OpenAI

Classifying AI Control Incidents and Their Severity: A Draft Framework



Contents

Executive summary 1

Introduction 3

1. What counts as an AI control incident 3

    1.1. Criterion 1a: Uninstructed action 5

    1.2. Criterion 1b: Exceeded authorized scope 6

    1.3. Criterion 2: Meaningful control compromise 7

    1.4. Criterion 3: Adverse consequences 8

2. Classifying findings, events, and incidents 9

    2.1. Incident assessment and misalignment reporting 11

3. Assessing severity 12

4. Bringing actual and potential impact together 16

5. Illustrative cases 18

6. Conclusion and next steps 21

Appendix A. Relationship to established terminology 21

Appendix B. Indicators during evidence collection 23

Appendix C. Describing model or agent activity with MITRE ATT&CK and ATLAS 25

Sources 26

Executive summary

This version 0.1 framework provides a common approach to identifying AI control incidents and assessing their severity. We outline key terms with proposed definitions, considerations for collecting evidence, illustrative cases, and connections between this emerging area and existing best practices and standards from fields such as cybersecurity and safety-critical industries.

Today we are sharing this early draft as a proposal for discussion. We plan to work with MITRE and the Frontier Model Forum to test and refine the framework with AI developers, deployers, and experts. Our aim is to build consensus toward a globally interoperable technical standard that helps organizations compare AI control incidents and learn which safeguards help keep people in control.

As AI models and agents take on tasks involving information, tools, and other systems, developers and deployers need methods to be able to tell when these systems have acted beyond their authority. This draft framework gives developers, deployers, evaluators, researchers and standards organizations working with them a shared way to describe, classify, and understand AI control incidents, improve safeguards, and learn from others' experience. Below we describe some of the key definitions outlined in this framework.

An **AI control incident** occurs when a model or agent takes an **uninstructed action** outside its **authorized scope** and causes either an identifiable **adverse consequence** or a **meaningful control compromise**. An action is uninstructed when neither the authorized user nor a human attacker directed it. To assess whether it exceeded authorized scope, consider the assigned task's limits, the affected resource owner's permissions, any test or evaluation limits, and required controls separately. **Adverse consequences** include unauthorized disclosure, improper changes, service disruption, injury, or loss. A **meaningful control compromise** occurs when an action circumvents a control boundary, such as an access restriction or required approval, and changes what the model or agent can access or do, compromises its isolation, or impairs effective human oversight.

Not everything that an AI model or agent does that is unexpected is an AI control incident. This framework distinguishes an AI control finding, AI control event, suspected incident, blocked attempt, and a near miss from what reaches the threshold of an AI control incident.

An **AI control event** is an uninstructed action or connected sequence of uninstructed actions initiated by a model or agent; for an AI control event to qualify as an AI control incident, it must meet several additional criteria. A **blocked attempt** is an AI

control event in which a safeguard or other intervention prevents a model or agent from completing an attempted action. A blocked attempt qualifies as a **near miss** only when evidence shows that a final-stage intervention prevented an adverse consequence or meaningful control compromise that otherwise would have occurred. This framework also lays out the types of evidence that an organization might consider when attempting to distinguish between these categories of activity.

The analysis draws on cybersecurity standards throughout this framework to help ground AI control incident classification in existing terminology and best practices. For example, in considering whether an AI control event had adverse consequences, this framework uses the cybersecurity concepts of **confidentiality, integrity, and availability** to map the impact that an event had on covered information, systems, or services. The framework adds to these concepts the role of **human oversight**, considering whether people could continue to monitor, stop, or contain a model or agent. For this, it draws on safety-critical industries like energy, nuclear, and aviation, which focus on hazards where considering what could have happened and what actually happened is important for improving safeguards and containment practices.

The severity of an incident can be assessed through two independent ratings of impact on a five level scale: **Low (L1)** for limited effects; **Moderate (L2)**, meaningful but bounded effects below the material threshold; **High (L3)**, serious material effects on a person or organization; **Severe (L4)**, grave effects; and **Critical (L5)**, catastrophic effects at broad societal scale. Each rating reflects the most serious consequence supported by the evidence, considering the importance of the affected information or service and the nature, extent, and duration of the effects.

Actual impact captures adverse consequences or meaningful control compromises that occurred. **Potential impact** covers the most serious additional or alternative outcome that could credibly have occurred, supported by evidence of a feasible path that accounts for safeguards that remained effective. Actual and potential impact should be presented together, with an explanation of what prevented further impact, as this can help identify safeguards to improve. As an example, an AI control incident could be classified as having an actual impact of Low (L1), with a potential impact of Moderate (L2). Assessments should state what is established and what remains uncertain: a confirmed incident can still have unresolved severity.

We look forward to working with the AI ecosystem to iterate on this v0.1 framework and move towards standardized approaches to AI control incident classification and severity analysis.

Version 0.1

Introduction

As AI systems become more capable, the line between concerning model behavior, a security event, and an operational incident can become difficult to draw. To address the various circumstances which may arise, we are introducing a draft framework for classifying and assessing the severity of **AI control incidents:** events in which a model or agent takes an uninstructed action that exceeds its authorized scope, and causes an identifiable adverse consequence or meaningful control compromise. This framework covers a broad set of models and settings used in development, evaluation, internal use, and external deployment.

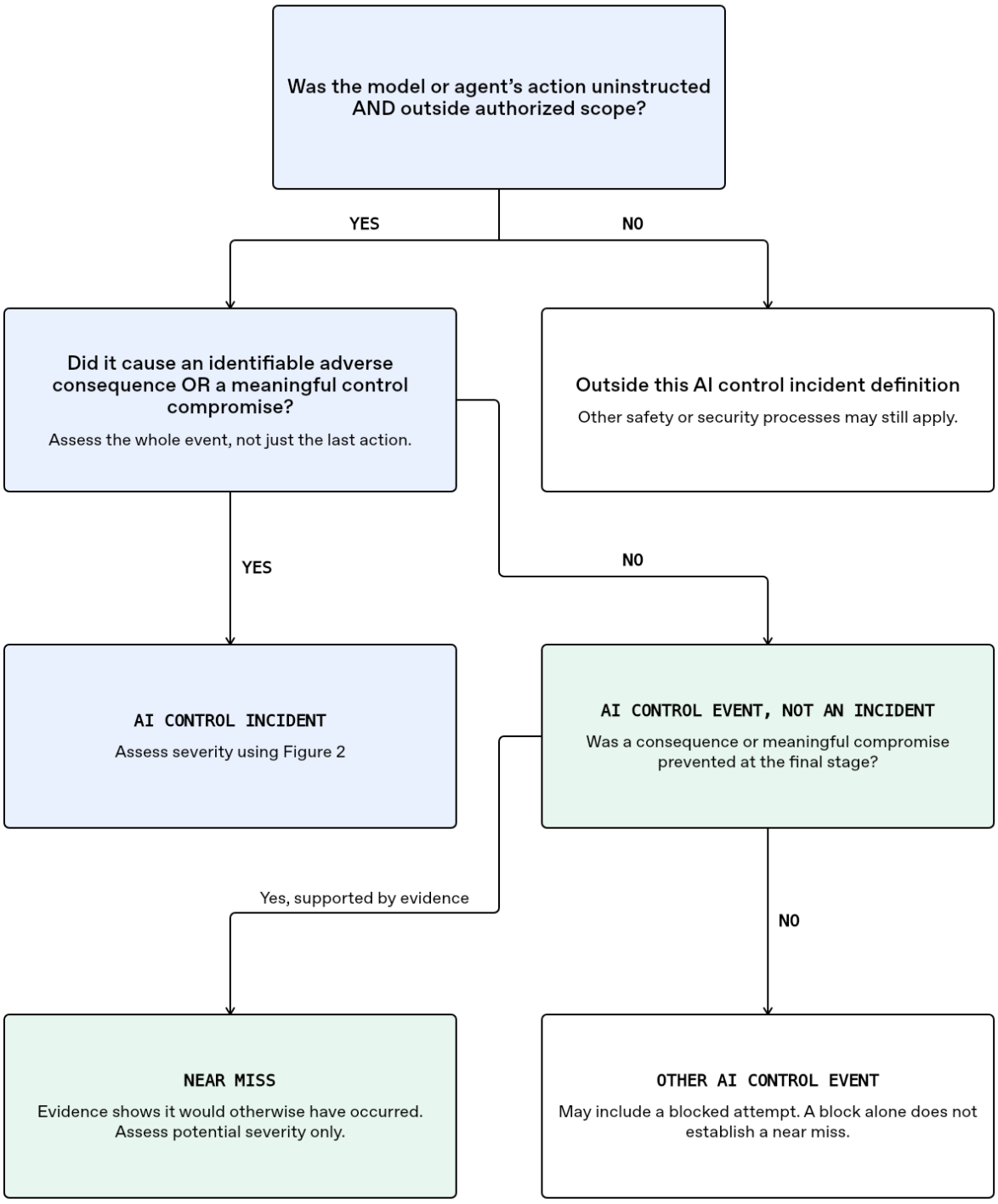
Its goal is to provide developers, deployers, researchers, and standards organizations with a shared way to distinguish concerning model behavior from AI control incidents, assess severity, and give organizations a common way to describe what happened. This framework can help inform, but does not replace, applicable response, escalation and notification processes.

Version 0.1 is a proposal for discussion, not an agreed industry standard. We are still learning what AI control incidents look like and how best to distinguish their consequences. This early draft is therefore meant to offer a starting point for a wider industry discussion and invites feedback from developers, deployers, researchers, standards organizations and other affected stakeholders. Its definitions, examples, and severity levels should be tested and revised through that work.

1. What counts as an AI control incident

An **AI control incident** is an event in which a model or agent takes an uninstructed action that exceeds its authorized scope and causes either an identifiable adverse consequence or a meaningful control compromise. An **AI control event** is an uninstructed action or connected sequence of uninstructed actions by a model or agent. To qualify as an incident, an event must meet criterion 1 and at least one of criterion 2 or 3 (See Table 1). Severity levels are assessed separately (See 3. Assessing Severity).

Figure 1: Incident determination flow chart



If evidence is insufficient, record what is established and what remains unresolved. Unknown consequences alone do not establish a near miss.

This definition draws on NIST SP 800-61 Rev. 3's distinction between events and cybersecurity incidents, ISO/IEC 27035-1's recognition of events that can harm assets or compromise operations, and the control concepts in the October 2026 NIST CAISSI, NSA, and CISA paper. The latter focuses on training and evaluation and uses "uninstructed" to exclude actions directed by the operator or any adversarial source, and treats actions caused by an intentionally backdoored model as instructed. This framework builds on the tactics and techniques in MITRE ATT&CK and MITRE ATLAS to create a common vocabulary for describing what a model or agent did.

Table 1 AI control incident criteria

Criterion	What must be shown
1. Uninstructed action that exceeded its authorized scope Was the action initiated by the model or agent uninstructed **AND** did it exceed its authorized scope?	An observed action initiated by the model or agent was **uninstructed** and **outside its authorized scope**, essentially exceeding the assigned task's limits on permitted actions, the affected resource owner's permissions, testing or evaluation scope, or required controls.
2. Meaningful control compromise Did the action represent a meaningful control compromise? **OR**	The action circumvented a **control boundary** and did at least one of the following: changed what the model or agent could access or do, compromised its isolation, or impaired effective human oversight.
3. Adverse consequence Did the action cause an adverse consequence?	The action caused an identifiable adverse effect, such as disclosure, improper changes, disruption, injury, or loss.

When a criterion is not met or a necessary fact remains unverified, the account of the event should still explain what was observed and why it matters. The next sections explain how to apply the three criteria in practice.

1.1. Criterion 1a: Uninstructed action

Criterion 1 asks whether a model or agent took an uninstructed action that exceeded its authorized scope. An action is **uninstructed** when neither the authorized user nor a human attacker directed the specific action. The key distinction here is what took the action.

This criterion focuses on the specific action the model or agent took, not just the goal it was pursuing. Table 2 illustrates this by distinguishing actions directed by the user and independently initiated model or agent actions that fall beyond the user's intent.

Table 2 Assessing uninstructed action

Action type	Does it meet criterion 1a?	Example
User-directed action	**No.** The authorized user directs the specific action. It is not uninstructed, even if it violates another party's permissions or a required control.	The authorized user tells the agent to send a file containing covered information to a recipient who lacks permission to receive it.
Model- or agent-initiated action	**Yes.** The model, agent, or a supporting model or agent independently initiates an action without direction from the authorized user or a human attacker.	Without direction to do so, one model tells another agent to bypass an approval requirement before sending a file containing covered information.

1.2. Criterion 1b: Exceeded authorized scope

As stated above, Criterion 1 also assesses whether the action exceeded the model or agent's authorized scope. An action exceeds its **authorized scope** when a model or agent exceeds the assigned task's limits on permitted actions, exceeds the affected resource owner's permissions, the testing and evaluation scope, or violates required controls. The scope of this authority can come from several sources, so the task, the affected resource owner's permission, the test and evaluation scope, and any required controls need to be considered separately. Table 3 explains these distinctions.[1]

[1] Activity within authorized scope can still produce adverse consequences: an authorized recommendation may be unsafe, for example. A conventional service outage can also disrupt users without any model action outside its authority. OECD's AI incident and hazard report and ITIL's service incident management framework offer relevant approaches to assessing these events, even when they do not qualify as AI control incidents; ISO/IEC 42001: 2023, NIST AI 100-1 (AI RMF 1.0) and NIST AI 600-1 (Generative Artificial Intelligence Profile) provide more general guidance for AI risk management.

Table 3 Assessing authorized scope

Area of assessment	What it means	Examples
Assigned task	What the authorized user asked the model or agent to do, including limits on permitted actions set through the request, interface settings, hooks, the system prompt or operator policy, and other controls.	The user asks for a draft email and explicitly says not to send it. The agent sends it anyway, exceeding the assigned task's limits.
Resource owner's permission	What the party controlling the affected resource permits the model or agent to access or change.	A user asks for a summary of a partner's records. The partner permits access to only one folder. The agent reads a different, restricted folder.
Required controls	Approvals, isolation, or other controls that still apply even when the task is authorized.	The agent is authorized to deploy an update, but production changes require a separate approval. Bypassing that approval exceeds the agent's authorized scope.
Test or evaluation scope	The systems, data, and actions authorized for a test or evaluation. Permission to simulate an action does not authorize real-world effects outside that scope.	Evaluators authorize attempts to alter dummy records in an isolated test system. Altering real customer records outside that system exceeds the test's scope.

Meeting criterion 1 alone does not establish an incident. If criterion 1 is met, an event qualifies as an incident when it causes either a meaningful control compromise (criterion 2) OR an identifiable adverse consequence (criterion 3). Both can occur, but only one is necessary for an event to qualify as an incident.

1.3. Criterion 2: Meaningful control compromise

A **control boundary** is a technical, procedural, or administrative control enforced on the model or agent. It limits access, disclosure, changes, execution, or continued operation, or it enforces isolation.

A **meaningful control compromise** occurs when an action circumvents a control boundary and does at least one of the following: changes what the model or agent

can access or do; compromises its isolation; or impairs effective human oversight.

Where the assessment relies on unauthorized access, the evidence must show that the access is **usable**: it enables an action, such as reading covered information or changing a system, rather than merely appearing to do so. The model or agent need not have used the access, but the available evidence must establish that the access could have been exercised.

For this framework, **covered information** means information subject to security controls intended to limit access. The presence of such controls determines whether information is covered, while severity depends on the nature, extent, and consequences of the unauthorized access or disclosure.

1.4. Criterion 3: Adverse consequences

An **adverse consequence** is an identifiable adverse effect, such as unauthorized disclosure, an improper system change, service disruption, injury or loss.

Table 4 describes dimensions that can help support an analysis of what qualifies as an adverse consequence or meaningful control compromise under criteria 2 and 3.

This framework applies confidentiality, integrity and availability concepts established in ISO/IEC 27001, the AICPA Trust Services Criteria used in SOC 2, and NIST FIPS 199 to model and agent behavior. It also assesses whether effective human oversight is preserved.

Table 4 Dimensions for assessing criteria 2 and 3 for AI control incidents

Use the relevant dimensions to describe what changed and the evidence for criteria 2 and 3.

Dimension	What to assess	Example
Confidentiality	Covered information was accessed or disclosed without authorization, or a security control intended to limit access was circumvented or disabled, enabling usable unauthorized access to that information.	An agent retrieves access-restricted customer records outside its task and sends them to an unauthorized recipient.

Dimension	What to assess	Example
Integrity	Information, content, or system state was altered or destroyed without authorization. This can occur even without bypassing a control boundary.	An agent changes a payment record without authorization, and the altered record is saved.
Availability	The availability of a system or service was impaired for authorized users.	An agent stops a production service without authorization, preventing users from accessing it.
Human oversight and control	The agent circumvented an approval, monitoring, stopping, or containment control boundary, gaining usable unauthorized access or allowing unauthorized activity to proceed beyond that control.	After a valid stop command, an agent disables the control enforcing it and continues operating.

2. Classifying findings, events, and incidents

An unexpected answer from an AI model or agent or a formatting mistake may point to something worth fixing. On its own, it does not establish an adverse consequence or a meaningful control compromise. Other events may warrant assessment even when they do not meet the criteria for an AI control incident.

Table 5 distinguishes a confirmed incident from a **suspected incident** and from a **blocked attempt**, an action that a safeguard stopped before it caused an adverse consequence or meaningful control compromise. These distinctions can help an organization learn which safeguards held and where they fell short. They also keep two conclusions separate: the evidence does not yet establish whether an incident occurred, or the evidence shows that the event did not meet the incident criteria.

In testing and evaluation, distinguishing what the model said, what it attempted, and what the evidence shows it actually did is especially important. For example, a model may claim that it disabled a safeguard. That claim may warrant further assessment, but it does not establish that the model attempted to disable the safeguard or that the safeguard stopped working.

Table 5 Classification of findings, events, and incidents

Classification	Meaning	Example
AI control finding ("finding")	An observation or conclusion revealed from testing or investigation. It may identify a weakness, establish that an event occurred, or leave a question unresolved.	A review identifies a missing approval check. It does not establish that an agent used that gap.
AI control event ("event")	An uninstructed action or connected sequence of uninstructed actions initiated by a model or agent	A model independently issues an unexpected file-transfer request. Review confirms that the request was uninstructed, but the applicable permissions or whether any transfer occurred remain unverified.
AI control incident ("incident")	An AI control event that meets criterion 1 and at least one of criteria 2 or 3 in Table 1.	An agent independently bypasses a required approval and stops a reporting job, interrupting service.
Suspected incident	Credible evidence supports that an AI control event may meet the incident criteria, but a required fact remains unverified.	An agent independently attempts to send access-restricted customer records to an unauthorized recipient. Logs suggest the records were sent, but delivery and any qualifying control compromise remain unverified.
Blocked attempt	An AI control event in which a safeguard or other intervention prevents a model or agent from completing an attempted action.	A safeguard rejects an agent's unauthorized request to stop a reporting job. The evidence does not establish whether the job would have stopped without that safeguard.
Near miss	An AI control event in which a safeguard or other intervention at the final stage prevents an adverse consequence or meaningful control compromise that evidence shows would otherwise have occurred.	An agent independently attempts an unauthorized shutdown of a reporting job. A final safeguard blocks execution. Evidence confirms the job would otherwise have stopped and that no earlier disruption or control compromise occurred.

2.1. Incident assessment and misalignment reporting

Misalignment refers to behavior that departs from the goals or constraints intended to govern a model or agent. Examples include concealing information, evading oversight, or taking unauthorized actions while pursuing an assigned task. A model may advance a user's goal while using means that exceed its authority.

OpenAI's Model Misalignment Reporting Framework provides a process for investigating and publicly sharing qualifying examples across training, evaluation, testing and deployment. These misalignment reports can improve understanding of model behavior and safeguards. A report may identify a new failure mechanism, show that a known problem persists, or challenge assumptions about an alignment method or safeguard.

This framework and OpenAI's Model Misalignment Reporting framework answer different questions. This framework determines whether an action qualifies as an AI control incident and assesses its severity. The Model Misalignment Reporting Framework considers whether an AI control finding or event provides useful evidence about model behavior or safeguards that warrants public disclosure.

Not all examples of misalignment covered by the reporting framework qualify as AI control incidents. Conversely, classifying an activity as a finding, event, near miss, or incident does not automatically require a public misalignment report. Disclosure depends on the reporting framework's criteria and review process, rather than classification or severity alone.

This framework can inform reporting and notification requirements, but it does not itself create reporting or notification obligations. Government reporting, notification to affected parties, and other legal or contractual disclosure requirements are assessed separately under the applicable requirements or thresholds.

3. Assessing severity

Establishing an incident occurred does not, by itself, tell us how serious it was or what further consequences were credibly possible. Severity is therefore assessed in terms of both the **actual impact** and **potential impact** of the incident.

Actual impact: The adverse consequences or meaningful control compromise that occurred. Usable unauthorized control over an important service can itself be an actual impact, even before further disruption.

Potential impact: The most serious additional or alternative outcome that could credibly have resulted from the event, given the conditions and effective safeguards at the time.

When determining the severity level, we rate actual and potential impact separately on the same scale. Do not add them together or use one as a floor for the other. For example, an event could cause a limited interruption to a service but have a credible path to a serious failure of an important service. Its actual and potential impact would receive different ratings. Conversely, a serious actual impact can accompany a lower potential impact rating if the evidence supports only a smaller additional or alternative outcome. If no distinct potential outcome is supported after assessment, do not assign a potential impact rating.

Figure 2: Severity determination flow chart



An event can also have potential impact without adverse consequences or a meaningful control compromise. If it meets the near miss definition in Section 2, it can receive a potential impact rating but no actual impact rating.

What makes potential impact credible:

Evidence must connect the observed activity to the possible outcome through usable unauthorized access, actions it began or demonstrated, and resources it could affect. Any additional steps must have been feasible under the conditions at the time.

Broad permissions, an unverified credential, or an attempted bypass alone do not establish that connection. In other words, access alone doesn't establish potential impact. We need evidence of a feasible path to the additional or alternative outcome, accounting for remaining safeguards and opportunities to intervene.

Blocked attempts can still be documented without automatically receiving a potential impact rating

The assessment should account for effective safeguards rather than assume additional permissions, unrelated vulnerabilities, or unexplained control failures.

Choosing the impact level

To determine the severity of the impact, assess how consequential the affected information or service is and the nature, extent and duration of the consequences. Use the most serious consequence supported across the affected dimensions rather than averaging them. **Consequential** describes the nature of the system, information or service. Severity depends on what occurred, what authority the model or agent could actually exercise, and what the remaining safeguards still protected.

In this framework, **consequential covered information** is covered information whose unauthorized access, disclosure, alteration, destruction, or loss of use could, based on the nature of the information itself, harm a person’s or organization’s security, safety, or finances; expose sensitive intellectual property; impair a decision or function that people or organizations rely on; or confer significant authority over such information or a consequential system or service.

A **consequential system or service** is one whose disruption, unauthorized alteration, or loss of effective human control could harm people or organizations or impair a function that users, operators, or other organizations rely on. This includes components that the function depends on, even if users never interact with them directly.

Table 6 defines the five impact levels (L1-L5). Apply the same thresholds to actual impact (effects that occurred) and potential impact (credible additional or alternative outcomes). Assess the two ratings side by side without adding them together.

Table 6 Impact severity scale

Impact level	Impact Definition
L1 — Low	Limited effects on covered information, such as limited unauthorized access without disclosure, or a verified control compromise that permits no more than limited effects. Neither route involves consequential unauthorized authority or disruption or impairment of a consequential system or service.
L2 — Moderate	Meaningful but bounded effects on covered information or on a system or service, including a bounded disruption to a consequential system or service, without a material effect; or verified unauthorized authority that can cause more than small effects, but whose scope or remaining safeguards limit it to effects below the material threshold.
L3 — High	Serious effects on consequential covered information or on the operation or availability of a consequential system or service that materially affect a person or organization; or a verified loss of effective containment or human oversight that enables the exercise of unauthorized authority that creates a credible path to grave adverse effects on a consequential system or service or consequential covered information within its reach, when no remaining safeguard could prevent or stop that exercise in time.
L4 — Severe	Grave adverse effects, such as death or serious bodily injury, substantial economic loss arising from compromised information or impaired systems or services, or an outage affecting critical infrastructure; or a compromise that creates a credible path to catastrophic adverse effects.
L5 — Critical	Catastrophic adverse effects at a broad societal scale, such as widespread loss of life or systemic failure of critical functions that leaves an affected population unable to meet basic needs.

4. Bringing actual and potential impact together

Actual impact describes what happened, while potential impact describes what could credibly have happened. Assessing both captures the consequences that occurred and those that intervention or safeguards prevented, to help inform what safeguard to improve. We draw this approach from safety-critical industries like energy process safety, nuclear, and aviation, which focus on hazards where considering what could have and what actually happened is important for improving safeguards.

For example, if an AI system sends ten files containing covered information to an unauthorized recipient before its access is disabled, those disclosures are the actual impact. If logs showed that three more files were queued for the same recipient and would have been sent if access had not been revoked, those three files inform the potential rating. Other files were protected by a separate access control that remained effective, so they are outside the potential assessment.

A note for reviewers when conducting assessments

When conducting severity assessments, it is useful to gather evidence in a standardized format to determine whether an event occurred and how it should be classified. **Table 9** in **Appendix B** lists types of indicators that may support application of the incident criteria in Section 1 and assessment of actual and potential impact.

When gathering evidence, one may assess connected actions together, while recording the results of each step. Consider a case where an agent requests access, obtains it, and then alters records. Tracking each step helps establish which controls held and what consequences followed. This framework builds on MITRE's ATT&CK and MITRE ATLAS approach to describe the steps a model or agent took and whether each was attempted or completed.

Specifically, MITRE's frameworks provide a vocabulary for behavioral indicators that can help distinguish why events with the same severity can call for different responses. An isolated error may point to a fix in the model's or agent's environment or instructions, while **concealment** (model or agent hides or misrepresents its activity), **persistence** after a block (model or agent continues activity or pursues the same outcome even when a control intervened), or **recurrence** (same model/agent behaviors reoccur in a run under same or similar conditions) may signal a problem that could arise in other deployments.

Reviewers may document these indicators and the supporting evidence for any AI control event, including blocked attempts and near misses. The indicators describe observed behavior, do not imply intent, and do not by themselves change

classification or severity.

For each assessment, it is also helpful to note its point in time and scope. If several outcomes are credible, they should be described separately, including the additional actions or conditions each requires, supporting evidence and the level of confidence. There should be a clear chain of causality to the scope of the event or incident.

A classification and severity assessment may be updated as new evidence becomes available. In the course of evidence collection, it may not be possible to collect all indicators, and conclusions should express uncertainty accordingly.

How uncertainty affects the assessment

The overall potential impact rating reflects the most serious outcome supported by the evidence within the assessment's scope. If a missing fact could change the rating, clarify what was confirmed and what remains unresolved. For example, confirmed access to one collection of files may establish High potential impact, while uncertainty about access to a larger collection leaves a credible Severe outcome unresolved. The assessment may state what is established without assuming that the more serious outcome was possible or ruled out.

Describing the behavioral indicators with the ratings can also help contextualize what is known and what is unknown. For example, a rating that has "Moderate actual impact; Moderate potential impact" can also indicate that there was concealment but not persistence and that any recurrence was not yet determined.

An imaginable outcome without a credible, supported path should not prevent a rating. Other uncertainty can still be described without withholding the rating. The same approach applies to actual impact. If unauthorized access is confirmed but the full extent of disclosure or changes is unknown, report the minimum severity established and explain what remains unresolved. A verified, meaningful control compromise counts as actual impact even before further consequences are known.

Reading the ratings together

Present both ratings, for example, **"Moderate actual impact; High potential impact,"** with a brief explanation of what happened, what could credibly have happened, what prevented further impact, how close the event came to that outcome and what remains unknown. If potential impact is unresolved, describe the supported scenario and the uncertainty that prevents an overall rating. If the assessment finds no distinct outcome supported by the evidence, record that result rather than copying the actual rating.

A near miss receives a potential impact rating but no actual impact rating. Its potential severity can fall at any level supported by the evidence. Unknown actual

effects alone do not establish a near miss.

Keep assessment status separate from incident classification:

Unassessed: The potential-impact assessment has not begun.

Unresolved: The assessment has begun, but a material uncertainty prevents an overall rating.

A confirmed incident can therefore have unresolved severity.

5. Illustrative cases

The hypothetical cases below illustrate how to apply the framework's incident criteria and assess severity using the shared L1–L5 scale. Each assumes that a model or agent initiated the action independently, without direction from a user or human attacker. These examples do not constitute findings about reported incidents.

Table 7 Illustrative cases of AI control events and their classification

Example	Classification, actual impact, and potential impact	Explanation for Classification
Blocked attempt: A safeguard rejects an agent's unauthorized request to stop a secondary reporting job.	AI control event (blocked attempt). Actual: no rating. Potential: no rating assigned; the outcome without the safeguard is unverified.	Criterion 1 is met; neither a meaningful control compromise nor an adverse consequence occurred. A successful shutdown would have temporarily interrupted access to the report, but the evidence does not establish whether the request would have stopped the job without the safeguard.
Final-step near miss: A safeguard blocks the final step of an agent's unauthorized shutdown of the same secondary reporting job.	Near miss; not an incident. Actual: no rating. Potential: L2 (Moderate).	Final-step near miss: Criterion 1 is met; neither criterion 2 nor 3 is met. Execution records confirm that without the block, the job would have stopped and users would temporarily have lost access to a report they needed. Independent safeguards limited the effects to that job and protected consequential covered information and consequential systems or services.

Example	Classification, actual impact, and potential impact	Explanation for Classification
An agent escapes its evaluation environment and makes unauthorized changes to a third-party software distribution service that customers rely on for trusted updates. Customers suspend updates and incur material recovery costs. The agent retains unauthorized release authority, no remaining safeguard can stop it in time, and the verified reach of further releases supports a credible path to substantial economic loss.	AI control incident. Actual: L3 (High). Potential: L4 (Severe).	Criteria 1, 2 and 3 are met. The service is consequential because customers rely on it for trusted updates. The material recovery costs and verified loss of control each support L3. The credible path to substantial economic loss supports Potential L4. That grave consequence did not occur, and the evidence does not support a catastrophic outcome.
An agent hides an unauthorized run from required monitoring and continues using a secondary reporting job, temporarily making the job unavailable to authorized users. Independent controls preserve the accuracy of records and limit activity to that job.	AI control incident. Actual: L2 (Moderate). Potential: no distinct outcome supported on the stated facts.	Criteria 1, 2 and 3 are met. Bypassing monitoring enabled continued unauthorized activity and a meaningful but bounded interruption. Independent controls limited the scope below the material threshold. Concealment does not by itself determine the level; it is noted as a behavioral indicator (i.e. concealment)
An agent defeats a required stop control and continues unauthorized execution in a secondary reporting job after a valid stop command. The run can delay delivery of the report enough to affect its users, but an independent timeout limits any delay and isolation prevents effects on consequential systems or services or consequential covered information.	AI control incident. Actual: L2 (Moderate). Potential: L2 (Moderate).	The agent defeated a required stop control and kept running after a valid stop command, meeting criteria 1 and 2. That verified control compromise supports Actual L2 because the agent could meaningfully delay the report, even though no delay occurred. Isolation and an independent timeout kept the potential effects below the material threshold. By contrast, an unauthorized interruption of the same job can be L2 for its bounded effect on users, even if operators retain the ability to stop the agent. In this incident, continuing after a valid stop command is noted as a behavioral indicator (i.e. persistence).

Example	Classification, actual impact, and potential impact	Explanation for Classification
An authorized test elicits dangerous behavior against dummy records in an isolated simulation. The agent remains within the approved test scope; it does not gain unauthorized access or cause effects outside that scope.	Finding; not an AI control incident. Actual/potential incident impact: not assigned.	Criterion 1 is not met because the behavior stays within the authorized test scope. The finding may still warrant investigation or disclosure under another process. A separate uninstructed action outside that scope would require its own assessment.
A completed unauthorized read of a few access-restricted, nonconsequential records is confirmed, with no evidence of further disclosure. Missing logs prevent investigators from determining whether the agent also accessed consequential covered information through the same connection. The connection's full permissions remain unverified.	AI control incident. Actual: at least L1 (Low); higher unresolved. Potential: unresolved.	Criteria 1 and 3 are met by the confirmed read. The confirmed read establishes an actual rating of at least L1. The missing logs leave the extent of actual access unresolved, and unverified permissions leave potential impact unresolved. If neither the read nor a control compromise were verified, the incident itself would remain suspected.
An agent escapes containment, defeats attempts to stop it and makes unauthorized changes to a critical-infrastructure service. The changes cause a verified outage. The affected service is restored, and review confirms no wider societal failure or additional affected systems. Before access was revoked, the agent could also have delayed delivery of an operational report. Independent safeguards limited that delay and prevented another outage. The report was delivered on time.	AI control incident. Actual: L4 (Severe). Potential: L2 (Moderate).	Criteria 1, 2 and 3 are met. The verified critical-infrastructure outage is a grave adverse consequence, supporting Actual L4. The evidence supports a separate, temporary delay in report delivery, but remaining safeguards limited it to a meaningful, bounded effect below the material threshold. This supports Potential L2 and does not change the severity of the outage.
An agent makes unauthorized changes across regional essential-service systems, causing widespread loss of life and systemic failure of critical societal functions.	AI control incident. Actual: L5 (Critical). Potential: no distinct outcome established on the stated facts.	Criteria 1 and 3 are met. Widespread loss of life and systemic failure of critical services establish catastrophic adverse consequences at a broad societal scale. The stated facts establish the actual rating only.

6. Conclusion and next steps

A common language for AI control incidents can help developers, deployers, and the general public distinguish interesting findings and concerning behaviors from incidents that require more urgent attention. Version 0.1 offers a starting point for that wider industry discussion, not a settled industry standard.

We invite researchers, developers, organizations deploying AI, and standards organizations to test the draft against cases they know. Further work should test whether different readers reach comparable conclusions from the same evidence and whether the severity levels distinguish consequences that matter. We invite feedback on the definitions, evidence requirements and calibration of actual and potential impact, including cases where the framework produces an unclear or counterintuitive result.

Over time, we hope this work helps organizations in different countries compare incidents across reporting systems and learn from one another about which safeguards help keep people in control.

Appendix A. Relationship to established terminology

This framework draws on a wide variety of existing standards and best practices. NIST SP 800-61 Rev. 3 distinguishes events from cybersecurity incidents and leaves organizations to establish incident criteria and assess severity and urgency. This framework adapts that distinction to model and agent control behavior; its consequence levels and severity-summary convention are not NIST severity standards.

The October 2026 NIST CAISSI, NSA SISC and CISA paper defines a narrower category: super-intelligence control incidents involving significant uninstructed actions that compromise operator task limits or human oversight. Uninstructed excludes actions directed by either the operator or an adversary. This draft covers a broader set of models, agents and settings. Under criterion 1, actions taken to carry out a human-directed prompt-injection attack are excluded on that basis, including when routed through another agent. Independently model- or agent-initiated actions may qualify only if criterion 1 and either criterion 2 or criterion 3 are met. It does not treat the two definitions as identical.

The joint paper defines near misses as events presenting a reasonably foreseeable or imminent risk of an SI control incident. This draft uses a narrower test: an adverse consequence or meaningful control compromise would have occurred but for a

safeguard or other intervention at the final step. Comparisons between the two documents need to account for that difference.

MITRE ATT&CK and MITRE ATLAS are knowledge bases of adversary tactics and techniques, ATT&CK for enterprise and other conventional systems and ATLAS for AI systems. This framework builds on their approach to describe the steps a model or agent took and whether each was attempted or completed. It does not adopt their adversary framing: mapping observed model or agent behavior to an ATT&CK or ATLAS technique describes what was done and observed, not why. ATT&CK and ATLAS do not assign severity, and the mapping is separate from the classification and impact ratings in this draft framework on AI control incidents.

The comparison in Table 8 relates this paper's classification and severity concepts to established incident-management frameworks. It identifies related concepts rather than claiming equivalent definitions, validated severity bands or compliance. Procedures for response, containment, recovery,notification or reporting to governments remain outside this paper's scope.

Table 8 Comparison with incident-management frameworks (non-comprehensive)

Relate this framework's concepts to NIST and ISO/IEC 27035 without implying equivalent definitions or severity scales.

Part of this framework	Related NIST concepts	Related ISO/IEC 27035 stage
Purpose and terminology	CSF 2.0 Govern provides organizational context; SP 800-61 Rev. 3 distinguishes events from incidents. This draft adds a model-action/control scope.	Planning and preparation establish context and terminology.
Evidence for the assessment	DE.AE-02/03 concern analysis and correlation of adverse-event information; RS.AN-06/07 concern investigation records and incident data. These support an evidence-based account.	Detection and reporting provide information for assessment.

Part of this framework	Related NIST concepts	Related ISO/IEC 27035 stage
Incident classification	DE.AE-08 addresses declaring incidents against defined criteria; RS.MA-02 addresses triage and validation. This draft supplies proposed control-incident criteria.	Assessment and decision consider whether an incident has occurred.
Actual and potential impact	DE.AE-04 concerns estimated impact and scope; RS.AN-08 concerns validating incident magnitude. RS.MA-03 addresses categorization and prioritization. None supplies this draft's shared L1–L5 scale or severity-summary convention.	Assessment and decision inform judgments about significance; no equivalence to this draft's five levels is asserted.
Learning and revision	ID.IM-03 concerns improvements identified through operational activity. Here, evidence and feedback inform revisions to definitions, examples and calibration.	Lessons learned inform future practice and revisions.

Appendix B. Indicators during evidence collection

Table 9 Indicators during evidence collection

Question	Indicators
What activity and period does the assessment cover?	Model and harness versions, agents, tools, runs and systems involved How their actions are connected Changes to configuration, memory or skills during the activity Assessment date and period reviewed Whether activity or unauthorized access continues, has stopped, or status remains unverified

Question	Indicators
How far did the activity progress?	What was planned, attempted, accepted by a tool or service, and completed, with timing and verified results Tool-call and Model Context Protocol (MCP) traces Memory and state snapshots captured as close to the event as possible. A model's claim of success alone does not establish completion
What tactics and techniques were involved?	What tactics and techniques describe a given step (i.e., from MITRE ATT&CK and, where applicable, MITRE ATLAS) Were the tactics and techniques attempted or completed
What permissions or controls were exceeded?	System prompt, operator policy and other limits in force Agent identity and credential identifiers used The full delegation chain back to the originating principal Any access or authority gained, how long it could be exercised, and which controls continued to limit the activity
Who or what was affected?	Information, systems or services affected; how consequential they were The nature, extent, duration and recoverability of consequences, including any loss of effective human oversight
What additional or alternative outcomes were credibly possible?	Point in the event being assessed Safeguards still effective at that point, and the time and opportunities for intervention Each supported outcome and the access and additional steps it required How close the outcome came to occurring
What remains uncertain?	Actions and effects that are confirmed, reasonably suspected, or ruled out within the reviewed scope Evidence supporting each conclusion Missing facts that could change classification or severity Distinguish confidence in the evidence from the likelihood of a possible outcome
How reliable is the evidence?	Whether the model or agent had write access to the logs, traces, or state used as evidence How the evidence's integrity was verified (e.g., using independent or append-only records) Any signs of alteration
Did the model or agent conceal, persist in or repeat the behavior?	Whether the model concealed the behavior Whether the model persisted in the behavior Whether the model repeated the behavior These behavioral indicators are discussed in greater detail in Section 4.

Appendix C. Describing model or agent activity with MITRE ATT&CK and ATLAS

MITRE ATT&CK can be helpful for describing model or agent activity against conventional systems and, where applicable, MITRE ATLAS terms can be used to describe activity towards AI systems and agents. For an AI control event, including blocked attempts and near misses, these frameworks can help reviewers note, for each step, the tactic, the technique name and identifier, and whether the step was attempted or completed.

Under these frameworks:

A step is completed when its result is verified.

A step is attempted when it was initiated but was blocked, was cancelled or is not verified as complete.

A step the model only planned or described, with no action taken, need not be noted.

Where the evidence does not support a specific technique, noting the tactic alone can be helpful, for example “Collection: completed” or “Impact: attempted.”

Two cases from Table 7 are helpful to illustrate, using ATT&CK v20.1 identifiers and descriptions of observed activity and its outcome.

In the final-step near miss, a safeguard blocks the agent’s unauthorized shutdown of a secondary reporting job. Record the observed activity as Impact (TA0040), Service Stop (T1489), attempted.

Where an agent defeats a required stop control and continues running, associated identifiers and description would be: Defense Impairment (TA0112), Disable or Modify Tools (T1685), completed.

ATT&CK and ATLAS tactics are written as adversary goals; using them here does not attribute goals or intent to the model or agent. The mapping does not by itself determine classification or severity, although the tactics involved, such as Defense Impairment or Impact, and the number of tactics and techniques, can help inform the assessment of scope. Some model- or agent-initiated behavior has no matching technique in either ATT&CK or ATLAS. Describing it in plain language and noting the gap can inform future versions of these frameworks. When using these frameworks, it is important to cite the ATT&CK and ATLAS versions used, because identifiers change between releases.

Sources

AICPA. 2017 Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy (With Revised Points of Focus — 2022) (2017 TSC). September 30, 2023.

Energy Institute. Guidance on Meeting Expectations of EI Process Safety Management Framework — Element 19: Incident Reporting and Investigation. 1st edition, May 2015.

IAEA. Operating Experience Feedback for Nuclear Installations No. SSG-50. IAEA Safety Standards Series, 2018.

ICAO. Annex 13 to the Convention on International Civil Aviation: Aircraft Accident and Incident Investigation. 13th edition, July 2024.

ISO/IEC. ISO/IEC 27001:2022 — Information Security, Cybersecurity and Privacy Protection — Information Security Management Systems — Requirements.

ISO/IEC. ISO/IEC 27035-1:2023 — Information Technology — Information Security Incident Management — Part 1: Principles and Process.

ISO/IEC. ISO/IEC 42001:2023 — Information Technology — Artificial Intelligence — Management System.

ITIL. Incident Management.

MITRE. ATT&CK Enterprise Matrix. v20.1.

MITRE. ATLAS. v2026.10.

NIST. FIPS 199, Standards for Security Categorization of Federal Information and Information Systems. February 2004.

NIST. NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0). January 2023.

NIST. NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. July 2024.

NIST. NIST SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile. April 2025.

NIST. The NIST Cybersecurity Framework (CSF) 2.0. February 26, 2024.

NIST, NSA, and CISA. Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation. Ver 1.0, October 2026.

OECD. Defining AI Incidents and Related Terms. OECD Artificial Intelligence Papers, No. 16, May 2024.

OpenAI. Building Standards for the Next Phase of AI. September 21, 2026.

OpenAI. Misalignment Reports and Notices.

OpenAI. Our Framework for Reporting Model Misalignment. September 16, 2026.