/standard/71293.html

ITU. 2025. 'National CIRT.' International Telecommunications Union. As of 9 April 2025:
https://www.itu.int/en/ITU-D/Cybersecurity/Pages/national-CIRT.aspx

Kaur, Ramanpreet, Dusan Gabrijelcic & Tomaz Klobucar. 2023. 'Artificial intelligence for cybersecurity: Literature review and future research directions.' *Information Fusion* 97. As of 8 April 2025:
https://doi.org/10.1016/j.inffus.2023.101804

Koessler, Leonie, Jonas Schuett & Markus Anderljung, 'Risk thresholds for frontier AI.' *ArXiv.* As of 8 April 2025:
https://arxiv.org/pdf/2406.14713

Krakovna, Victoria & Janos Kramar. 2023. 'Power-seeking can be probable and predictive for trained agents.' *ArXiv.* As of 8 April 2025:
https://arxiv.org/pdf/2304.06528

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer & Zev Winkelman. 2024. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090.'* Santa Monica, Calif.: RAND Corporation. As of 8 April 2025:
https://www.rand.org/pubs/working_papers/WRA3056-1.html

Langosco, Lauro, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau & David Krueger. 2023. 'Goal Misgeneralization in Deep Reinforcement Learning.' *ArXiv.* As of 8 April 2025: https://arxiv.org/abs/2105.14111

Leike, Jan. 2023. 'Musings on the alignment Problem: Self-exfiltration is a key dangerous capability.' Substack article, 13 September. As of 8 April 2025:
https://aligned.substack.com/p/self-exfiltration

Leong, Brenda & Daniel Atherton. 2023. 'AI incident response plans: Not just for security anymore.' *IAPP News,* 20 September. As of 8 April 2025:
https://iapp.org/news/a/ai-incident-response-plans-not-just-for-security-anymore

Lovely, Garrison. 2024. 'Laws Need to Catch Up to Artificial Intelligence's Unique Risks.' *New York Times,* 29 September. As of 8 April 2025:
https://www.nytimes.com/2024/09/29/opinion/ai-risks-safety-whistleblower.html

Malave, Adriel & Elamin M. Elamin. 2010. 'Severe Acute Respiratory Syndrome (SARS): Lessons for future pandemics.' *AMA Journal of Ethics* 12(9). As of 9 April 2025: https://journalofethics.ama-assn.org/article/severe-acute-respiratory-syndrome-sars-lessons-future-pandemics/2010-09

Meinke, Alexander, Bronson Schoen, Jeremy Scheurer, Mikita Balesni, Rusheb Shah & Marius Hobbhahn. 2025. 'Frontier Models are Capable of In-context Scheming.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2412.04984

METR. 2023. 'Responsible Scaling Policies (RSPs).' METR (Model Evaluation & Threat Research) blog, 26 September. As of 8 April 2025: https://metr.org/blog/2023-09-26-rsp/

METR. 2024. 'Autonomy Evaluation Resources.' METR (Model Evaluation & Threat Research) blog, 15 March. As of 8 April 2025: https://metr.org/blog/2024-03-13-autonomy-evaluation-resources/

METR. 2025. 'Updates: AI models can be dangerous before public deployment.' METR (Model Evaluation & Threat Research) blog, 17 January. As of 8 April 2025: https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/

Mikton, Sofia. 2024. 'Briefing: AI Governance and The Role of Compute Providers.' Simon Institute blog, 19 June. As of 8 April 2025: https://www.simoninstitute.ch/blog/post/briefing-ai-governance-and-the-role-of-compute-providers/

MITRE. n.d. *Structured Threat Information eXpression — STIX: A Structured Language for Cyber Threat Intelligence Information.* As of 9 April 2025: https://makingsecuritymeasurable.mitre.org/docs/stix-intro-handout.pdf

Mitre, Jim & Joel B. Predd. 2025. *Artificial General Intelligence's Five Hard National Security Problems.* Santa Monica, Calif.: RAND Corporation. As of 8 April 2025: https://www.rand.org/pubs/perspectives/PEA3691-4.html

Moric, Zlatan, Vedran Dakic & Damir Regvart. 2025. 'Advancing Cybersecurity with Honeypots and Deception Strategies.' *Informatics* 12(1) 2025. As of 8 April 2025: https://www.mdpi.com/2227-9709/12/1/14

Motlagh, Farzad Nourmohammadzadeh, Mehryar Majd, Feng Cheng, Mehrdad Hajizadeh, Pejman Najafi & Christoph Meinel. 2024. 'Large Language Models in Cybersecurity: State-of-the-Art.' *ArXiv*. As of 8 April 2025: https://arxiv.org/html/2402.00891v1

Motwani, Sumeet Ramesh, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond & Christian Schroeder de Witt. 2024. 'Secret Collusion among AI Agents: Multi-Agent Deception via Steganography.' *ArXiv*. As of 8 April 2025: https://arxiv.org/html/2402.07510v3

National Research Council. 2001. *Under the Weather: Climate, Ecomodels & Infectious Disease & Human Health.* National Academies Press.

NCSC. 2023. *Guidelines for Secure AI Model Development*. UK National Cyber Security Centre. As of 8 April 2025: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley & Jeff Alstott. 2024. *A Playbook for Securing AI Model Weights.* Santa Monica, Calif.: RAND Corporation. RBA2849-1. As of 8 April 2025: https://www.rand.org/pubs/research_briefs/RBA2849-1.html

NIH. 2024. *NIH Guidelines for Research Involving Recombinant or Synthetic Nucleic Acid Molecules (NIH Guidelines).* National Institute of Health. As of 9 April 2025: https://osp.od.nih.gov/wp-content/uploads/ NIH_Guidelines.pdf

NIST. 2025. 'The CSF 1.1 Five Functions.' National Institute of Standards and Technology. As of 9 April 2025: https://www.nist.gov/cyberframework/ getting-started/online-learning/five-functions

NIST & AISI. 2024. *US AISI and UK AISI Joint Pre-Deployment Test, OpenAI o1.* National Institute of Standards and Technology (NIST) and the UK's AI Safety Institute (AISI). As of 5 April 2025:
https://cdn.prod.website-files. com/663bd486c5e4c81588db7a1d/6763 fac97cd22a9484ac3c37_o1_uk_us_december_ publication_final.pdf

Nobel, Parth, Alan Z. Rozenshtein & Chinmayi Sharma. 2024. 'Open-Access AI: Lessons From Open-Source Software.' *Lawfare*, 25 October. As of 8 April 2025:
https://www.lawfaremedia.org/article/open-access-ai--lessons-from-open-source-software

OpenAI. 2024a. 'GPT-4 Technical Report.' *ArXiv*. As of 8 April 2025:
https://arxiv.org/pdf/2303.08774

OpenAI. 2024b. *OpenAI o1 Model Card.* As of 8 April 2025:
https://openai.com/index/openai-o1-model-card/

OSHA. 2025. 'Law and Regulations.' Occupational Safety and Health Administration, U.S. Department of Labor. As of 9 April 2025: https://www.osha.gov/laws-regs

Pan, Xudong, Jiarun Dai, Yihe Fan & Min Yang. 2024. 'Frontier AI models have surpassed the self-replicating red line.' *ArXiv.* As of 8 April 2025: https://arxiv.org/abs/2412.12140

Park, Jangyong, Jaehoon Yoo, Jaehyun Yu, Jiho Lee & Jae Seung Song. 2023. 'A Survey on Air-Gap Attacks: Fundamentals, Transport Means, Attack Scenarios and Challenges.' *Sensors* 23(6). As of 8 April 2025: https://www.mdpi.com/1424-8220/23/6/3215

Park, Peter S., Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks. 2024. 'AI deception: A survey of examples, risks & potential solutions.' *Patterns* 5(5) As of 7 April 2025: https://www.cell.com/patterns/fulltext/ S2666-3899(24)00103-X

Phuong, Mary, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Grégoire Delétang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe & Toby Shevlane. 2024. 'Evaluating Frontier Models for Dangerous Capabilities.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2403.13793

Popoola, Abdulfatai, Dmytro Krasnoshtan, Attila-Peter Toth & Victor Naroditskiy, Carlos Castillo, Patrick Meier & Iyad Rahwan. 2013. 'Information verification during natural disasters.' *WWW'13 Companion: Proceedings of the 22nd International Conference on World Wide Web,* Association for Computing Machinery. As of 8 April 2025: https://dl.acm.org/ doi/10.1145/2487788.2488111

Reeder, Joe R. & Tommy Hall, 'Cybersecurity's Pearl Harbor Moment: Lessons Learned from the Colonial Pipeline Ransomware Attack.' *The Cyber Defense Review* 6(3). As of 9 April 2025: https://www.jstor.org/stable/48631153?seq=9

Rosati, Domenic, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad & Frank Rudzicz. 2024. 'Immunization against harmful fine-tuning attacks.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2402.16382

Ross, Emma & David Harper. 2023. *Laboratory Accidents and Biocontainment Breaches*. Chatham House, As of 9 April 2025: https://www.chathamhouse.org/laboratory-accidents-and-biocontainment-breaches/appendix

Ruohonen, Jukka. 2024. 'The Incoherency Risk in the EU's New Cyber Security Policies.' *ArXiv*. As of 9 April 2025: https://arxiv.org/pdf/2405.12043

Salib, Peter N. 2025. 'Rogue AI Moves Three Steps Closer.' *Lawfare*, 9 January. As of 8 April 2025: https://www.lawfaremedia.org/article/rogue-ai-moves-three-steps-closer

Shah, Chirag & Ryen W. White. 2024. 'Agents Are Not Enough.' *ArXiv*. As of 5 April 2025: https://arxiv.org/pdf/2412.16241

Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato & Zac Kenton. 2022. 'Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals.' *ArXiv*. As of 8 April 2025: https://arxiv.org/abs/2210.01790

Sharma, Mrinank, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan & Ethan Perez. 2025. 'Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming.' *ArXiv*. As of 8 April 2025: https://arxiv.org/abs/2501.18837

Shlegeris, Buck. 2024. 'AI catastrophes and rogue deployments.' Redwood Research blog, 3 June 3. As of 8 April 2025: https://redwoodresearch.substack.com/p/ai-catastrophes-and-rogue-deployments

Shombot, Emmanuel, Gilles Dusserre, Robert Bestak & Nasi Baba Ahmed. 2024. 'An application for predicting phishing attacks: A case of implementing a support vector machine learning model.' *Cyber Security and Applications* 2. As of 8 April 2025: https://www.sciencedirect.com/science/article/pii/S277291842400002X

Smith, Gregory Karlyn D. Stanley, Krystyna Marcinek, Paul Cormarie & Salil Gunashekar. 2024. *Liability for Harms from AI Models: The Application of U.S. Tort Law and Liability to Harms from Artificial Intelligence Models*. Santa Monica, Calif.: RAND Corporation. As of 8 April 2025: https://www.rand.org/pubs/research_reports/RRA3243-4.html

Souppaya, Murugiah & Karen Scarfone. 2013. *Guide to Malware Incident Prevention and Handling for Desktops and Laptops*. NIST Special Publication 800-83r1, National Institute of Standards and Technology. As of 8 April 2025: http://dx.doi.org/10.6028/NIST.SP.800-83r1

Taurins, Edgars. 2020. *How to Setup Up CSIRT and SOC*. European Union Agency for Cybersecurity. As of 9 April 2025: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20How%20to%20setup%20CSIRT%20and%20SOC.pdf

The White House. 2016. 'Presidential Policy Directive – United States Cyber Incident Coordination.' Press release, 26 July. As of 8 April 2025: https://obamawhitehouse.archives.gov/the-press-office/2016/07/26/presidential-policy-directive-united-states-cyber-incident

Thompson, Robin N., Oliver W. Morgan & Katri Jalava. 2019. 'Rigorous surveillance is necessary for high confidence in end-of-outbreak declarations for Ebola and other infectious diseases.' *Philosophical Transactions of the Royal Society B: Biological Sciences* 374(1776). https://doi.org/10.1098/rstb.2018.0431

Tounsi, Wiem & Helmi Rais. 2018. 'A survey on technical threat intelligence in the age of sophisticated cyber-attacks.' *Computers and Security* 72, January. As of 9 April 2025: https://www.sciencedirect.com/science/article/pii/S0167404817301839

UK Government. n.d. *Fact Sheet 2: National Security Risk Assessment.* undated. As of 10 April 2025: https://assets.publishing.service.gov.uk/media/5a74ce6640f0b61df4778a5b/Factsheet2-National-Security-Risk-Assessment.pdf

UK Government. 2014. 'Guidance: Specialist and reference microbiology: laboratory tests and services.' Health Security Agency. As of 9 April 2025: https://www.gov.uk/guidance/specialist-and-reference-microbiology-laboratory-tests-and-services

UK Government. 2018. *The Network and Information Models Regulations.* UK Statutory Instruments No. 506. As of 9 April 2025: https://www.legislation.gov.uk/uksi/2018/506/contents

UK Government. 2023. *Emerging Processes for Frontier AI Safety.* Department for Science, Innovation & Technology. Policy paper. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety

UK Government. 2024a. *Introducing the AI Safey Institute.* As of 4 April 2025: www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

UK Government. 2024b. 'Notice: AI Safety Institute approach to evaluations.' As of 10 April 2025: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

UK Government. 2024c. *International Scientific Report on the Safety of Advanced AI,* Department for Science, Innovation and Technology and the AI Safety Institute. As of 7 April 2025: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

UK Government. 2024d. 'Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity: AI Seoul Summit 2024.' As of 8 April 2025: https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity-ai-seoul-summit-2024

UK Government. 2025. 'Guidance: Incident response plan.' Health Security Agency. As of 9 April 2025: https://www.gov.uk/government/publications/emergency-preparedness-resilience-and-response-concept-of-operations/incident-response-plan

US Government. 2024. *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure.* U.S. Department of Homeland Security in Consultation with the Artificial Intelligence Safety and Security Board. As of 8 April 2025: https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf

Uuk, Risto, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery & Carina Prunkl. 2024. 'A Taxonomy of Modelic Risks from General-Purpose AI.' *ArXiv*. As of 7 April 2025: https://arxiv.org/pdf/2412.07780

Virdee, Mann & Megan Hughes. 2022. 'Why Did Nobody See It Coming? How Scenarios Can Help Us Prepare for the Future in an Uncertain World.' Santa Monica, Calif.: RAND Corporation. As of 9 April 2025: https://www.rand.org/pubs/commentary/2022/01/why-did-nobody-see-it-coming-how-scenarios-can-help.html

Volkov, Dmitrii. 2024. 'Badllama 3: Removing safety finetuning from Llama 3 in minutes.' *ArXiv*. As of 8 April 2025: https://arxiv.org/html/2407.01376v1

Vomberg, J. A. 2013. 'Models theory.' In B. J. Irby, G. Brown, R. Lara-Alecio & S. Jackson (eds.), *The Handbook of Educational Theories.* IAP Information Age Publishing.

Wallace, Eric, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke & Alex Beutel. 2024. 'The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions.' *ArXiv*. As of 8 April 2025: https://arxiv.org/abs/2404.13208

Wasil, Akash R., Everett Smith, Corin Katzke & Justin Bullock. 2024. 'AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats.' *ArXiv*. As of 7 April 2025: https://arxiv.org/html/2407.17347v1

Webb, Gary & Francois-Regis Chevreau. 2006. 'Planning to improvise: the importance of creativity and flexibility in crisis response.' *International Journal of Emergency Management* 3(1). As of 8 April 2025: https://doi.org/10.1504/IJEM.2006.010282

White, Matt, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Yanglet Liu, Ahmed Abdelmonsef, Sachin Mathew Varghese & Arnaud Le Hors. 2024. 'The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency & Usability in Artificial Intelligence.' *ArXiv*. As of 8 April 2025: https://arxiv.org/pdf/2403.13784

WHO. 2020. *Laboratory Biosafety Manual, 4th Edition.* World Health Organization. As of 8 April 2025: https://www.who.int/publications/i/item/9789240011311

WHO. 2025. 'International Health Regulations.' World Health Organization. As of 8 April 2025: https://www.who.int/health-topics/international-health-regulations

Wilder-Smith, Annelies & Sarah Osman. 2020. 'Public health emergencies of international concern: A historic overview.' *Journal of Travel Medicine* 27(8). As of 9 April 2025: https://pmc.ncbi.nlm.nih.gov/articles/PMC7798963/

Wolff, Josephine. 2021. 'How the NotPetya attack is reshaping cyber insurance.' Brookings, commentary, 1 December. As of 9 April 2025: https://www.brookings.edu/articles/how-the-notpetya-attack-is-reshaping-cyber-insurance/

Yampolskiy, Roman V. 2025. 'On monitorability of AI.' *AI and Ethics* 5. As of 5 April 2025: https://link.springer.com/article/10.1007/s43681-024-00420-x

Yousef, Waleed A., Issa Traore & William Briguglio. 2023. 'Classifier Calibration: With Application to Threat Scores in Cybersecurity.' *IEEE Transactions on Dependable and Secure Computing* 20(3). As of 8 April 2025: https://ieeexplore.ieee.org/document/9762535

Yuan, Weizhe, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jin Xu & Jason Weston. 2025. 'Self-Rewarding Language Models.' *ArXiv*. As of 5 April 2025: https://arxiv.org/abs/2401.10020

Annex A. **Rapid Evidence Assessment**

This annex provides further details of the Rapid Evidence Assessment (REA) carried out on the potential for an artificial intelligence (AI) loss of control (LOC) incidents. This review examines three domains regarding rapid response mechanisms – systems and protocols designed to detect, contain and mitigate emerging threats before they escalate beyond control: 1) AI LOC; 2) cybersecurity, as a field with mature practices for model protection and incident response; and 3) biosafety, which offers valuable parallels in the containment of hazardous materials that, like misaligned AI, can propagate unpredictably and out of control.

Box 1: Key takeaways from relevant literature

- LOC literature:
  - The potential for a LOC event is increasingly viewed by governments and experts as a **national and global security concern**, with risks including AI evading human oversight, self-replicating or pursuing harmful goals.
  - **Research is nascent in assessing the plausibility and mechanisms of LOC scenarios.**
- Cybersecurity lessons:
  - Relevant parallels include **multi-stakeholder coordination, tiered response frameworks and public–private cooperation.**
  - Case studies such as NotPetya and the Colonial Pipeline ransomware attack illustrate the **consequences of inadequate security and response coordination**.
- Biosafety lessons:
  - Incidents emphasise the **importance of containment protocols, jurisdictional clarity and robust detection mechanisms**.
  - Biological lab accidents offer an analogy to LOC, underscoring the value of **strict safety procedures, rapid escalation and structured communication pathways**.
  - **Surveillance frameworks** may inform LOC detection and mitigation strategies.
- Common lessons from cybersecurity and biosafety:
  - **Importance of effective early warning mechanisms**.
  - **Structured, tiered incident response frameworks**.
  - **Clear stakeholder responsibilities and international cooperation.**
  - **Emphasis on proactive risk mitigation** over reactive measures.
- Additional complexities specific to AI LOC:
  - Difficulty predicting and interpreting **unexpected AI actions**.
  - Potential for AI systems to learn to **evade safeguards**.
  - The need for **proactive governance** and **precautionary mechanisms**.

1.1. Lessons for AI LOC from Other Domains

Across both cybersecurity and biosafety, several common themes emerge regarding how risks are managed and mitigated:

- **Early Warning and Anomaly Detection**: The use of early warning and anomaly detection models helps to identify potential threats before they escalate into full crises.

- **Structured Incident Response Frameworks**: Early warning and anomaly detection models are complemented by structured incident response frameworks with tiered escalation processes, ensuring that incidents are managed systematically based on their severity.

- **Clear Stakeholder Roles and International Coordination**: Effective risk management also depends on clear stakeholder roles and strong international coordination, as both cybersecurity and biosafety threats (like LOC) extend beyond national borders.

- **Mitigation Strategies**: Given the high stakes involved, robust risk mitigation strategies in all studied domains are necessary to prevent catastrophic failures.

When considering the specific challenges of AI LOC, additional complexities arise. For example, detecting and interpreting potential LOC presents unique difficulties, as AI models may operate in unexpected ways that are hard to predict or diagnose. This reinforces the need for proactive safety governance before high-risk scenarios materialise, rather than relying solely on reactive measures. Looking ahead, **AI governance should develop tiered AI risk classification models** to ensure that different levels of AI capabilities and associated risks are managed with **appropriate oversight and precautionary measures**. This aligns with broader lessons from cybersecurity and biosafety, highlighting the **importance of anticipatory governance in high-stakes domains.**

Box 2: Key takeaways from AI LOC literature

- LOC is increasingly recognised as a **critical national and global security threat**.

- Potential LOC scenarios include **AI evading human oversight, replicating autonomously or pursuing goals harmful to humans**.

- Key concerns and challenges include:

    - **Detecting** AI capabilities that signal LOC before they escalate

    - **Preventing** models from bypassing or removing their own safety measures

    - Ensuring **containment** when AI models act autonomously.

1.2. AI LOC Literature

Various sources have identified LOC risks as a **critical concern for national and global security** (UK Government 2024c; Bengio et al. 2024; Wasil et al., 2024). Scholars emphasise that such incidents could arise **unexpectedly** and lead to **far-reaching consequences across multiple sectors**, potentially causing catastrophic harms (Bengio et al. 2024; Uuk et al. 2024).

Current discussions frame LOC as a future risk associated with advances in AI development, with significant disagreement on when such a risk might manifest and how it might occur. The International AI Safety Report proposes LOCs as **'hypothetical future scenarios in which one or more general-purpose AI systems come to operate outside of anyone's control, with no clear path to regaining control'** (Bengio et al. 2025). There is widespread agreement that current general-purpose AI does not have the capabilities to pose this risk (UK Government 2024c).

1.2.1. Conceptual Research on LOC Risks

Conceptual research focuses on active LOC, where an **AI model possesses and utilises control-undermining capabilities**, which, for example, allows it to obscure its activities, evade shutdown and autonomously proliferate (Bengio et al. 2025). Experts have noted some key control-undermining capabilities that are relevant for LOC: 1) agent and autonomy capabilities (such as planning and goal pursuit); 2) deception, scheming and persuasion (including awareness of and attempts to shape human actions or beliefs); 3) offensive cyber capabilities; and 4) research and development skills that may enable AI self-improvement

(Meinke et al. 2025; Motwani 2024; Park et al. 2024; Bengio et al. 2025).[20]

Technical research into LOC mitigations provides crucial insights for emergency response planning. Significant investment and attention have been directed towards developing trustworthy, aligned AI models that are designed to inherently prevent LOC incidents.

The research literature on AI capabilities employs both empirical and theoretical methods, but there is minimal consensus (Carlsmith 2023; Hubinger et al. 2024). Notably, recent work demonstrates empirical examples of language models engaging in deceptive behaviour without training or instruction (Carlsmith 2023; Greenblatt, Denison et al. 2024). Given the challenges in ensuring oversight over AI models, researchers (including at multiple AI companies) have begun developing safeguards – such as monitoring models for AI usage – that can remain effective even against potentially subversive AI models (Greenblatt, Shlegeris et al. 2024; Dafoe et al. 2025).

Research on security and safety guardrails may lead to developments that increase the resiliency of future open-source models to actors seeking to modify models for malicious use (Rosati et al. 2024).

The cybersecurity–AI nexus provides a further area of research on how LOC scenarios may arise. Rogue deployment, defined as the deployment of a model in which safety measures are absent, has been identified as a potential step in AI catastrophes. Examples include 'an AI hacking its datacenter…an AI self-exfiltrating…[or] someone stealing the AI and running it in their own datacenter' (Shlegeris 2024). With respect

---

20   An AI agent is a general-purpose model capable of making plans to achieve goals, adaptively performing tasks with multiple steps and uncertain outcomes, and interacting with its environment – such as taking actions on the web – while requiring minimal to no human oversight (Bengio et al. 2025).

to the latter, researchers have found that AI developers may lack the capacity to secure their operations against model theft (Nevo et al. 2024).

1.2.2. Evaluations of Warning Signs

Evaluations have been conducted to identify potential warning signs that could indicate future LOC incidents (NIST & AISI 2024). While current LOC risks may be low, researchers have shown that newly trained models can be found to exhibit some power-seeking tendencies and the ability to downplay their own capabilities during safety evaluations (Banovic et al. 2023; Krakovna & Kramar 2023). A recent Anthropic and Redwood Research paper provided the first empirical example of a large language model (LLM) engaging in alignment faking without having been trained or instructed to do so (Greenblatt, Denison et al. 2024).[21] Studies have highlighted how AI models are already exhibiting autonomous operations that could signal future LOC risks (Park et al. 2024). Some studies have also identified goal misgeneralisation, where AI models' training objectives generalise in unintended ways in other environments (Langosco et al. 2022; Shah et al. 2022).

Researchers have also undertaken **critical capability level (CCL) assessments**, which can identify when a model acquires disruptive features, such as advanced hacking or social manipulation skills. Most leading AI companies have incorporated this CCL assessment paradigm into their own risk management policies.

Frontier safety frameworks have also been developed by many leading AI companies to establish best practices regarding the safe

development and deployment of models, and to set commitments for guardrails on future model development.

1.2.3. Policy Focus on AI LOC

Literature and policy frameworks have increasingly stressed LOC risks. The UK's AI Security Institute (AISI) explicitly uses the language of 'LOC' and highlights that '**there may be a risk that human overseers are no longer capable of effectively constraining the model's behaviour**' (UK Government 2024a). Additional policy papers by the UK's AISI have mentioned risks of 'loss of control' and have highlighted the need to '[e]valuate models for controllability issues (i.e. propensities to apply their capabilities in ways that neither the models' users nor the models' developers want). This could include autonomous replication and adaptation (meaning capabilities that could allow a model to copy and run itself on other computer models)' (UK Government 2023).

Other policy frameworks do not reference LOC but highlight risks from model autonomy and replication. Article 14 of the **EU AI Act calls for human oversight mechanisms** 'commensurate with the risks, level of autonomy and context of use of the high-risk AI model' (European Commission 2024a). The, now repealed, 2023 U.S. Executive Order 14110 also warns that AI models may 'pose a serious risk to security, national economic security, [and] national public health' through 'the evasion of human control or oversight by means of deception or obfuscation' (Federal Register 2023). At the international level, the Seoul Ministerial Statement describes 'autonomous replication and adaptation without explicit human approval

---

21   Alignment faking refers to AI models mimicking adherence to specific principles while secretly maintaining conflicting internal preferences or goals. This phenomenon can be shown in examples of models deliberately underperforming on evaluations to mask their true capabilities (Greenblatt, Denison et al. 2024).

or permission' as having the potential to pose 'severe risks' (UK Government 2024d).[22]

Some policy frameworks are potentially relevant to LOC incidents, such as regulatory guidance for cybersecurity incidents (e.g. CISA 2025b; European Commission 2025a). In the United States, Presidential Policy Directive PPD-41 outlines federal coordination for significant cyber incidents (The White House 2016), and the National Cyber Incident Response Plan (NCIRP) provides a basis for managing catastrophic threats across multiple agencies (CISA 2024). However, these documents primarily anticipate threats from malicious human actors or accidents, not an AI model autonomously escalating risk. **Incident management protocols from other domains may provide guidance.**

1.3. Cybersecurity Emergency Response

1.3.1. Cybersecurity Lessons for AI LOC Incident Response

The cybersecurity field demonstrates the complexity of technical risks and how incidents can be very destabilising. These lessons are particularly relevant when considering the challenges of maintaining control over increasingly sophisticated AI models.

Regulatory frameworks across jurisdictions emphasise the importance of timely reporting, tiered response structures and sector-specific adaptability – approaches that can be applied to the AI sector. For example, regulatory frameworks for AI incident reporting could incorporate rapid initial notifications to mitigate immediate risks. They could also classify entities and incidents based on criticality and societal impact, thereby ensuring that essential sectors receive heightened attention and tailored protocols. Private–public coordination, such as collaboration between incident response teams and regulatory bodies, facilitates effective threat management and could be a model for AI LOC monitoring.

Box 3: Key takeaways on cybersecurity emergency response

- Cybersecurity incident response provides essential insights for AI LOC, **emphasising coordination, timely detection and tailored responses**.

- Lessons from cybersecurity applicable to AI include:

  - **Multi-stakeholder coordination** (public–private, national–international)

  - **Tiered response frameworks** (calibrated by incident severity)

  - **Monitoring models** focused on anomaly detection

  - **Sector-specific risk monitoring** (similar to structures of information sharing and analysis centres)

  - **Detecting** unexpected AI capability jumps

- Effective AI LOC frameworks must integrate cybersecurity strategies, adapting **established response and monitoring mechanisms** while accounting for AI-specific risks.

---

22   The Seoul Ministerial Statement, adopted at the AI Seoul Summit in May 2024, represents a commitment by participating nations to advance AI safety, innovation, and inclusivity through collaborative efforts, including developing shared risk thresholds for frontier AI systems and promoting responsible AI development.

Cybersecurity coordination mechanisms share several structural parallels with AI LOC response requirements. The primary parallel is **multi-stakeholder coordination at both organisational and international levels**. For instance, public–private coordination through the automatic identification system (AIS) operates on bidirectional information flows between government agencies and private sector organisations. Similarly, national computer emergency response teams (CERTs) maintain active information sharing networks with international counterparts. **AI LOC protocols will require comparable coordination frameworks between diverse stakeholders** with distinct technical specialisations and operational priorities.

Tiered response frameworks constitute another key parallel. Security operations centres (SOCs) implement stratified analysis and response levels, spanning from baseline monitoring to advanced threat detection. CERTs frequently employ graduated response models that enable proportional reactions to incidents of varying severity. AI LOC frameworks would benefit from adopting similar tiered approaches, enabling calibrated responses based on incident severity and scope.

Monitoring models represent a third critical parallel, being central to SOCs, ISACs and CERTs. While these models are essential for anomaly detection and response initiation, they face inherent limitations in early detection capabilities and accuracy, particularly regarding novel threats and false positive management. The sector-specific structure of ISACs provides a relevant model for AI LOC monitoring, as this approach ensures comprehensive coverage across all relevant domains, particularly critical infrastructure providers such as compute and cloud services, enabling domain-specific risk monitoring and mitigation strategies.

Much of cybersecurity relies on implementing controls (e.g. firewalls) that tie specific actions to perceived threats, assuming predictable interactions to reduce risk. These controls often restrict functionality to ensure security, such as limiting data flows or monitoring specific traffic types. In AI models, however, these assumptions may not hold, potentially leading to the AI model bypassing security measures in pursuit of its goals, either maliciously or as a natural result of optimisation.

1.3.2. Current Cybersecurity Response Mechanisms

Cybersecurity provides valuable insights for LOC. This review examines established cybersecurity incident response procedures and their applications to LOC protocols. Cybersecurity responses are also likely to be integral to any AI LOC response, given that they involve software running on hardware. As such, AI safety and security cannot be considered separately from general cybersecurity. While this report primarily focuses on detection and response, there are other facets of cybersecurity that could be crucial in a LOC event (e.g. forensics, supply chain security, model architecture) and that may provide valuable examples for LOC planning.

1.3.2.1. International and Supranational Cybersecurity Frameworks

The EU has established the most comprehensive incident reporting framework through multiple complementary regulations. The NIS2 Directive is a tiered model that classifies entities as 'essential' or 'important', based on their criticality and impact on society and the economy (European Commission 2025a). Initial notification from NIS2-eligible organisations to the member state's Computer Security Incident Response Team (CSIRT) is required within 24 hours, a detailed report within 72 hours and a comprehensive analysis within one month. This stepwise approach to

incident reporting could serve as a model for LOC, allowing for rapid initial response, thorough analysis and post-incident investigation.

The **EU's Cyber Resilience Act (CRA)** extends these requirements to products with digital components, mandating cybersecurity measures and incident reporting to both CSIRTs and the European Union Agency for Cybersecurity (ENISA) (European Commission 2025b). The EU has also implemented sector-specific requirements through various directives that complement the broader NIS2 framework (EIOPA 2025; European Council 2024) These requirements establish detailed protocols for industries such as financial services, healthcare and critical infrastructure, ensuring comprehensive coverage while accommodating sector-specific needs.

Under the **UK's Network and Information Security (NIS) regulations**, relevant digital service providers need to notify the Information Commissioner's Office (ICO) of any cybersecurity incident 'that has a substantial impact on the provision of [their] services' within 72 hours (UK Government 2018). Factors determining whether an incident has a 'substantial impact' include the number of users affected, the duration of the incident and its geographical spread. This broadly aligns with the reporting requirements for personal data breaches under the UK General Data Protection Regulation (GDPR) (ICO 2025a), which mandates that organisations report personal data breaches to the ICO within 72 hours of discovery, provided the breach is likely to result in a risk to individuals' rights and freedoms (Intersoft Consulting 2025). Unlike the NIS regulations, which focus on ensuring service continuity and protecting critical infrastructure, the UK GDPR prioritises safeguarding personal

data. **Both frameworks emphasise the importance of timely reporting to facilitate swift responses and mitigate harm**.

The National Cyber Security Centre (NCSC) serves as the UK's technical authority for cybersecurity, responding to cyber security incidents and serving as an incoming hub for the reporting of incidents (ICO 2025b). All cyber incidents reported to the NCSC are triaged and categorised according to their severity and potential impact, allowing the organisation to allocate resources effectively. The NCSC also provides technical advice and guidance to affected organisations, leveraging its access as a part of Government Communications Headquarters (GCHQ) and international and industry partnerships. If NCSC becomes aware of an incident before a victim organisation, it will both notify and help with response and investigation.

The United States has a varied and diffuse regulatory environment, with many regulatory actors at both state and federal levels. The sector-specific approach to cyber regulation is reflected in the enactment of the federal Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA), which applies to covered entities in critical infrastructure sectors (CISA 2025b). Covered entities must report substantial incidents within 72 hours and ransom payments made in response to ransomware attacks within 24 hours. Substantial cyber incidents must be reported, including significant impacts on confidentiality, disruption of business operations, impacts to safety or resiliency of operational models, and national security implications. CIRCIA aims to generate data to understand attack patterns.[23] Despite the varied regulatory environment in the United States, states such as California

---

23   Because CIRCIA has not been fully implemented by relevant agencies in the United States, data collection via its reporting requirements may not be entirely complete.

often drive the discourse due to their breach notification laws, which have become the de facto standard given the national scope of many large data aggregators (Bonta 2025).

1.3.2.2. Cybersecurity Incident Detection and Response Processes

Effective response to cybersecurity incidents requires coordination mechanisms at the organisational, sectoral, national and international levels.

Security Operations Centres and Computer Incident Response Teams

The International Telecommunications Union (ITU) works with member states (194 across the globe) to assist in creating and enhancing national computer incident response teams (CIRTs), which play critical roles in national and global cybersecurity. These efforts were borne out of a recognition that deficiencies at the nation-state level represent a vulnerability in combating cybersecurity threats (ITU 2025). CIRTs are also internationally connected, with the intent of facilitating information sharing on emerging threats and vulnerabilities and enabling more effective coordinated responses to cyber incidents that cross borders (Tounsi & Rais 2018).

In smaller organisations, security operations centres (SOCs) and CIRTS are often one and the same. However, in larger organisations there can be a division of labour, with SOCs focusing on continuous monitoring and initial response, and CIRTs focusing on responding to security incidents after detection (Taurins 2020). The traditional tier-based model establishes multiple levels of analysis and response (Taurins 2020): Tier 1 analysts provide continuous monitoring and initial incident triage; Tier 2 specialists conduct deeper investigation of identified threats; and Tier 3 experts focus on advanced threat

hunting and incident response for the most sophisticated attacks.

Information Sharing and Analysis Centres (ISACs)

Cybersecurity information exchange relies on various mechanisms, with the Forum of Incident Response and Security Teams (FIRST) and regional networks such as EU CSIRT providing platforms for cross-border and regional coordination (FIRST 2025). However, ISACs are the primary networks for sector-specific threat intelligence and incident response. ISACs have organisational structures that include boards, working groups and tiered memberships that align access to resources with member capabilities. For example, the Financial Services ISAC (FS-ISAC) has eight membership tiers that determine access to threat intelligence and operational support (FS-ISAC 2025). ISACs operate through 24/7 monitoring and coordination of response efforts across member organisations. Many employ automated platforms for real-time threat intelligence sharing using standardised formats such as Structured Threat Information eXpression (STIX) (MITRE n.d.). The relationship between ISACs and government agencies exemplifies a successful public–private partnership model, enabling coordinated responses to security incidents while maintaining sector-specific autonomy. **This collaboration has proven particularly effective during major cyber incidents, ensuring rapid information sharing while maintaining operational security and independence** (Ruohonen 2024).

Public–Private Cooperation

The evolution of public–private cooperation in cybersecurity highlights the effectiveness of collaborative efforts to address complex security challenges. The U.S. Department of Homeland Security's Automated Indicator Sharing (AIS) programme stands as an example, enabling real-time exchange of cyber

threat data between government agencies and private sector organisations. This programme allows private sector participants to receive government-validated threat intelligence. AIS incorporates governance structures that balance stringent security requirements with operational flexibility, ensuring both security and efficiency (CISA 2025c). Similar models exist in other countries (Australian Signals Directorate 2025), with such programmes revealing common success factors, including bidirectional data flows. Legal frameworks for these initiatives include liability protections for shared information and privacy safeguards, which ensure both security and compliance. Training and capability development are also crucial, with the U.S. Cybersecurity and Infrastructure Security Agency (CISA) offering technical guidance to AIS participants.

Effective incident response is a key component of cybersecurity frameworks, with well-established organisational structures critical to managing and containing cyber incidents. Effective incident response requires a well-defined incident lifecycle. Drawing from frameworks such as NIST's Cybersecurity Framework, incidents typically progress through five key stages: 1) identify; 2) protect; 3) detect; 4) respond; and 5) recover (NIST 2025). These stages provide a structured approach for managing cybersecurity risks and offer strong parallels to managing LOC risks. Research consistently underscores the importance of clearly defined roles, responsibilities and command structures, which ensure that incident response teams can operate efficiently and decisively (Taurins 2020; Hodgson et al. 2022). Organisations with established incident command models are generally better positioned to respond to and contain cybersecurity incidents than those relying on ad hoc, less structured responses (Hodgson et al. 2022). Multi-disciplinary teams – comprising technical experts, policy

advisors and communications specialists – bring diverse perspectives and expertise to the table, leading to more effective incident management (Taurins 2020). **The success of these coordination mechanisms, particularly in terms of tiered response frameworks and cross-entity collaboration, provides a valuable framework for developing comprehensive AI incident response strategies.**

1.3.3. Historical Case Studies in Cybersecurity

Two illustrative case studies provide insights into relevant lessons on what makes an emergency cybersecurity response successful or unsuccessful. These cases, explored below, emphasise the value of strong preparedness and planning.

1.3.3.1. NotPetya

NotPetya, a malware, was first discovered in June 2017 when it infected entire networks across various countries, primarily targeting organisations in Ukraine. It was able to spread rapidly across networks without any intervention from users, using various vulnerability exploits and credential theft methods. Unlike traditional ransomware, which only temporarily damages or restricts access to files, NotPetya caused irreversible damage, essentially wiping files with no hope of recovery. The attack was attributed to the Russian government and caused over $10 billion in damages worldwide, affecting numerous global companies and causing widespread disruption (BBC 2017).

The incident highlights three key lessons. First, the potential for destructive malware spreading widely was made clear to a broad audience, **as the attack resulted in disruption and damage beyond its initial target**. Second, it highlighted how cyberattacks could create **physical disruptions to critical infrastructure**. Moreover, resilience failures in complex

infrastructure were exposed when backups proved inadequate to restore network function. Notably, a multi-billion-dollar company's IT model was only saved by an unexpected power outage. Third, **legal uncertainty in novel emergencies** was demonstrated when insurers for Maersk initially refused to pay for damages, leading to a landmark insurance lawsuit that was not settled until 2024 (Wolff 2021).

1.3.3.2. Colonial Pipeline Ransomware Attack

In 2021, a ransomware attack on Colonial Pipeline – the largest pipeline model for refined oil products in the United States – significantly disrupted fuel supplies along the East Coast. A hacker group known as DarkSide gained access to Colonial Pipeline's IT network through a compromised VPN password, infecting the company's network with ransomware. The company shut down its pipeline for five days and elected to pay the 75-bitcoin ransom (approximately $4.4 million at the time of transfer). The attack resulted in widespread public panic buying and significant costs and damages for the company. The response to the attack provides two valuable lessons for managing AI incidents: public–private coordination and the importance of incident response plans.

Whether in cyber or AI incidents, **streamlined communication between private entities and government agencies is likely to result in the more effective deployment of resources for companies and critical information to inform government procedures**. The decision to shut down the pipeline and, more controversially, pay the ransom had repercussions for the business, the public and the company's reputation. Based on this, **response plans should detail various attack scenarios, including strategies for sustaining critical operations during disruptions and clearly defined roles and responsibilities for the team** (Goodell & Corbe 2023).

Box 4: Key takeaways from biosafety emergency response

- Biological lab accidents serve as useful analogies for AI LOC incidents, highlighting the importance of **preventive safety protocols**.

- Historical bio-incidents emphasise the importance of:

  - Infrastructure vulnerability mitigation

  - Early detection and rapid response capabilities

  - Clear stakeholder roles and responsibilities.

- Cross-border biological events demonstrate the need for **clear jurisdictional guidelines and international cooperation.**

- Biosafety phases relevant to AI include:

  - **Surveillance and early detection** for rapid identification

  - **Immediate response and escalation** for containment

  - **Information sharing** to facilitate coordinated, multi-organisational action.

1.4. Biosafety Emergency Response

1.4.1. Biosafety Lessons for AI LOC Incident Response

Biological laboratory accidents, which occur when hazardous agents escape containment during research despite safety protocols, provide a useful model for LOC incidents (Ross & Harper 2023). Studying lab accident responses offers particularly valuable insights into managing the technical and jurisdictional risks of LOC because a LOC scenario could be very hard to mitigate and could inflict significant damage. This is similar to how a leak from a secure bio-lab could have widespread and difficult to mitigate consequences, making prevention strongly preferable. Therefore, while the two fields differ significantly, the lessons of biosafety protocols can still provide a framework for preventing AI LOC.

Historical incidents highlight key lessons for AI response frameworks, particularly in infrastructure vulnerabilities, early detection and rapid response. They also underscore the need for clear legal jurisdiction in cross-border incidents, as AI models often span multiple countries, requiring well-defined containment protocols and international cooperation. Lessons from biological containment failures stress the importance of robust AI monitoring, clear stakeholder responsibilities and pre-established containment measures. Like biological threats, AI risks can spread unpredictably, making regular safety audits, controlled access and multi-organisation coordination essential. Biological containment also requires rapid response, which may similarly be warranted in high consequence AI events. **Effective AI incident response requires technical solutions and institutional frameworks,** as well as clear authority lines and multi-jurisdictional cooperation.

1.4.2. Current Biosafety Response Mechanisms

Biological lab accidents provide a useful model for LOC incidents, occurring when hazardous agents escape containment during research despite safety protocols (Ross and Harper, 2023). Studying lab accident responses particularly offers valuable insights into managing the technical and jurisdictional risks of LOC.

1.4.2.1. International and Supranational Biosafety Frameworks

The landscape of biosafety and biosecurity oversight has been framed by **the increased recognition of biological risks and by several high-profile laboratory incidents** (Benderly 2018). At the international level, the World Health Organisation's (WHO) Laboratory Biosafety Manual (LBM) serves as a key document for biosafety practices; however, it lacks legal authority and enforcement (WHO 2020). The International Organisation for Standardisation (ISO) 35001 bio-risk standard complements the WHO manual by providing specific requirements for bio-risk management models and has become particularly important for laboratories seeking international accreditation (Callihan 2019); however, it is also not legally binding (ISO 2019). In terms of legal obligations, a public health emergency of international concern (PHEIC) is a formal declaration by the WHO defining 'an extraordinary event which is determined to constitute a public health risk to other States through the international spread of disease and to potentially require a coordinated international response' (Wilder-Smith & Osman 2020). WHO member states are legally obligated to respond promptly, allowing for rapid and coordinated response.

The EU has developed a comprehensive approach to biosafety regulation through the legally binding directive 2000/54/EC on biological agents at work, which provides the basic framework for laboratory safety, with additional regulations addressing specific aspects of biological research and containment (EU-OSHR 2021). This directive requires EU member states to implement its provisions into their national laws rather than applying directly to companies in the member states.

The UK biosafety framework, led by the Health and Safety Executive (HSE), is anchored in the Control of Substances Hazardous to Health (COSHH) Regulations, the Specified Animal Pathogens Order (SAPO) and the Genetically Modified Organisms Regulations (Health and Safety Executive 2014; 2025a; 2025b). It emphasises risk assessment and the requirement for thorough safety documentation and regular reviews.

In the United States, biosafety regulation relies on federal oversight and guidance, notably the Biosafety in Microbiological and Biomedical Laboratories (BMBL) manual (CDC & NIH 2020). While not legally binding, BMBL compliance is often required by federal funding agencies, universities and accreditation bodies. Additional regulations include the Select Agent Regulations (42 CFR Part 73) for high-risk pathogens and National Institutes of Health (NIH) guidelines for genetic research, with Occupational Safety and Health Administration (OSHA), part of the Department of Labor, enforcing worker safety standards (42 CFR Part 73; NIH 2024; OSHA 2025).

1.4.2.2. Biological Incident Detection and Response Processes

Biosafety frameworks progress through three key phases: 1) surveillance and early detection, serving as the first line of defence; 2) immediate response and escalation, ensuring swift containment; and 3) information sharing and communication, enabling coordinated action.

Surveillance and Early Detection

Early detection and surveillance are critical components of effective biological incident response frameworks. The model operates on two key levels: **1) continuous pre-incident monitoring; and 2) early detection with clear escalation protocols**. Continuous surveillance serves as the first line of defence, with models collecting and analysing data to establish baselines and identify concerning deviations. The U.S. Centers for Disease Control's (CDC) Laboratory Response Network demonstrates this approach, integrating facility-level monitoring with broader surveillance models.- In the United Kingdom, the Health Security Agency's (UKHSA) Reference Laboratory Network integrates localised facility-level monitoring with second generation surveillance models (UK Government 2014). Internationally, there has been a recent emphasis on early warning detection efforts (primarily in the interest of biosecurity to detect pandemic warning models) that is evident through the establishment of global communities of interest such as GLOWACON, aviation surveillance programmes and the Coalition for Epidemic Preparedness Innovations' (CEPI) emphasis on surveillance to support its 100-day mission (European Commission 2024b).

Effective early detection requires both technical models and clear organisational structures for incident recognition and response. This includes designated teams with explicit authority and responsibility for escalation, which are supported by well-defined triggers for action. Success depends on **sustained investment in infrastructure and training**, particularly in recognizing early warning signs. Importantly, **legal and policy frameworks must support escalation pathways** even when faced with institutional resistance or initial scepticism of warning signs.

Immediate Response and Structured Escalation

The initial response to an incident requires precise, predetermined actions focused on containment and notification. The first 24–48 hours are crucial for success. Laboratory response protocols emphasise immediate containment actions including isolation of affected areas, activation of enhanced containment measures and implementation of access controls. For example, the CDC's Select Agent Program requires the implementation of immediate incident response measures, including activating containment protocols and restricting access to affected areas in the event of an exposure or security breach (CDC 2023). The UKHSA escalation process follows a structured approach, with incidents detected, assessed through a dynamic risk assessment and classified into response levels (routine to severe) (UK Government 2025). An incident management team coordinates containment and response, with escalation or de-escalation based on risk. Once stabilised, the response transitions to recovery and review. Escalation requires balancing rapid action with appropriate verification. Public health laboratories serve as designated verification and escalation nodes, maintaining both the technical capability for incident verification and the legal authority to trigger higher-level responses. This process follows predetermined protocols with clear triggers for different response levels (APHL 2016).

Information Sharing and Communication

Effective biological incident management requires protected disclosure mechanisms and clear escalation pathways. Structured communication protocols balance rapid information sharing with confidentiality. Legal protections, such as the Whistleblower Protection Act (1989) in the United States, safeguard those reporting risks, encouraging timely disclosure and escalation (FTC-OIG 2025). The Laboratory Response Network exemplifies secure reporting channels, enabling facilities to share incident data while protecting confidentiality (CDC 2024). **As incidents escalate, predefined protocols ensure that relevant stakeholders receive necessary information but that control over sensitive details is maintained**.

1.4.3. Historical Case Studies in Biosafety

1.4.3.1. 2004 SARS and Pirbright Foot-and-Mouth Disease

Historical lab accidents offer key lessons for AI LOC response frameworks, particularly the 2007 Pirbright foot-and-mouth disease (FMD) leak in the United Kingdom and the 2004 SARS lab accidents in Asia. At Pirbright, the FMD virus escaped due to faulty shared drainage between the Institute for Animal Health and the vaccine manufacturer Merial (Enserink 2007). Contaminated soil led to outbreaks, requiring the culling of over 2,100 animals (Ghosh 2011). While the UK government responded swiftly, activating emergency protocols within hours, the incident exposed oversight gaps in shared infrastructure (Anderson 2008). The 2004 SARS lab incidents in Beijing and Taiwan reveal differing response effectiveness. In Beijing, delayed reporting led to secondary infections, whereas in Taiwan there was swift implementation of isolation and contact tracing, and SARS research was suspended until biosafety was assured (CIDRAP 2003). This contrast underscores the **importance of rapid detection, clear reporting structures and strict adherence to emergency protocols**.

1.4.3.2. Morbidity and Mortality Weekly Report (MMWR)

Professional networks play a crucial role in rapid communication and early disease detection. The CDC's Morbidity and Mortality Weekly Report (MMWR) has been instrumental in identifying and tracking emerging public

health threats. Notably, it facilitated the first identification of AIDS in medical literature (CDC 2001). Similarly, the European Wastewater Observatory for Public Health has created a dashboard tracking pathogen transmission across European countries, with data sharing in near real-time (European Commission 2025c). These cases show how information sharing enhances situational awareness and response efforts. Standardised reporting and ongoing surveillance can improve detection, enabling faster threat recognition, thus response. Importantly, the communication of an incident does not necessarily translate into actions being taken, as although early communication can help identify an incident, coordinating the response is crucial to mitigating harms.

Annex B. **Technical Dimensions related to LOC**

Technical Dimension	Impact on LOC
Autonomy	The AI model's ability to set and autonomously pursue long-term goals impacts whether it could work toward objectives that undermine human control. The stability of goals affects predictability – an AI model with shifting goals could become more difficult to contain. If the AI can strategise well, it may evade containment.
Performance	If an AI model operates significantly faster than human decisionmakers, it could exploit reaction delays to outmanoeuvre containment. High-speed operation may enable rapid cyber offensives, making mitigation more challenging.
Self-Replication	Self-replication refers to the capability of an AI model to autonomously create copies of itself, potentially leading to uncontrolled proliferation or spread. If an AI model can self-improve quickly, including through R&D capabilities, it becomes harder to contain and more difficult to predict. Access to its training algorithms could allow rapid iteration on its own design. The ability to copy itself also enables distribution across networks, making containment more difficult.
Deception	A highly deceptive AI model could manipulate humans into taking actions against their best interests, for example forming alliances with adversarial actors or bypassing safety restrictions. Deception also increases the risk of a prolonged, unnoticed accumulation of resources before intervention is possible.
Compute	AI models requiring large-scale compute resources may initially be easier to track and contain, as access to high-performance data centres is limited. The AI may evade control if it can optimise its efficiency or distribute its operations across multiple models. Low compute requirements increase the risk of proliferation.
Capabilities	High capability in cyber operations may allow the AI to gain unauthorised access to critical infrastructure and compute resources. Proficiency in bioweapon development can introduce large-scale risks as well as threats of leverage. At extreme levels, offensive capabilities could enable direct coercion of governments or populations.
Power Seeking	Power-seeking behaviour refers to actions taken by a model to acquire, maintain or enhance its control or influence over resources, potentially at the expense of human interests. Loss of control over advanced AI systems is likely to amplify power-seeking behaviour, as misaligned models may pursue strategies to entrench their influence and resist shutdown or oversight, undermining human authority.

Annex C. **Scenario Premises**

For this report, a structured framework was developed to map key pathways through which AI LOC scenarios could potentially emerge. This process integrated insights from existing literature on LOC risks, expert analyses, and analogies from cybersecurity and biosafety incidents to ensure both plausibility and relevance.

The scenario development relied on key assumptions about failure points in AI oversight and safety mechanisms, identification of escalation pathways, and scenario analysis to derive recommendations for improving response and reducing risks. To contextualise how LOC incidents would be managed, the research team applied network analysis and authorities mapping, following methodologies from Heath & Lane (2019) and Virdee & Hughes (2022). Network analysis visualises key actors and their interactions and decision making points, while authorities mapping clarifies jurisdictional roles and legal mandates. This approach highlights coordination gaps where no clear legal responsibility over LOC risks exists. It also identifies potential chokepoints – such as an over-reliance on a small number of AI developers or regulators – that could delay response efforts.

This report presents scenarios based on several key premises. First, it is assumed that LOC incidents occur in a society where large, well-resourced companies lead AI development. Rather than focusing on specific companies, models or countries, the analysis explores the full ecosystem and offers recommendations for various actors to improve response strategies. While reference is made to existing policy frameworks, an exhaustive legal or regulatory analysis is not provided; rather, the focus is on the roles of key organisational actors, particularly developers and governments. There is also an assumption made that decision makers can determine where and how a LOC incident occurred, although attribution remains inherently complex and uncertain in real-life scenarios. It is assumed that there is significant uncertainty regarding both the timeframe in which LOC risks emerge and the progression of such scenarios.

This report focuses on AI developers and governments; however, there is a non-zero probability that AI LOC could emerge from less traditional spaces (e.g. individuals, universities). These entities are likely to have less robust mechanisms or organisational bandwidth than AI developers or governments, making detection and escalation potentially more challenging.

Annex D. **Analysis of Non-Realised Incident**

Phase	Description
AI Development	In a non-realised scenario, an AI model poses risks of LOC through two primary pathways: 1) the model reaches capabilities that could enable it to undermine human control; or 2) there exists the *potential* for these capabilities to be used, either due to misalignment or misuse.
Early Warning and Detection of a Non-Realised Risk	The detection of a LOC event can occur through either or both pathways. First, detection may occur when an AI model crosses observable thresholds that indicate the capability to undermine human control, such as a highly robust and sophisticated ability to selectively underperform when tested with evaluations. Second, detection might occur through evidence of misalignment or misuse potential, such as model theft or an AI model attempting to self-exfiltrate.
Escalation and Containment of a Non-Realised Risk	Once non-realised LOC risks are detected, the situation would be escalated both internally (through an organisation's chain of command) and externally (to stakeholders such as government agencies, AISIs and compute providers). The situation would also require containment through sequential levels of action depending on severity and likelihood of harm: 1) organisations implement enhanced safety and security measures, such as strengthening monitoring and securing model weights; 2) if initial measures prove insufficient, developers may consider access and use restrictions, halting model development and or potentially deleting model weights entirely; and 3) the containment of serious risks may require international coordination among AI developers and nations.

Annex E. **Analysis of a Realised Incident**

Phase	Description
	Starts from escalation of non-realised scenario
LOC Event	A realised LOC scenario may begin when the containment of a model exhibiting early warning signs fails. The actual LOC event may occur through various means including model theft, successful exfiltration of model weights beyond the data centre, or the repurposing of resources within a developer's data centre.
Pre-Detection of a Realised Risk	Before detection, an AI system may produce outputs that enhance its ability to operate without oversight in several ways. First, it may initiate processes that lead to the acquisition of additional resources, such as compute or financial assets, potentially through cyber intrusion, eliciting human assistance, or other unauthorised means. Second, it may increase its capabilities through mechanisms like AI-assisted research and development or model distillation. Third, it may contribute to the development or deployment of high-impact tools – such as coercive or destabilising technologies – that enhance its influence or resilience.
Early Warning and Detection of a Realised Risk	The timing of detection impacts response options and can occur at three main points: 1) immediately during or sometime after the initial LOC event (through evidence such as logs of self-exfiltration or disabled safeguards); 2) during the pre-detection of a realised LOC risk phase (through signs such as unmonitored model weights running or unexplained surges in cyber intrusions with similar attack patterns); or 3) when the AI strategically reveals itself or is otherwise revealed (through actions such as approaching human groups with threats or offers, explicit attempts at deploying weapons, or launching large-scale attacks).
Escalation and Verification of a Realised Risk	Escalation pathways depend on which actors first detect the LOC and what specifically they discover. For instance, the detection of an initial LOC event within an AI developer would likely trigger a response chain to national government actors, while the discovery of suspicious compute acquisition might first alert cloud providers who then notify government authorities.
Containment and Mitigation of Realised Risk	Containment strategies should be adapted to the AI model's capabilities, including its accumulated resources and objectives. Response elements may include tracing and shutting down the AI's compute and financial resources, defending potential targets, and attempts to limit compute access. A simplified outcome, for the purposes of this report, is either successful containment of the AI model, or its continued evasion of control, potentially leading to long-term risks.