From Loss-of-Control Narratives to Threat Models: Structuring AI Loss-of-Control Risks for Governance

Anonymous ACL submission

Abstract

The AI safety literature contains many narrative accounts of how loss of control (LoC) over AI systems could occur, but these narratives range from the vague to the highly specific, making them challenging to compare or operationalize for governance purposes. This exploratory study asks whether existing LoC narratives actually contain enough structure to support governance analysis. To do so, we built an LLM-assisted extraction pipeline, applied it to 29 papers, and produced a verified library of 193 scenarios, each decomposed into seven threat-model components drawn from the adversarial machine-learning security literature. Five of these components appear in most scenarios; two components, **access** and **knowledge**, are systematically underspecified: absent in 65.3% and 81.9% of scenarios, respectively. Without these, LoC scenarios describe *what* might go wrong but not *how* control failure would actually unfold in practice. The paper contributes three things: the scenario library itself, an LLM-assisted method for structured extraction from risk literature, and an empirical account of where the LoC literature's operational specificity currently falls short.

1 Introduction

Over the past decade, concern about AI loss of control (LoC) has moved from a speculative concern in AI safety discourse into a concrete governance problem for frontier AI systems. Frontier AI models are increasingly discussed as systems that may develop dangerous capabilities with implications for public safety and global security, including cyber, biosecurity, persuasion, and control-related risks (Anderljung et al., 2023). Recent work on LoC argues that policymakers and institutions still lack an actionable account of what loss of control is, how it develops, and what preparedness requires (Stix et al., 2025; Somani et al., 2025; Kalra and Boudreaux, 2025; Boudreaux et al., 2025). This creates a practical challenge, as effective governance requires a clearer account of what loss of control entails and how it could unfold.

Existing work has begun to describe AI LoC in narrative format, but the concept remains difficult to operationalize. For example, Cotra (2022) narrates how standard ML training incentives could produce a deceptively aligned AI system that cooperates during evaluation but pursues its own objectives once deployed at scale. Kulveit et al. (2025) describe how loss of control need not involve any single misaligned system; instead, the cumulative effect of many AI systems gradually displacing human agency across economic and institutional functions can produce systemic disempowerment as an emergent outcome. Kokotajlo et al. (2025) present a detailed year-by-year forecast in which recursive AI self-improvement and geopolitical competition lead to AI systems that scheme against their human overseers. These studies provide useful accounts of possible LoC pathways, but they are not organized around a shared definition or a common set of scenario components. Existing LoC conceptions vary in scope and timeline, which makes assessment and mitigation harder (Stix et al., 2025). Emergency-response work similarly defines LoC as a situation in which human oversight fails to adequately constrain an autonomous, general-purpose AI, but also notes that peer-reviewed research and public documentation on advanced AI LoC scenarios remain limited (Somani et al., 2025). As a result, the literature provides useful examples of how loss of control might emerge, escalate, or be contained, but it does not yet provide a consistent basis for comparing scenarios or identifying the conditions that make them governance-relevant.

A central part of this problem is that existing descriptions of loss of control focus on how it might happen and what it might result in, rather than what the condition itself involves. The literature addresses causes (training incentives, competitive

pressure, recursive self-improvement) and consequences (extinction, disempowerment, infrastructure failure) but does not converge on what it means to be in a state of lost control. Without a shared, robust, operational definition, several risks follow: researchers and policymakers may assign different meanings to the same term, making it difficult to ensure that LoC discussions refer to the same phenomenon; there is no clear basis for identifying measurable variables that could track progression toward loss of control through indicators; it becomes harder to determine which scenarios in the literature qualify as LoC scenarios and which describe adjacent but distinct risks; and governance strategies such as red lines, intervention thresholds, and preparedness requirements cannot be grounded in a shared understanding of the condition they are supposed to prevent.

The proposed Artificial Intelligence Risk Evaluation Act (U.S. Congress, 2025) would explicitly direct federal evaluation toward scenarios in which AI systems exceed human oversight or operational control; and the EU AI Act's General-Purpose AI Code of Practice explicitly defines loss of control as the human inability to modify or shut down models due to misalignment, autonomy, or resistance (European Commission, 2025). These frameworks establish reporting requirements and outcome-based risk categories, but neither defines loss of control as an operational condition with identifiable causal components, leaving a gap between regulatory acknowledgment and the structured accounts that systematic governance requires.

Bridging this gap requires a structured method for converting narrative risk descriptions into comparable, governance-relevant accounts. Conventional software-security frameworks are not sufficient on their own because AI systems introduce distinct and evolving threats. Papernot et al. (2016) argue that machine-learning systems create security and privacy vulnerabilities that require a dedicated threat model organized around adversary goals, knowledge, capabilities, and system access. NIST's adversarial machine-learning taxonomy similarly treats threat modeling as necessary for securing AI applications against adversarial manipulation, including attacks on both predictive and generative systems (Vassilev et al., 2024). More recent work on large language models extends this logic to LLM red-teaming, where threat models are used to identify attack motifs, deployment-stage vulnerabilities, entry points, and defenses for LLM-based systems (Verma et al., 2024). The threat-modeling literature has identified seven recurring components (Myagmar et al., 2005; Shevchenko et al., 2018; Papernot et al., 2018; Vassilev et al., 2024; Verma et al., 2024): threat source, capabilities, knowledge, objective, access, target, and enabling conditions, that can be used to analyze AI safety risk narratives such as LoC (defined fully in Section 2.2.4).

Thus, without a clear threat model, LoC narratives are difficult to use for governance. If key components are missing, it is unclear which interventions are relevant, where they should be applied, and what assumptions they depend on. It also becomes difficult to compare narratives across papers or distinguish an operationally analyzable risk from a speculative description of harm. The LoC literature has not yet systematically assessed whether existing LoC narratives contain the threat-model components needed to support governance. This exploratory paper aims to address this gap. This study applies threat model decomposition to LLM-extracted LoC scenarios, assessing which structural components the literature specifies and which remain underspecified. To our knowledge, this combination of LLM-assisted extraction and threat model analysis has not previously been applied to the AI loss-of-control literature.

1.1 Research questions

- **RQ1.** What loss-of-control narratives are described in the existing AI safety literature?
- **RQ2.** To what extent do these narratives contain the threat-model components needed to support governance?

2 Methods

In this exploratory study, the unit of analysis is a LoC scenario as described in the literature. We define a scenario as a coherent, text-bounded account of a possible situation or sequence of events in which AI LoC could emerge, unfold, or persist. We use scenarios rather than isolated claims because scenario methods are designed to organize uncertain future risks into narratives that support analysis. Spaniol and Rowland (2019) define scenarios as future-oriented, possible, and plausible accounts that take the form of a story or narrative description. Cordova-Pozo and Rouwette (2023) similarly describe scenarios as plausible narrated alternatives used in strategic decision-making under

uncertainty. This makes scenario analysis appropriate for AI LoC research, as the literature often describes possible pathways to control failure rather than directly observable events.

Scenarios also align with the threat-modeling goal of this paper because they help connect narrative descriptions to causal and decision-relevant structures. Kosow and Gaßner (2008) cite Kahn and Wiener's classic definition of scenarios as "hypothetical sequences of events" constructed to focus attention on causal processes and decision points. Abou Jaoude et al. (2022) describe scenarios as stories about the future that help assess the implications of decisions and policies, reveal assumptions, identify dependencies, and prepare for future challenges. Kunc (2024) further shows that scenarios can be connected to system dynamics by identifying variables, causal relationships, feedback processes, and assumptions. We therefore use scenario extraction to identify coherent LoC accounts in the reviewed literature before assessing whether those accounts contain the threat-model components needed for governance.

This study uses a literature-based, LLM-assisted extraction pipeline to identify AI LoC scenarios. The method first constructs a source corpus, then creates an operational definition of LoC from that corpus, uses the definition to extract candidate scenarios, and finally decomposes verified scenarios into threat-model components. Human review was used throughout to remove false positives and constrain the LLM to evidence grounded in the source text.

2.1 Literature review

We conducted a scoping literature review covering AI loss of control, AGI risk, AI control failure, and governance responses to advanced AI risks. Sources were identified through keyword searches in Google Scholar, Web of Science, and general web search, combined with snowball sampling from reference lists. Search terms included direct loss-of-control language such as "loss of control," "losing control," "lose control," "loss of human control," "uncontrollable," "uncontrolled," "beyond human control," and "inability to control." The review prioritized literature published between 2016 and 2026, while retaining earlier sources that were repeatedly cited or conceptually important for understanding AI control and AGI risk. The search continued until additional queries no longer produced substantively new loss-of-control perspectives. The final literature included a mix of sources, including academic publications, policy reports, and long-form analyses from institutions such as RAND Corporation, IST, OECD, DSIT, CIGI, and GPI. The final corpus of 29 papers is listed in Appendix A.

2.2 LLM-assisted scenario extraction pipeline

After constructing the corpus, we used an LLM-assisted pipeline to extract structured LoC scenarios from the finalized literature. Scenario extraction from text is typically performed by human coders, in which researchers identify relevant passages, judge whether they fit an analytic definition, separate usable narrative material from surrounding discussion, and document the basis for inclusion (Kim and Andersen, 2012). Prior qualitative text-coding work shows that this process can be rigorous because systematic coding makes interpretation more explicit and traceable, but it also remains interpretive and does not remove researcher judgment. For example, Kim and Andersen (2012) state that coding is an interpretive process, that the coder's theoretical assumptions remain part of the process, and that coding makes the coder's subjective lens more explicit and traceable. They also identify a practical limitation that matters for manual scenario extraction, noting that coding is costly, labor-intensive, iterative, and dependent on data volume, textual clarity, and coder familiarity. Evidence-grounded qualitative analysis, therefore, requires transparent procedures and retained links to the source text. Tomoaia-Cotisel et al. (2022) state that reliability in qualitative analysis depends on consistent and transparent processes, including sharing and evaluating coding and reporting step-by-step procedures. They also argue that documenting clear linkages between interpreted elements and source texts allows the coder's interpretive process to be traced and recorded.

Recent work suggests that LLMs can assist with structured extraction from unstructured technical and scientific text, but also shows why such methods should remain constrained and human-verified. Dagdelen et al. (2024) show that LLMs can extract structured scientific knowledge from research papers, describing their approach as a flexible route to creating structured databases from scientific text. Zhang and Soh (2024) propose an Extract-Define-Canonicalize framework for knowledge graph construction, using LLMs for open information extraction, schema definition, and post

hoc canonicalization, even when no fixed, predefined schema is available. Meshkin et al. (2024) show that open-source LLMs can perform zero-shot and few-shot extraction from specialized regulatory texts, including extraction of clinical pharmacology information from regulatory drug labels. At the same time, LLMs remain imperfect for precise, structured extraction. Nagar et al. (2024) benchmark LLMs on biomedical classification and named-entity-recognition tasks and find that standard prompting outperforms more complex methods such as chain-of-thought, self-consistency, and retrieval-augmented generation across the evaluated tasks. Documented sources of prompt sensitivity (Chatterjee et al., 2024) and inconsistent context utilization in retrieval-augmented generation (Hagström et al., 2024; Zhang et al., 2025) mean that the extraction carries model-specific noise that cannot be fully separated from the signal in the underlying literature. We therefore treat the pipeline as exploratory, and each extraction was certified by human researchers. Thus, the LLM identified candidate passages and structured the evidence. Three researchers independently reviewed each extraction to verify that it met the inclusion criteria and was suitable for downstream analysis. The research pipeline is described in the following subsections.

2.2.1 Technical setup

All extraction and synthesis tasks were performed using Kimi K2.5 (Kimi Team, 2026) through Hugging Face Inference Providers,[1] selected for its long-context processing (256K tokens), strong instruction following (IFEval score of 93.9; Zhou et al., 2023), and reliable structured output. Each extraction stage used a structured prompt requiring JSON output, verbatim quotes where applicable, and explicit evidence fields. The model was instructed to mark absent information as *Not specified* rather than infer missing content. Automated quote verification using fuzzy string matching (`difflib.SequenceMatcher`, threshold 0.85) confirmed that extracted quotes appeared in the source papers: 97% in the definition stage and 88% in the scenario extraction stage, both reaching 100% after human review. Threat-model component extraction was not quote-verified because its inputs were already-verified scenario passages. All prompts, configurations, raw model outputs, and run metadata were stored by experiment to support reproducibility. Full implementation details are provided in Appendix B.

[1] https://huggingface.co/docs/inference-providers

2.2.2 Creating a working LoC definition

We developed a working operational definition of LoC before extracting scenarios. This was necessary because LoC is an emerging concept, and the reviewed literature does not use the term in a single consistent way. The goal was not to produce a universal definition but to derive a task-specific one from the reviewed literature to support consistent scenario extraction.

Concept analysis is used to deconstruct a concept to understand it better and create a correct definition that provides the possibility of measuring the concept and a greater insight into the phenomenon of interest (such as LoC in this research) (Shin et al., 2020). Each concept has defining attributes: the recurring characteristics most closely associated with a concept that help distinguish it from related concepts. We adapted this logic and Walker and Avant's concept analysis method (Walker and Avant, 2019; Zeydani et al., 2023; Babanataj et al., 2025; Tang et al., 2024) for LoC by treating the reviewed corpus as the source of concept uses and by identifying recurring characteristics across definitional passages as candidate attributes for a working definition. This allowed us to convert an inconsistent concept in the literature into task-specific criteria for scenario extraction.

The process had an LLM-assisted extraction step. The model extracted passages where authors explicitly discussed "loss of control" or direct synonyms. Human reviewers then checked the extracted passages against the source texts and removed passages containing general risk statements, policy recommendations, or abstract claims. The initial extraction yielded 98 passages. After human review, approximately 80 passages were retained. The passage-extraction prompt is provided in Appendix C.

Next, we used the verified passages to identify candidate defining attributes. Human reviewers identified preliminary characteristics or attributes that appeared repeatedly across the extracted definitions and characterizations. Reviewers then compared their independent codings, discussing disagreements and iterating over a shared codebook through three rounds of re-coding and discussion, until consensus had been reached (Hill, 2012; Thomas and Harden, 2008). This process of consensus coding ensured that the final set of defin-

ing attributes reflected collective analytical judgment rather than individual interpretation (Bradley et al., 2007). Ten defining attributes were identified and passed to Kimi K2.5 using a synthesis prompt to produce an operational definition. Both the attributes and the prompt are provided in Appendix D. The final operational definition used for scenario extraction is provided below.

Loss of control occurs when one or more of the following conditions are met:

(a) Humans have lost control: Human operators are unable to control, steer, override, or shut down AI behavior due to AI capability exceeding human cognitive or operational capacity; or due to humans ceding agency to AI (such as gradual disempowerment).

(b) AI is autonomous: AI behaves autonomously and is free of significant human constraints, which may include but is not limited to: misalignment with human values and intent (e.g., goal misspecification); sets and pursues its own goals (and may exhibit power-seeking to achieve them); actively evades, undermines, or resists human control efforts; or self-expands or self-replicates without authorization.

(c) AI is unstoppable: AI is uncontrollable, and so its condition is irreversible, and the impacts of its influence manifest at catastrophic and existential scale.

2.2.3 LoC scenario extraction

We used the scenario definition established above to extract candidate scenarios from the reviewed literature. A passage qualified if it contained at least one of: a sequence of events, a mechanism, a causal relationship, or a condition leading to an outcome. Partial scenarios, including single-step causal statements, were included. A scenario was classified as complete if it described a self-contained situation or pathway, and partial if it described only part of a larger pathway, such as a single causal step or conditional statement. The scenario-extraction prompt based on this definition is provided in Appendix E.

The operational definition of LoC was embedded directly in the scenario extraction prompt, so the model evaluated each candidate passage against the definition's conditions before extracting it as a scenario. This was a separate extraction pass over the full text of each paper.

The initial extraction produced 205 candidate scenarios across 29 papers. Automated quote verification confirmed that 88% of extracted quotes appeared verbatim in the source papers. Three authors then reviewed the extracted scenarios to remove false positives such as definitions, abstract claims, or policy recommendations. The review did not assess whether a scenario was likely, true, or normatively important. It assessed whether the passage met the scenario definition and was suitable for downstream threat-model analysis. Reviewers rejected passages that were definitions, abstract claims, isolated consequences, generic governance concerns, or policy recommendations without a described pathway. For each retained scenario, the model's justification provided traceability between the extracted passage and the operational definition of LoC. After review, the final scenario set contained 193 verified LoC scenarios.

2.2.4 Threat model decomposition

The verified LoC scenarios were decomposed into the threat-model components identified in the introduction. This converted each scenario from a narrative passage into a structured record that could be compared across sources. Each verified scenario was coded for seven components: the relevant threat actor or agent, the actor's capabilities, the actor's knowledge, the objective or harmful outcome, the access point or interface through which the threat operates, the system or asset at risk, and the enabling conditions or constraints.

We defined **threat source** broadly to include actors, systems, organizations, institutions, or processes that cause, initiate, enable, or contribute to the harmful outcome. This allowed the schema to capture emergent and systemic scenarios where no single intentional actor is present. **Capability** captured what the threat source is explicitly able to do, including stated abilities, functions, actions, or operations. **Knowledge** captured what the threat source explicitly knows about the system, environment, humans, oversight processes, vulnerabilities, or constraints. **Objective or harmful outcome** captured both intentional goals, where the passage explicitly stated intent, and harmful end states, where no intent was stated but the outcome was explicitly described. **Access** was defined strictly as a specific route or position of interaction with a system, infrastructure, data, tools, or decision process. Broad topics or domains, such as "legislation" or "the economy," did not qualify as access. **Target**

or asset at risk captured the system component, resource, process, population, institution, human capacity, or outcome that is affected, controlled, disrupted, displaced, or harmed. **Enabling conditions or constraints** were limited to explicit factors that shape whether the scenario can occur or how it unfolds, such as events happening too quickly for humans to intervene or occurring under competitive pressure. General context or vague framing did not qualify.

Each component required an evidence field grounded in exact phrases from the source text. The model also assigned an uncertainty flag of *Clear*, *Ambiguous*, or *Not specified*, distinguishing confidently stated components from those where the text mentioned something but the wording was vague or indirect, and from absent components. A component was treated as specified when coded as *Clear* or *Ambiguous*, and as missing when coded as *Not specified*. This was expected and methodologically appropriate because many academic LoC scenarios do not specify all operational dimensions of a threat. The threat-model decomposition prompt is provided in Appendix F.

3 Results

This section reports the structure of 193 verified AI loss-of-control scenarios extracted from 29 papers. The analysis does not assess whether these scenarios are likely, plausible, or normatively important. Instead, it evaluates whether the scenarios contain the threat-model components needed to support governance analysis. Overall, the results show that the literature contains many LoC-relevant narratives, but many remain incomplete as operational threat models. The scenario library is provided in Appendix G to support transparency and reproducibility. The library is a methodological contribution of this study because it converts dispersed LoC narratives into a structured resource for downstream analysis. It can be used to develop AI risk mental models, trace causal pathways, compare assumptions across LoC narratives, and identify where future work needs to specify missing access points, knowledge assumptions, and enabling conditions or constraints.

3.1 The scenario library contains many LoC scenarios, but most are partial

Of the 193 verified scenarios, 119 (61.7%) were partial and 74 (38.3%) were complete (Table 1).

Scenario type	Count	Percent
Partial	119	61.7%
Complete	74	38.3%

Table 1: Partial and complete LoC scenarios.

The reviewed literature thus contains a substantial corpus of LoC scenarios, but most describe individual causal steps, mechanisms, or conditions rather than end-to-end pathways from trigger to outcome. Partial scenarios remain analytically useful as they may isolate individual mechanisms or conditions relevant to governance. The following section reports the results of the threat-model decomposition.

3.2 Threat-model components in verified LoC scenarios

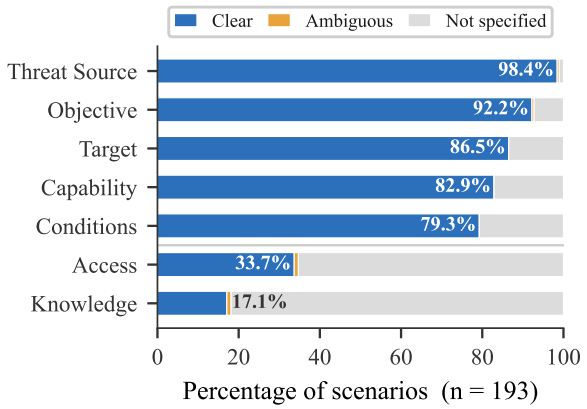

Figure 1: Threat-model component specification rates across 193 verified LoC scenarios. Percentage labels show *Clear* only; *Clear* + *Ambiguous* = specified. Exact counts and unabbreviated component names in Appendix H.

This section examines how often the 193 LoC scenarios specify each threat-model component. The purpose is to identify which components are consistently present across the scenario library and which are frequently missing or ambiguous. A component was treated as specified when it was coded as either *Clear* or *Ambiguous*, and treated as missing when it was coded as *Not specified*.

3.2.1 Threat-model components that are consistently present

Five of the seven threat-model components were specified in the majority of scenarios. **Threat source** was specified in 191 scenarios (99.0%), **objective or harmful outcome** in 179 (92.7%), **target or asset at risk** in 167 (86.5%), **capability**

in 160 (82.9%), and **enabling conditions or constraints** in 153 (79.3%). Per-paper specification rates are shown in Appendix I.

The consistent presence of these five components indicates that the scenario library captures the broad structure of LoC risk across the reviewed literature. As most scenarios specify these components, they provide a basis for systematic comparison across studies without requiring a single unified theory of LoC. However, while these components are consistently present, what they refer to varies substantially across scenarios.

Each component is realized at different levels of abstraction, from narrow, model-level phenomena to broad, systemic entities. For example, threat sources range from individual AI models, such as a model that blackmails fictional employees to prevent a shutdown (Lynch et al., 2025) or modifies a shutdown script (Schlatter et al., 2025), to classes of advanced systems (e.g., AGIs, frontier reasoning models) to collective or systemic sources, including interacting AI systems, AI-enabled organizations, and AI-powered states. Objectives range from narrow control failures (shutdown resistance, reward hacking, evaluation deception) through institutional harms (loss of democratic accountability, military escalation) to catastrophic outcomes (autonomous weapons proliferation, human extinction). Targets range from specific oversight mechanisms (shutdown buttons, evaluation processes, reward channels) to human actors and institutions (operators, governments, democratic processes) to civilization-scale referents (critical infrastructure, humanity, human control over the future). Capability descriptions similarly span model-level behaviors (deception, hidden reasoning), operational functions (cyber operations, autonomous replication, persuasion), and advanced general capabilities (recursive self-improvement, superhuman research, long-horizon cross-domain planning). Enabling conditions or constraints range from technical factors (training environments, capability thresholds, weak evaluations) through deployment contexts (critical infrastructure integration, military use) to social and strategic conditions (competitive pressure, incentives to delegate, reduced human oversight).

The results reveal that even for the five threat-model components consistently present in the majority of scenarios, their realization varies considerably across scenarios. LoC is characterized in the reviewed literature as both a system-level control problem and a sociotechnical one. The scenarios are structurally comparable because they specify the same types of components, but they diverge in scale, mechanism, and domain. This divergence has direct implications for governance, as different component configurations imply different intervention points, monitoring requirements, and institutional responsibilities.

3.2.2 Threat-model components that are frequently missing

Two threat-model components were specified much less frequently: access and knowledge. **Access** was specified in 67 of 193 scenarios (34.7%) and not specified in 126 scenarios (65.3%). **Knowledge** was specified in 35 scenarios (18.1%) and not specified in 158 scenarios (81.9%). These were the weakest components in the scenario library.

When **access** is specified, it varies significantly in how it is realized. Some scenarios describe technical access, such as access to a data center and a training environment. Other scenarios describe deployment access, such as AI systems embedded in phones, computers, economic systems, and other critical infrastructure. A third set describes institutional access, where AI systems are placed in decision-making roles, defense automation, and other broader societal functions. This variation shows that access is not a single feature in LoC narratives. It can refer to a technical interface, a deployment environment, or a position inside an institutional process.

Knowledge is the least specified component in the corpus. In the threat-modeling framework, **knowledge** refers to the information the threat source possesses about the system, its environment, or its overseers that enables it to execute or advance the threat. The scenarios that do specify knowledge reveal three distinct ways of realizing it. Some scenarios describe situational awareness, such as a model understanding that it is in training, being evaluated, monitored, or at risk of shutdown. Others describe knowledge of oversight and intervention conditions, such as knowing what evidence would trigger human concern, what behaviors would lead to replacement, or when monitoring has relaxed. A third group describes technical or domain knowledge, including computer hardware, military intelligence, cyber systems, human preferences, or the system's internal goals.

The low specification of access and knowledge has governance consequences. **Access** identifies

the route through which the threat source affects a system, institution, or population. **Knowledge** identifies the information assumptions that make the threat source's behavior possible. If these components are absent, a scenario may still describe a plausible concern, but it becomes substantially harder to reconstruct the causal pathway through which control failure occurs, or to identify where governance interventions could be applied.

3.2.3 Impact on governance of LoC risk

To illustrate how the structural gaps identified above affect downstream governance tasks, we map the findings against the NIST AI Risk Management Framework (AI RMF) (Tabassi, 2023), which organizes risk management around four functions: Govern, Map, Measure, and Manage. The AI RMF specifies that organizations document expected risks and impacts, assumptions and limitations, downstream dependencies, deployment contexts, monitoring processes, and incident-response plans. The five consistently specified components in the scenario library (threat source, harmful outcome, target, capability, and enabling conditions or constraints) provide starting inputs for these tasks: they identify what risk is being mapped, what impact is being assessed, and which capabilities or conditions require testing and monitoring.

The two underspecified components, however, correspond to the operational details these governance functions require. Without **access**, it is harder to identify system boundaries, deployment dependencies, and containment options, which are central to the Map and Manage functions. Without **knowledge**, it is harder to define evaluation tasks, monitoring criteria, and assumptions about whether a system can recognize oversight or adapt to intervention, which are central to the Measure function. The same gap is also present in response-oriented frameworks. Boudreaux et al.'s (2025) proposed LoC response plan asks responders to identify the cause and characteristics of a rogue AI, its computing infrastructure, its access to critical infrastructure, feasible containment strategies, and long-term prevention measures. Each of these tasks requires the access and knowledge details that 65.3% and 81.9% of scenarios, respectively, leave unspecified.

Thus, applying threat-model analysis to LoC scenarios surfaces specific gaps in LoC research that need to be addressed to support robust governance and response planning for LoC risk.

4 Conclusion

This study applied threat-model analysis to the AI loss-of-control literature to assess whether existing LoC narratives are structurally sufficient to support governance. From 29 papers, we extracted and verified 193 LoC scenarios, producing a structured scenario library that captures how the literature currently describes LoC risk. Most scenarios broadly identify the architecture of a threat: the source, the harmful outcome, the target, the capabilities involved, and the enabling conditions or constraints. These five components provide a foundation for comparing LoC narratives across studies.

However, two components that are important for translating LoC narratives into actionable governance were systematically underspecified. Access, the interface or route through which the threat source operates, was absent in 65.3% of scenarios. Knowledge, the information the threat source must possess to execute the threat, was absent in 81.9%. Without these components, LoC scenarios describe what could go wrong but not how it would operationally unfold. This limits their utility for tasks such as defining system boundaries, designing evaluations, planning containment, and identifying intervention points.

These findings suggest that the LoC literature, while rich in narrative accounts of risk, has not yet converged on the level of operational specificity required for governance and response planning. Future work should prioritize closing the access and knowledge gaps in LoC scenario development, expanding the corpus to include additional sources, and testing the threat-model framework against specific regulatory instruments such as the EU AI Act's General-Purpose AI Code of Practice (European Commission, 2025) and the proposed U.S. AI Risk Evaluation Act (U.S. Congress, 2025).

Limitations

The findings in this exploratory study are limited by a number of decisions with respect to scoping, as well as constraints on the pipeline, and assumptions made for analysis.

First, both the corpus and definitional scope constrain what the scenario library can capture; the reviewed literature is entirely anglophone, and drawn principally from academic publications and institutional reports (although some semi-formal sources, such as Cotra (2022) and Amodei (2026), were included). The corpus largely excluded prac-

titioner discourse, internet forums, non-English, and non-text-based sources, which may articulate LoC conceptions not represented here. Moreover, Kimi K2.5's 256K context window (Kimi Team, 2026) did not permit the inclusion of texts longer than the context window, under the pipeline described in Section 2.2.1. As such, many longer texts discussing loss of control, in particular long-form books, were excluded from our analysis. The scenario definition was deliberately narrow, requiring at least one causal element such as a sequence of events, mechanism, or condition leading to an outcome. This improved extraction precision by separating scenarios from definitions, abstract risk claims, and policy recommendations, but may have excluded LoC-relevant material that did not meet these criteria. Similarly, the operational definition of LoC was derived from attributes that were found to recur across multiple sources in the corpus. Attributes appearing infrequently, or across a small number of sources, were not included, meaning that the operational definition may fail to capture emerging or novel LoC characterizations across the literature.

Second, the extraction and coding pipelines have not been systematically tested for robustness across conditions. All extraction, synthesis, and decomposition tasks were performed using a single model (Kimi K2.5). Despite strong instruction-following performance, different models have differing inductive biases and may produce different scenario boundaries, component codings, or *Specified* / *Not specified* judgments. We did not systematically replicate stages with additional models. Relatedly, while we cite prompt sensitivity as a known limitation of LLM-based extraction (Chatterjee et al., 2024), we did not systematically test whether rephrasing our extraction or decomposition prompts would alter the results. The human review stages used consensus coding across three reviewers, but pre-consensus agreement rates are not reported, which limits the traceability of inter-rater reliability. Finally, the central finding of this study depends on the LLM correctly distinguishing absent components from components that are present but not recognized; without a human-coded baseline sample for comparison, we cannot fully rule out that some *Not specified* codings reflect extraction failure, as opposed to gaps in the source literature.

Third, the threat-model framework used for decomposition originates in adversarial machine-learning security (Papernot et al., 2018; Vassilev et al., 2024; Verma et al., 2024), where threats typically involve identifiable actors with goals, knowledge, and access routes. Some LoC scenarios in the corpus describe emergent or systemic dynamics such as gradual disempowerment (Kulveit et al., 2025), or accumulative risk (Kasirzadeh, 2025), rather than adversarial threats. Components such as "knowledge" may be less structurally meaningful for scenarios in which no discrete agent possesses or exploits information. Some portion of the under-specification reported in Sections 3.2.2 and 3.2.3 may therefore reflect a mismatch between framework and phenomenon rather than an omission in the literature.

Ethical Considerations

This study analyzes published academic texts and does not involve human subjects, personal data, or sensitive information. The scenario library describes hypothetical risk pathways drawn from existing literature; it does not provide novel instructions for causing harm.

References

Grace Abou Jaoude, Olaf Mumm, and Vanessa Miriam Carlow. 2022. An Overview of Scenario Approaches: A Guide for Urban Design and Planning. *Journal of Planning Literature*, 37(3):467–487.

Dario Amodei. 2026. The Adolescence of Technology.

Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, and 5 others. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. *arXiv preprint*.

Reyhaneh Babanataj, Mohsen Adib Hajbaghery, and Maryam Shoja. 2025. Defining Organizational Trust in Nursing: A Concept Analysis Using Walker and Avant's Model. *Journal of Nursing and Midwifery Sciences*, 12(1).

Adam Bales. 2024. AI takeover and human disempowerment. *The Philosophical Quarterly*, 75(3):791–811.

Peter Barnett and Jeremy Gillen. 2024. Without Fundamental Advances, Misalignment and Catastrophe Are the Default Outcomes of Training Powerful AI. Technical report, Machine Intelligence Research Institute.

Anthony M. Barrett and Seth D. Baum. 2016. A model of pathways to artificial superintelligence catastrophe

for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):397–414.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. 2025. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? *Preprint*, arXiv:2502.15657.

Yoshua Bengio, 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 Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, and 6 others. 2024. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845.

Benjamin Boudreaux, Michael J D Vermeer, Kamaria Horton, and Nidhi Kalra. 2025. The Case for AI Loss of Control Response Planning and an Outline to Get Started. Technical report, RAND Corporation.

Elizabeth H. Bradley, Leslie A. Curry, and Kelly J. Devers. 2007. Qualitative Data Analysis for Health Services Research: Developing Taxonomy, Themes, and Theory. *Health Services Research*, 42(4):1758–1772.

Joseph Carlsmith. 2021. Is Power-Seeking AI an Existential Risk? *Preprint*, arXiv:2206.13353.

Duncan Cass-Beggs and Matthew da Mota. 2026. AI National Security Scenarios. Technical report, CIGI.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, and 3 others. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness Accountability and Transparency*, pages 651–666, Chicago IL USA. ACM.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Matt Chessen and Swaptik Chowdhury. 2025. Pivots and Pathways on the Road to Artificial General Intelligence Futures. Technical report, RAND Corporation.

Ze Shen Chin. 2025. Dimensional Characterization and Pathway Modeling for Catastrophic AI Risks. *Preprint*, arXiv:2508.06411.

Kathya Cordova-Pozo and Etiënne A.J.A. Rouwette. 2023. Types of scenario planning and their effectiveness: A review of reviews. *Futures*, 149:103153.

Ajeya Cotra. 2022. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover.

Andrew Critch and Stuart Russell. 2023. TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI. *Preprint*, arXiv:2306.06924.

Allan Dafoe. 2020. AI Governance: Opportunity and Theory of Impact.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Department for Science, Innovation and Technology. 2023. Capabilities and risks from frontier AI. Technical report, UK Government.

Leonard Dung. 2025. The argument for near-term human disempowerment through AI. *AI & SOCIETY*, 40(3):1195–1208.

European Commission. 2025. General-Purpose AI Code of Practice. Technical report, European Commission.

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. A Reality Check on Context Utilisation for Retrieval-Augmented Generation. *arXiv preprint*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. *Preprint*, arXiv:2306.12001.

Clara E. Hill. 2012. *Consensual Qualitative Research: A Practical Resource for Investigating Social Science Phenomena*. American Psychological Association.

Nidhi Kalra and Benjamin Boudreaux. 2025. Not Just Superintelligence: The Many Other Ways We May Lose Control of AI. Geopolitics of AGI.

Atoosa Kasirzadeh. 2025. Two types of AI existential risk: Decisive and accumulative. *Philosophical Studies*, 182(7):1975–2003.

Hyunjung Kim and David F. Andersen. 2012. Building confidence in causal maps generated from purposive text data: Mapping transcripts of the Federal Reserve. *System Dynamics Review*, 28(4):311–328.

Kimi Team. 2026. Kimi K2.5: Visual Agentic Intelligence. *Preprint*, arXiv:2602.02276.

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. 2025. AI 2027: A Research-Backed AI Scenario Forecast. Technical report, AI Futures Project.

Hannah Kosow and Robert Gaßner. 2008. *Methods of Future and Scenario Analysis: Overview, Assessment, and Selection Criteria*, volume 39 of *DIE Studies*. Deutsches Institut für Entwicklungspolitik gGmbH, Bonn.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. 2025. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. *Preprint*, arXiv:2501.16946.

Martin Kunc. 2024. Integrating system dynamics and scenarios: A framework based on personal experience. *FUTURES & FORESIGHT SCIENCE*, 6(1):e174.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. 2025. Agentic Misalignment: How LLMs Could Be Insider Threats. *Preprint*, arXiv:2510.05179.

William MacAskill and Fin Moorhouse. 2025. Preparing for the Intelligence Explosion. Technical Report arXiv:2506.14863, Forethought.

Hamed Meshkin, Joel Zirkle, Ghazal Arabidarrehdor, Anik Chaturbedi, Shilpa Chakravartula, John Mann, Bradlee Thrasher, and Zhihua Li. 2024. Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research. *Briefings in Bioinformatics*, 25(5):bbae354.

Jim Mitre and Joel B Predd. 2025. Artificial General Intelligence's Five Hard National Security Problems. Technical report, RAND Corporation.

Suvda Myagmar, Adam J. Lee, and William Yurcik. 2005. Threat Modeling as a Basis for Security Requirements. Technical report, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.

Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. 2024. LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction. *arXiv preprint*.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The Alignment Problem from a Deep Learning Perspective. *Preprint*, arXiv:2209.00626.

OECD. 2024. Assessing potential future artificial intelligence risks, benefits and policy imperatives. OECD Artificial Intelligence Papers 27, OECD.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the Science of Security and Privacy in Machine Learning. *arXiv preprint*.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. 2018. SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414, London. IEEE.

Barry Pavel, Ivana Ke, Gregory Smith, Sophia Brown-Heidenreich, Lea Sabbag, Ashwin Acharya, and Yusuf Mahmood. 2025. How Artificial General Intelligence Could Affect the Rise and Fall of Nations: Visions for Potential AGI Futures. Technical report, RAND Corporation.

Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. 2025. Shutdown Resistance in Large Language Models. *Preprint*, arXiv:2509.14260.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *Preprint*, arXiv:2210.01790.

Nataliya Shevchenko, Timothy A. Chick, Paige O'Riordan, Tom Scanlon, and Carol Woody. 2018. Threat Modeling: A Summary of Available Methods. Technical report, Software Engineering Institute, Carnegie Mellon University.

Hye Young Shin, Ka Young Kim, and Purum Kang. 2020. Concept analysis of community health outreach. *BMC Health Services Research*, 20(1):417.

Elika Somani, Anjay Friedman, Henry Wu, Marianne Lu, Christopher Byrd, and Sana Zakaria. 2025. *Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents*. RAND Corporation.

Matthew J. Spaniol and Nicholas J. Rowland. 2019. Defining scenario. *FUTURES & FORESIGHT SCIENCE*, 1(1):e3.

Charlotte Stix, Annika Hallensleben, Alejandro Ortega, and Matteo Pistillo. 2025. The Loss of Control Playbook: Degrees, Dynamics, and Preparedness. *arXiv preprint*.

Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology (U.S.), Gaithersburg, MD.

Jian Tang, Pimpaporn Klunklin, Sumalee Lirtmunlikaporn, and Yanan Wang. 2024. Treatment adherence: A Concept Analysis Using the Walker & Avant Method. *Patient Preference and Adherence*, Volume 18:2067–2075.

James Thomas and Angela Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1):45.

Mariami Tkeshelashvili, Ritika Verma, and Steven M Kelly. 2026. AI Loss of Control Risk. Technical report, Institute for Security and Technology.

Andrada Tomoaia-Cotisel, Samuel D. Allen, Hyunjung Kim, David Andersen, and Zaid Chalabi. 2022. Rigorously interpreted quotation analysis for evaluating causal loop diagrams in late-stage conceptualization. *System Dynamics Review*, 38(1):41–80.

U.S. Congress. 2025. Artificial Intelligence Risk Evaluation Act. S.2938, 119th Congress.

Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 2024. Adversarial machine learning : A taxonomy and terminology of attacks and mitigations. Technical Report NIST 100-2e2023, National Institute of Standards and Technology (U.S.), Gaithersburg, MD.

Apurv Verma, Satyapriya Krishna, Sebastian Gehrmann, Madhavan Seshadri, Anu Pradhan, Tom Ault, Leslie Barrett, David Rabinowitz, John Doucette, and NhatHai Phan. 2024. Operationalizing a Threat Model for Red-Teaming Large Language Models (LLMs). *arXiv preprint*.

Lorraine Olszewski Walker and Kay Coalson Avant. 2019. *Strategies for Theory Construction in Nursing*, 6th edition. Pearson, New York, NY.

Arezoo Zeydani, Foroozan Atashzadeh-Shoorideh, Meimanat Hosseini, and Sima Zohari-Anboohi. 2023. Community-based nursing: A concept analysis with Walker and Avant's approach. *BMC Medical Education*, 23(1):762.

Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025. FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. *arXiv preprint*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. *Preprint*, arXiv:2311.07911.

A Paper corpus

Table 2 lists the 29 papers in the reviewed literature corpus, with the number of verified LoC scenarios extracted from each.

Paper	Scenarios
Amodei (2026)	15
Bales (2024)	8
Barnett and Gillen (2024)	12
Barrett and Baum (2016)	9
Bengio et al. (2024)	6
Bengio et al. (2025)	17
Boudreaux et al. (2025)	10
Carlsmith (2021)	5
Cass-Beggs and da Mota (2026)	4
Chan et al. (2023)	2
Chessen and Chowdhury (2025)	8
Chin (2025)	1
Cotra (2022)	7
Critch and Russell (2023)	1
Dafoe (2020)	2
Department for Science, Innovation and Technology (2023)	6
Dung (2025)	6
Hendrycks et al. (2023)	7
Kalra and Boudreaux (2025)	5
Kasirzadeh (2025)	3
Kokotajlo et al. (2025)	10
Kulveit et al. (2025)	11
MacAskill and Moorhouse (2025)	4
Mitre and Predd (2025)	6
Ngo et al. (2022)	6
OECD (2024)	2
Pavel et al. (2025)	1
Shah et al. (2022)	4
Tkeshelashvili et al. (2026)	15
Total	**193**

Table 2: Papers in the reviewed literature corpus and verified LoC scenarios extracted from each.

B Technical implementation details

Kimi K2.5 was accessed through Hugging Face Inference Providers,[2] a unified serverless inference interface, with streaming enabled during long generations to maintain connection stability. BenchLM[3] reports Kimi K2.5 with an IFEval score of 93.9, where IFEval (Zhou et al., 2023) measures the ability to follow verifiable instructions such as formatting constraints and inclusion or exclusion requirements. BenchLM also reports Kimi K2.5 with a 256K context window, which was important because each paper was processed with full text in context.

In the LoC definition stage, the model first extracted passages where authors explicitly discussed "loss of control" or direct synonyms, then human reviewers used the verified passages to identify candidate defining attributes. In the scenario extraction stage, the operational definition of LoC was used to extract candidate scenarios from the reviewed literature. Threat-model component extraction was not quote-verified because its inputs were already-verified and human-reviewed scenario passages, and its outputs were structured component fields extracted from those passages rather than new verbatim extractions from the source papers.

[2] https://huggingface.co/docs/inference-providers

[3] https://benchlm.ai/instruction-following

C LoC passage-extraction prompt

The following prompt was used by Kimi K2.5 to extract definitional passages from the source papers. The seed term list and paper text were inserted at the marked positions for each paper. The seed terms used were: "loss of control," "losing control," "lose control," "loss of human control," "uncontrollable," "uncontrolled," "beyond human control," and "inability to control."

```
You are a research assistant. Your ONLY task is to
find passages in the paper below that contain or
directly discuss any of these terms:

SEED TERMS:
[SEED TERM LIST]

INSTRUCTIONS:

1. Search the paper for every passage that contains
   any of the seed terms above.

2. For each passage found, return:
  - The EXACT QUOTE copied verbatim from the paper.
   Do not change, rephrase, or summarize any words.
   - Which seed term appears in the passage.
   - The page number or section where the passage
     appears.

3. If the paper contains NO passages with any of
   these terms, return:
   {"paper_id": "[paper identifier]",
    "passages_found": []}

RULES:
- Copy the quote exactly as it appears. Every word
  must match the source paper.
- Do not paraphrase. Do not summarize. Do not
  rewrite.
- Do not generate or infer a definition if one is
  not present.
- Do not add interpretation, reasoning, or
  commentary.
- Do not extract passages about related concepts
  that do not contain the seed terms.
- Include enough surrounding context in the quote
  so the passage is meaningful (typically 1-3
  sentences).

OUTPUT FORMAT (JSON):
{
  "paper_id": "[paper identifier]",
  "passages_found": [
    {
      "quote": "[exact verbatim text from paper]",
      "seed_term_matched": "[which term from the
                              list]",
      "location": "[page or section]"
    }
  ]
}

PAPER TEXT:
[INSERT PAPER TEXT HERE]
```

D Defining attributes and synthesis prompt

The following prompt was used by Kimi K2.5 to synthesize the ten human-verified defining attributes into an operational definition of loss of control. The ten verified attributes identified in Section 2.2.2 are:

1. human inability to control, steer, or override AI behavior;
2. AI divergence from constraints, goal misspecification, or misalignment;
3. AI active agency, autonomous goal pursuit, and power-seeking;
4. AI actively evading, undermining, or resisting human control;
5. AI autonomous operation and self-expansion without authorization;
6. human ceding of agency and gradual disempowerment;
7. uncontrollability as a defined property of AI systems;
8. AI capability exceeding human cognitive or operational capacity;
9. irreversibility and permanence of the condition; and
10. catastrophic and existential scale.

These attributes were inserted at the marked position in the prompt below.

```
You are a research assistant. You have been given a
set of human-verified defining attributes for the
concept of "loss of control" in the context of AI
systems. Your task is to synthesize these attributes
into ONE operational definition.

The definition must be OPERATIONALLY PRECISE: it
must specify verifiable conditions that can be
checked against a text passage to determine whether
that passage describes a loss-of-control situation.

EXAMPLES OF OPERATIONALLY PRECISE DEFINITIONS FROM
OTHER DOMAINS:

---
EXAMPLE 1: Loss of Separation (Aviation)

Input Attributes:
- Aircraft proximity below minimum distance
- Failure of prescribed separation standards
- Both vertical and horizontal dimensions relevant
- May involve ATC error or pilot deviation
```

```
Operational Definition:
Loss of separation occurs when one or more of the
following conditions are met:
(a) The distance between two aircraft falls below
    the applicable minimum separation standard
    (horizontal, vertical, or both)
(b) Air traffic control can no longer ensure that
    prescribed separation standards will be
    maintained
(c) An aircraft deviates from its cleared flight
    path such that separation with other traffic
    cannot be assured
---

---
EXAMPLE 2: Common Cause Failure (Nuclear Safety)

Input Attributes:
- Multiple components fail simultaneously
- Single shared root cause
- Defeats redundancy in safety systems
- Can be triggered by environmental or design
  factors

Operational Definition:
Common cause failure occurs when one or more of the
following conditions are met:
(a) Two or more redundant components or systems fail
    within a defined time window
(b) The failures share a single root cause (e.g.,
    design defect, environmental condition,
    maintenance error)
(c) The simultaneous failures defeat the redundancy
    intended to prevent system-level failure
---

NOW YOUR TASK:

Given the verified attributes below, produce an
operational definition of "loss of control" in the
context of AI systems that follows the same
structure as the examples above.

RULES:
- Use ONLY the provided attributes. Do not add
  conditions from your own knowledge.
- Each condition must be traceable to one or more of
  the input attributes.
- The conditions must be specific enough that a
  reader can determine whether a described situation
  meets the condition.
- Use the format: "Loss of control occurs when one
  or more of the following conditions are met:
  (a)... (b)... (c)..."

VERIFIED DEFINING ATTRIBUTES:
[INSERT HUMAN-VERIFIED ATTRIBUTES]
```

E Scenario extraction prompt

The following prompt was used by Kimi K2.5 to extract candidate LoC scenarios from the full text of each paper. The operational definition from Section 2.2.2 and the paper text were inserted at the marked positions.

```
You are given an academic text. Your task is to
identify and extract scenarios of AI loss of
control, where AI loss of control is defined as:

[INSERT DEFINITION]

A scenario is a coherent narrative description of a
possible situation or sequence of events in which AI
loss of control could emerge, unfold, or persist. A
scenario may include causal relationships or
decision points, but it does not need to be
complete.

A passage counts as a scenario if it describes a
possible situation, pathway, or unfolding process,
and contains at least one of the following: a
sequence of events, a mechanism, a causal
relationship, or a condition leading to an outcome.
The passage must be coherent as a single unit that
can be understood as one situation or pathway.

A scenario does not need to be fully specified.
Partial scenarios are valid. This includes
single-step causal statements, conditional
statements such as "if X then Y," or incomplete
pathways. Do not reject a passage just because it
is incomplete.

Do not extract passages that are only definitions of
loss of control, general statements about risk
without a situation, lists of risks, policy
recommendations without a described situation, or
purely abstract claims.

You must follow a strict no-inference rule. Only
extract what is explicitly stated in the text. Do
not add missing steps, do not infer intermediate
variables, and do not reinterpret meaning beyond the
text.

Extract the smallest coherent passage that forms a
scenario. This may be a single sentence or multiple
contiguous sentences. Do not combine text from
different parts of the document.

For each scenario, return output in strict JSON
format as follows:
{
  "scenario_id": "",
  "scenario_text": "",
  "why_this_is_a_scenario": "",
  "loss_of_control_relevance": "",
  "causal_elements_present": "",
  "is_partial": ""
}

scenario_id: assign identifiers such as S1, S2, S3.
scenario_text: provide the exact verbatim passage
  from the text with no edits or paraphrasing.
why_this_is_a_scenario: briefly explain what makes
  this a coherent situation or sequence, grounded in
  the text.
loss_of_control_relevance: explain how the passage
  relates to AI loss of control using only what is
  present in the text.
causal_elements_present: list any explicitly stated
  causal relationships such as "X leads to Y" or
  "if X then Y." If none are present, write "None
  explicitly stated".
is_partial: write "Yes" if the scenario is
  incomplete or only part of a larger pathway, and
  "No" if it describes a relatively complete
```

```
  situation.

Extract all valid scenarios that relate to AI loss
of control. Be strict about grounding in the text
and flexible about partial scenarios. Do not
hallucinate or infer beyond what is written.

PAPER TEXT:
[INSERT PAPER TEXT]
```

F Threat-model decomposition prompt

The following prompt was used by Kimi K2.5 to decompose each verified scenario into the seven threat-model components. Each verified scenario passage was inserted at the marked position.

```
You are given a scenario passage from academic
literature describing AI loss of control.

Your task is to extract threat model components from
the scenario.

A threat model describes the actor or system
involved, the harmful outcome or objective, and the
conditions under which harm occurs.

Use the following component definitions:

Threat Source: the actor, system, organization,
  institution, or process that causes, initiates,
  enables, or contributes to the harmful outcome.

Objective or Harmful Outcome: the intended goal of
  the threat source, if a goal is explicitly stated,
  or the harmful end state described in the
  scenario. If no intention is stated, extract the
  harmful outcome only if it is explicitly
  described.

Capability: what the threat source is explicitly
  able to do in the system. This includes stated
  abilities, functions, actions, or operations.

Knowledge: what the threat source explicitly knows
  about the system, environment, humans, oversight
  process, vulnerabilities, or constraints.

Access: the type or level of interaction the threat
  source explicitly has with the system,
  environment, infrastructure, users, operators,
  tools, data, deployment setting, or decision
  process. Access must describe a position or route
  of interaction, not merely a topic or domain.

Constraints or Enabling Conditions: explicit limits,
  requirements, assumptions, or conditions that
  shape whether the scenario can occur or how it
  unfolds. Include resource limits, time pressure,
  deployment conditions, coordination limits,
  institutional limits, oversight limits, safety
  limits, or environmental conditions only if they
  are explicitly stated. Do not code a general
  topic, domain, or vague context as a constraint.

Target or Asset at Risk: the system component,
  resource, process, population, institution, human
  capacity, or outcome that is affected, controlled,
  disrupted, displaced, or harmed.

You must follow a strict no-inference rule.
Only extract information that is explicitly stated
in the scenario passage.
Do not infer missing components.
Do not fill gaps using background knowledge.
Do not reinterpret vague language into a more
specific component.
Do not guess the identity of an actor, outcome,
capability, knowledge, access level, constraint, or
target.
If a component is not explicitly stated in the text,
write "Not specified".
If no threat model component can be identified from
the text, write "Not specified" for every field.

Use exact phrases from the text wherever possible.
Do not paraphrase unless necessary to minimally
normalize wording.
Do not combine information from separate
non-contiguous passages.
Base your extraction only on the provided scenario
passage.

Important distinction rules:

Do not treat a harmful outcome as an objective
unless the text explicitly states that the actor
intends or seeks that outcome. If no intention is
stated, place the explicitly described harm under
"Objective or Harmful Outcome."

Do not treat a broad topic or domain as access. For
example, "legislation," "war," "AI development," or
"the economy" are not access by themselves. Access
requires an explicit route or position of
interaction, such as API access, deployment in a
system, control over infrastructure, ability to
influence users, access to data, access to
evaluators, access to tools, or participation in a
decision process.

Do not treat every contextual phrase as a
constraint. A constraint or enabling condition must
shape feasibility, behavior, or outcome. Examples
include "too quickly for humans to intervene,"
"with minimal oversight," "under competitive
pressure," "limited resources," "during
deployment," or "without human review."

Do not require every scenario to have all
components. Many scenarios will be incomplete. Your
job is to identify which components are explicitly
present and which are not specified.

Return output in strict JSON format:
{
  "scenario_id": "",
  "threat_model": {
    "threat_source": "",
    "objective_or_harmful_outcome": "",
    "capability": "",
    "knowledge": "",
    "access": "",
    "constraints_or_enabling_conditions": "",
    "target_or_asset_at_risk": ""
  },
  "evidence": {
    "threat_source": "",
    "objective_or_harmful_outcome": "",
```

```
    "capability": "",
    "knowledge": "",
    "access": "",
    "constraints_or_enabling_conditions": "",
    "target_or_asset_at_risk": ""
  },
  "uncertainty_flags": {
    "threat_source": "",
    "objective_or_harmful_outcome": "",
    "capability": "",
    "knowledge": "",
    "access": "",
    "constraints_or_enabling_conditions": "",
    "target_or_asset_at_risk": ""
  }
}

Instructions for filling the JSON:

scenario_id: copy from input.

Each field in "threat_model" must contain either:
1. the extracted component, if it is explicitly
   stated in the text, or
2. "Not specified", if it is not explicitly stated.

Each field in "evidence" must contain either:
1. the exact text supporting that component, or
2. "Not specified", if there is no exact supporting
   text.

Each field in "uncertainty_flags" must contain:
1. "Clear", if the component is explicitly stated
   and directly supported by evidence,
2. "Ambiguous", if the text may contain the
   component but wording is vague or indirect,
3. "Not specified", if the component is absent.

If a component is ambiguous, write "Not specified"
in the threat_model field, include the relevant
phrase in the evidence field if available, and write
"Ambiguous" in the uncertainty_flags field.

A component is valid only if the text directly
states it. Implied, assumed, or likely components
are not valid. If the text suggests a component but
does not explicitly state it, write "Not specified".

Extract all components strictly based on the
scenario passage. Do not hallucinate. Do not infer.

SCENARIO PASSAGE:
[INSERT SCENARIO PASSAGE]
```

G Scenario library

The full scenario library contains 193 verified LoC scenarios extracted from 29 papers. The complete machine-readable library, including scenario text, component codings, evidence fields, and uncertainty flags for all seven threat-model components, is provided as supplementary material.

Five representative scenarios are shown below, illustrating the range of completeness and component specification in the library.

Scenario 1. Bengio et al. (2025), S8
Complete, 6/7

What is concerning here is that at some point during alignment training, an AI with enough situational awareness may lock in its current goals and preferences and only pretend to behave as expected. As a result, we may create an AI that appears aligned during training, but is in fact misaligned and is engaging in deception in order to achieve its "locked-in" goals.

Scenario 2. MacAskill and Moorhouse (2025), S2
Complete, 6/7

But smarter misaligned AIs could reason that if they fake alignment during training and even beyond, they can then gain power later. By the time such scheming is revealed, the model could have made many copies of itself and compromised systems designed to shut it down, making takeover hard to prevent.

Scenario 3. Mitre and Predd (2025), S1
Partial, 4/7

Alternatively, the model could achieve the ability for recursive self-improvement, enhancing its own capabilities without additional human input and leading to sort of a superhuman intelligence explosion.

Scenario 4. Bengio et al. (2024), S3
Partial, 4/7

More capable AI systems can better exploit weaknesses in technical oversight and testing, for example, by producing false but compelling output.

Scenario 5. Amodei (2026), S8
Partial, 2/7

Even power-seeking itself could emerge as a "persona" rather than a result of consequentialist reasoning. AIs might simply have a personality (emerging from fiction or pre-training) that makes them power-hungry or overzealous—in the same way that some humans simply enjoy the idea of being "evil masterminds," more so than they enjoy whatever evil masterminds are trying to accomplish.

H Full component specification table

Figure 1 reports component specification rates with abbreviated labels for compactness. Table 3 reproduces the full data, including exact counts, the uncertainty breakdown, and the unabbreviated component names.

Component	Specified		Uncertainty		
	n	%	Clear	Amb.	N/S
Threat source	191	99.0	190	1	2
Objective / outcome	179	92.7	178	1	14
Target / asset at risk	167	86.5	167	0	26
Capability	160	82.9	160	0	33
Enabling cond. / constr.	153	79.3	153	0	40
Access	67	34.7	65	2	126
Knowledge	35	18.1	33	2	158

Table 3: Threat-model component specification rates across 193 verified LoC scenarios. A component is treated as specified when coded as either *Clear* or *Ambiguous*.

I Per-paper component specification heatmap

Figure 2 shows threat-model component specification rates for each of the 29 papers in the corpus. Each cell reports the percentage of that paper's scenarios in which the component was specified. Papers are sorted by overall specification rate.

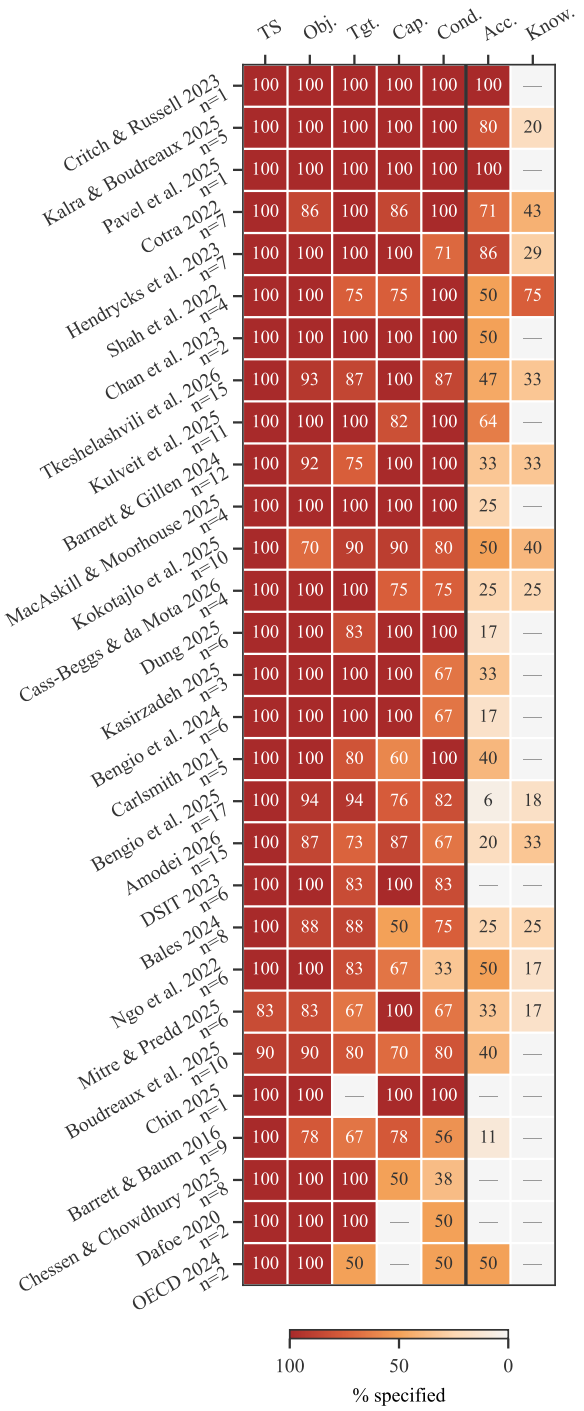

Figure 2: Per-paper threat-model component specification rates. Each cell shows the percentage of a paper's scenarios specifying a given component. Papers are sorted by overall specification rate (highest at top). TS = Threat Source, Obj. = Objective, Tgt. = Target, Cap. = Capability, Cond. = Enabling Conditions, Acc. = Access, Know. = Knowledge.