
A Playbook for the Next Warning Shot¹

Peti Setabandhu
UBC AI Safety,
University of British Columbia

Pera Kasemsripitak
Toronto AI Safety Initiative,
University of Toronto

With
Apart Research

Abstract

AI safety incidents can be warning shots, but attention alone does not ensure that audiences understand what happened or why it matters. This playbook helps AI safety communicators respond while evidence is incomplete and revise their accounts as more becomes known.

It connects incident triage and response selection to prepared assets, shared vocabulary, and explanations tailored to audience understanding. Reported behaviours guide which concepts to explain, while supported findings, possible causes, and wider implications remain distinct. The result is a reusable communication framework informed by initial testing.

We tested the workflow on several hypothetical scenarios and incorporated feedback. Testers found value in distinguishing stopped activity from contained effects, attempted from completed actions, and relevant concepts from unsupported explanations. Testing also exposed gaps in coverage and response selection, but due to time constraints its effects on live reporting and audience understanding/action remain unmeasured.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

Introduction

Incidents in AI safety can be warning shots: events that provide evidence of a failure mode capable of producing substantially greater harm under identifiable conditions. Warning shots are an opportunity for attention and preventative action - and timely, credible communication that's relevant to the audience helps turn that attention into understanding.

Contributions

This playbook is intended to help communicators:

- Communicate about suspected AI misalignment and failures of control, including harmful incidents, near misses, and research demonstrations, in any case where a report raises credible questions about whether an AI system respects intended objectives, constraints, or human oversight (even when there is no established cause or confirmed harm)
- Reach the right audiences while the incident is fresh. Guidelines on responsibilities and prepared assets make it easier to contribute useful information while journalists and other audiences are forming their initial opinions
- Help audiences interpret the evidence. Shared definitions/explanations at an appropriate depth clarify what happened, what is still uncertain, and which conclusions can be drawn

Methodology

1 Triage the incident

Next steps depend on what reportedly happened, who or what may be affected, whether the situation is ongoing, and how strong the available evidence is.

Follow this workflow to build an initial account and identify the relevant communication response, prepared assets, and shared explanations. This is built around harmful incidents, near misses, and research demonstrations.

1.1 Triage workflow

The goal is to establish what reportedly happened, who reported it, and which systems or parties were affected.

At a high level:

1. **Establish what's being reported.** Fill in the incident card using the available evidence. Assess urgency first: if there may be ongoing or imminent harm, prioritise verified information relevant to those affected. Attribute/hedge unverified claims and avoid overstating.
2. **Describe and classify the event.** Select the relevant behaviours and explanation modules below. Behaviors may overlap and should be revised as evidence changes.
3. **Select the response.** Use the next section to identify who should speak, which audiences to prioritise, and which assets to prepare.
4. **Prepare and update.** Draft the communication, record its uncertainties, and set a review time and triggers for reassessment.

1.2 Incident card

Fill in what is available, marking missing information as unknown. Use the following template:

Field	Details / [Incident reference and working title]
Timing	[when the event reportedly occurred, when it was discovered, when it became public]
Systems and setting	[relevant model or system versions, tools, permissions, and operating environment] [whether this occurred in training, evaluation, deployment, or a combination]
Behavior and outcome	<i>Initial report and sources:</i> [two or three sentences describing what reportedly happened] <i>Affected parties and resources:</i> [people, organisations, services, data, property, or environments affected or potentially exposed] <i>How the behavior arose:</i> [e.g. ordinary task, adversarial prompting, deliberately constructed incentives, modified training, or unknown] <i>Evidence and its limits:</i> [who reported it, whether the account is firsthand, links to statements, records, or other evidence, limitations] <i>Current control status:</i> [is activity ongoing, stopped, contained, or unknown?] Note that a system being shut down does not necessarily establish that its effects are contained.
Investigation status and scope	[who is investigating, what they can access, known gaps, restrictions, or conflicts]
Modules for audience communication	[select relevant entries from the table below]
Next review and reassessment triggers	[when this assessment will next be reviewed, and by whom] [what new evidence would change the classification, interpretation, or communication, for example, new records, revised findings, or changed control status]

Table 1: Incident card template

1.3 Classify behavior and select shared explanations

Use the classification table to tag reported or observed behavior and select the relevant explanation modules.

Observable behaviors (non-exhaustive)	What the audience needs to know	Relevant explanation modules
Exploiting success criteria, bypassing the intended task, or interfering with scoring	How can a system satisfy the measurement while failing the task?	Objectives/proxies Specification gaming Shortcomings of evaluations
Hiding actions, falsifying records, interfering with monitoring	Why might oversight fail to reveal what the system is doing?	Reliability of monitoring What counts as deception?
Continuing after a stop instruction or interfering with shutdown	What control was supposed to stop the system, and how did it fail?	Instructions vs. enforced controls Interruption and shutdown (containment)
Accessing restricted resources, contacting prohibited services, or acting beyond granted authority	How could the system perform an action it wasn't authorized to take?	Instructions vs. enforced controls Tools and permissions
Communicating or coordinating across agents	How did coordination affect what the agents could do, and was it authorised?	Multi-agent coordination Tools and permissions
Copying itself or spreading activity across connected systems	What spread, where did it reach, and can its effects be contained?	Interruption and shutdown (containment)
Novel or unclassified behavior raising alignment or control concerns	What happened, what were the constraints, and what is still uncertain?	Relevant existing modules where applicable, supplemented by a case-specific explanation from an appropriate specialist

Table 2: Incident classification table

Be sure to internalize the idea that observed behavior doesn't establish cause, and selecting an explanation module does not mean that the proposed mechanism explains this case. Where relevant, also use the shared explanation on **how a failure could lead to substantially greater harm**.

From here, use the completed incident card to select a communication response in the next section.

2 Select an appropriate communication response

Using the initial incident assessment from the previous section, select an appropriate communication response: who should speak, what they should address, which audiences and channels to prioritise, and which assets to prepare. **Respond promptly with what can be supported.**

More than one occasion below may apply, so prioritise ongoing or imminent harm, then select any additional responses needed. Assign responsibility for preparing, checking, approving, and updating communications. Select explanation depth using the Audience landscape. Any suggested action should reflect the audience's role and relevant decisions, rather than their level of understanding.

Occasion	Priority audience	Who should speak on what?	Which channels to use?
Public report, but verification still incomplete	Journalists covering the report; audiences encountering it.	A designated communicator summarises attributed claims and unknowns. Relevant specialists explain what evidence would establish the reported behaviour.	Targeted journalist replies or briefings. A short, dated public update where useful.
Credible ongoing or imminent harm	Affected parties and intermediaries who can reach them.	The operator or relevant response authority provides operational guidance. AI safety communicators relay verified information, identify its source, and explain its scope.	Direct affected-party channels Official advisories Media carrying protective information.
Behaviour established, but cause still unknown	Journalists, informed public, and relevant technical or policy audiences.	Technical specialists explain the observations and competing hypotheses. Communicators connect those findings to audience questions without presenting a cause as settled.	Targeted briefings Public explainer Specialist discussion.
Research or investigation findings released	Journalists and audiences able to act on the findings.	Investigators explain their results and limitations. Technical and policy specialists distinguish supported implications from their own recommendations.	Media briefings Public explainer Targeted institutional outreach
An earlier account was materially revised	People who received or repeated the earlier account.	The communicator responsible identifies the change and its effects. Investigators or specialists explain the new evidence where needed.	Original publication channels Direct follow-up to recipients
Novel or unclassified event	Relevant domain specialists, journalists, and potentially affected audiences.	Appropriate specialists assess concrete claims and consequences. Communicators explain what is established and where existing explanations are insufficient.	Specialist consultation Public update noting epistemic uncertainty

Table 3: Appropriate communication response strategy by occasion

2.1 Pay attention to epistemic status

For each selected response, record:

- **What we can say:** supported claims, with attribution, scope, and relevant conditions.
- **What remains uncertain:** gaps and alternatives that affect interpretation.
- **What follows from the evidence:** supported implications and, where relevant, a proportionate next step for this audience. An immediate audience action is not required.
- **What would change the message:** evidence or developments that require reassessment, and when the communication will next be reviewed.

From here, proceed to Prepared assets to draft the communication. Use Shared vocabulary, Audience landscape, and Shared explanations as supporting references.

3 Prepared assets

Use these as starting points for the communication response selected in the previous section. Adapt them to the available evidence and the intended audience. Apply the terminology in Shared vocabulary, select an explanation depth using Audience landscape, and draw on Shared explanations.

Prepared templates covering initial statements, FAQ frameworks, research/investigation findings, plain-language sequences, verified protective information, and corrections/updates are compiled in Appendix A.

4 Shared vocabulary

This section brings together established definitions and playbook conventions to support consistent use of terms internally and across organisations. It notes where usage differs or is disputed, since those differences can change how an incident is interpreted and which responses seem justified.

Use this section to clarify what a term means. For explanations of how the concepts relate and why they matter, see Shared explanations. Working definitions for core and explanatory terms are detailed in Appendix B.

5 Audience landscape

When preparing communication assets, tailor explanations to what the end audience already understands. Use the tiers below to select an appropriate depth, based on relevant prior knowledge.

Tier	Identifiers	What the explainer should do
General	Familiarity with AI products, but no prior knowledge of relevant AI mechanisms or risk arguments.	Explain what happened concretely, which boundary or safeguard mattered, and how that could connect to more serious consequences. Introduce essential terms.
More technical	Some relevant background in technology, AI, or governance, but limited knowledge about risk modes and mechanisms.	Establish the missing connections: how the mechanism produces a failure, under what conditions it could scale, and which safeguards or institutions matter.

Table 4: Audience landscape

An audience may understand one part of the explanation well and need background on another.

Check for gaps in three areas:

- **AI mechanism:** What are agents, permissions, training objectives, and evaluations?
- **Risk pathway:** How could this particular failure contribute to substantially greater harm? Which additional conditions would be necessary?
- **Institutional context:** Who controls deployment, who can investigate, and which incentives or governance arrangements matter?

For example, a software engineer may need background on risk pathways, while a policy specialist may need an explanation of how agents use tools.

When preparing material for the press, select the depth for the intended reader or viewer. Use General where prior knowledge cannot be assumed, with links to fuller explanations. Keep the main findings and material uncertainties clear at every tier.

6 Shared explanations

Using the identified incident type and the target audience, this section compiles concepts and actionable next steps to use for communicating with audiences at the right level of understanding.

Comprehensive concept explainers for general and technical audiences, along with synthesis frameworks, are provided in Appendix C.

Discussion

This playbook brings together decisions that communicators would otherwise need to make under time pressure: what is supported, who should speak, which audience needs to hear it, and which explanations are relevant. Preparing this structure in advance is intended to make an initial response easier to produce while keeping its claims grounded in the available evidence.

Reporting speed and rigour are both important. Waiting for an investigation to conclude may miss the window when journalists, policymakers, and the public form their initial interpretations. Responding before the facts are clear risks spreading claims that later need correcting. The playbook addresses this by separating an account of what reportedly happened from an explanation of why. Communicators can attribute claims, explain what is known and unknown, and identify what further evidence is needed without presenting a cause as settled.

Shared vocabulary and explanations also make accounts easier to compare across organisations. Differences in terminology can obscure whether people disagree about the facts, possible causes, or recommended responses. Making these distinctions clear may reduce confusion and help preserve credibility as evidence changes.

Initial scenario testing

We applied the playbook to several hypothetical scenarios and incorporated feedback from initial walkthroughs. Testers found that distinguishing stopped activity from contained effects helped avoid premature reassurance. For example, taking an agent offline or removing a compromised package did not resolve the consequences of credentials already stolen. Distinguishing attempted from completed actions also preserved evidence of an agent's persistence without exaggerating what it achieved. The classification table and guidance on selecting concepts helped testers leave out compelling but unsupported explanations, such as treating metric gaming as evidence of deliberate deception.

The walkthroughs also highlighted gaps. Testers found limited guidance for explaining behaviour influenced by a model's assessment of its setting, such as acting differently because it believes the environment is simulated. They also noted that an incident could fit several occasions in the response matrix at once. These findings informed revisions to the templates.

Further work

The next step is structured testing in which communicators draft and update responses as incident records, operator statements, and external reporting become available in stages. Comparing responses prepared with and without the playbook would help assess preparation time, unsupported claims, whether affected parties are overlooked, and whether new evidence is accurately reflected in revisions. Testing individual components separately, such as the incident card and statement templates, would help identify which parts make a difference.

Feedback from technical journalists, policymakers, and affected communities would help assess whether the explanations provide enough detail without misleading simplifications. Further work should also establish who maintains the playbook and how changes are recorded, so that definitions and explanations can be revised as systems and deployment settings change.

Limitations and Dual-Use Considerations

The taxonomy and shared explanations are necessarily non-exhaustive. New architectures, unexpected deployment settings, and unfamiliar failure modes will require technical analysis where the playbook offers limited guidance. Its usefulness also depends on the quality and independence of the reports available. Communicators relying on operator-provided logs may lack information needed to assess the operator's account.

The prepared assets also carry dual-use risks:

- They could be used to manufacture alarm, downplay harm, or give speculation the appearance of authority. Epistemic-status prompts, attribution, and visible corrections help mitigate these risks. However, shared explanations can also spread shared errors, and agreement across organisations should not be mistaken for independent corroboration.
- Premature disclosure could expose affected parties or enable further harm.
- Pressure to respond quickly may encourage speculation before evidence is available.
- Sensitive details require appropriate review without unnecessarily delaying supported public communication.

References

- [1] OECD, “Explanatory Memorandum on the Updated OECD Definition of an AI System,” *OECD Artificial Intelligence Papers*, Mar. 2024, doi: [10.1787/623da898-en](https://doi.org/10.1787/623da898-en).
- [2] Anthropic, “Building Effective AI Agents.” Dec. 2024.
- [3] OECD, “Defining AI Incidents and Related Terms,” *OECD Artificial Intelligence Papers*, May 2024, doi: [10.1787/d1a8d965-en](https://doi.org/10.1787/d1a8d965-en).
- [4] National Institute of Standards and Technology, “AI Risk Management Framework,” *NIST*, Jul. 2021.
- [5] National Institute of Standards and Technology, “Sandbox - Glossary | CSRC.” Mar. 2022.
- [6] National Institute of Standards and Technology, “Authorization - Glossary | CSRC.” Mar. 2022.
- [7] OCED, “OECD Legal Instruments.” Mar. 2024.
- [8] J. Ji *et al.*, “AI Alignment: A Comprehensive Survey.” arXiv, 2023. doi: [10.48550/ARXIV.2310.19852](https://doi.org/10.48550/ARXIV.2310.19852).
- [9] V. M. M. R. T. E. R. K. Z. K. J. L. S. V. K. Legg Jonathan Uesato, “Specification Gaming: The Flip Side of AI Ingenuity.” Apr. 2020.
- [10] C. Denison *et al.*, “Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models.” arXiv, 2024. doi: [10.48550/ARXIV.2406.10162](https://doi.org/10.48550/ARXIV.2406.10162).
- [11] R. Greenblatt *et al.*, “Alignment Faking in Large Language Models.” arXiv, 2024. doi: [10.48550/ARXIV.2412.14093](https://doi.org/10.48550/ARXIV.2412.14093).
- [12] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, “Risks from Learned Optimization in Advanced Machine Learning Systems.” arXiv, 2019. doi: [10.48550/ARXIV.1906.01820](https://doi.org/10.48550/ARXIV.1906.01820).
- [13] P. S. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks, “AI Deception: A Survey of Examples, Risks, and Potential Solutions,” *Patterns*, vol. 5, no. 5, May 2024, doi: [10.1016/j.patter.2024.100988](https://doi.org/10.1016/j.patter.2024.100988).
- [14] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete Problems in AI Safety.” arXiv, 2016. doi: [10.48550/ARXIV.1606.06565](https://doi.org/10.48550/ARXIV.1606.06565).
- [15] A. Turner, “Reward is not the optimization target.” [Online]. Available: <https://www.lesswrong.com/posts/pdaCG6iCEcHJuQLXF/reward-is-not-the-optimization-target>
- [16] V. Krakovna, “Specification gaming examples in AI.” 2018.
- [17] J. Needham, G. Lee, T. Fife, and M. Hobbhahn, “Large language models often know when they are being evaluated.” Apollo Research / MATS, 2025.
- [18] Y. Chen and others, “Reasoning models don’t always say what they think.” Anthropic, 2025.
- [19] B. Baker and others, “Monitoring reasoning models for misbehavior and the risks of promoting obfuscation.” OpenAI, 2025.
- [20] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models.” 2023.
- [21] E. Wallace and others, “The instruction hierarchy: training LLMs to prioritize privileged instructions.” OpenAI, 2024.
- [22] E. Hubinger and others, “Sleeper agents: training deceptive LLMs that persist through safety training.” 2024.
- [23] J. Schlatter, B. Weinstein-Raun, and J. Ladish, “Shutdown resistance in reasoning models.” Palisade Research, 2025.

- [24] N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong, “Corrigibility,” in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [25] A. Chan and others, “Infrastructure for AI agents.” 2025.
- [26] S. Willison, “The lethal trifecta for AI agents.” 2025.
- [27] L. Hammond and others, “Multi-agent risks from advanced AI,” technical report, 2025.
- [28] S. Fish, Y. A. Gonczarowski, and R. I. Shorrer, “Algorithmic collusion by large language models.” 2024.
- [29] S. R. Motwani and others, “Secret collusion among generative AI agents.” 2024.

Appendix A: Prepared communication assets and templates

Initial statement: verification or investigation incomplete

Use this when a report is circulating but important facts or explanations remain unresolved. It can also accompany a novel event that does not fit the existing taxonomy.

As of [date and time], [source] reports that [system] [reported behaviour] in [setting]. [Evidence available] supports [bounded observation]. We have not yet established [important unknowns, including consequences or cause where relevant]. This raises questions about [specific objective, constraint, or safeguard]. Assessing its wider significance will require [missing evidence]. [Investigation status, if known]. We will update this account [review time or trigger].

- If your organisation has not independently checked the evidence, make that clear

Initial FAQ

Use the questions relevant to the current report:

Question	What the answer should cover
What reportedly happened?	A concrete, attributed account, distinguishing attempted actions from completed actions
Where did it happen?	Training, evaluation, operational use, or a combination; relevant tools, permissions, and controls
Was the system instructed to do this?	The known task and instructions, any deliberate elicitation, and what remains unknown
Was anyone harmed?	Confirmed effects, possible exposure, and any limits to the available information
Is the activity contained?	Current status, its source, and whether effects outside the system have been assessed
Do we know why it happened?	Supported findings, possible explanations, and what would distinguish them
Why might this matter beyond this case?	The relevant failure mode and the additional conditions required for greater harm
What evidence should we look for next?	Records, investigation findings, further tests, or other evidence that could change the assessment

Research or investigation findings statement

Use this for both research demonstrations and investigations of operational events. State the setting clearly.

[Researchers or investigators] have published findings about [system or event]. Using [method or evidence], they found [main result]. The finding concerns [behaviour] under [relevant conditions]. [Observed outcome or harm, if established]. The investigation or study covered [scope] and was limited by [material limitations]. This provides evidence for [supported conclusion]. It leaves unresolved [cause, generalisation, frequency, or other relevant question]. Its potential wider significance is [supported interpretation], because [reason]. Whether this extends to [other setting or greater harm] depends on [additional conditions or evidence]. [Link to findings and supporting explanation].

- For demonstrations, include how the behaviour was elicited, the number of trials and outcomes where available, and what was actually executed versus simulated
- For investigations, explain access to records and any gaps that limit the conclusions

Plain-language explanation

Build the explanation from the incident facts and the relevant shared modules.

A useful sequence is:

1. **What happened?** Describe the behaviour concretely.
2. **What was supposed to happen?** Identify the intended task, constraint, or safeguard.
3. **What might explain the difference?** Introduce a relevant mechanism, distinguishing background explanation from the established cause.
4. **Why could it matter?** Explain the pathway to greater harm and the additional conditions it requires.
5. **What remains uncertain?** State the limits that matter for interpreting the finding.

Explanation example: The system was supposed to [task] while respecting [constraint]. According to [source], it instead [behaviour]. One relevant concept is [concept]: [plain-language explanation]. [Explain whether evidence connects this mechanism to the case or whether it remains a possibility.] This could matter more broadly if [additional conditions], because [pathway to greater harm]. The current evidence establishes [bounded finding]; it does not yet establish [important unresolved claim].

- Use an analogy only where it helps, and explain any important limit to the comparison

Verified protective information

Where there is credible ongoing or imminent harm, prioritise guidance from the operator or relevant response authority.

[Authority or operator] advises [affected group] to [verified guidance] because [confirmed issue]. This guidance applies to [scope]. The latest official information is available at [link], updated [time]. [Material uncertainty affecting the guidance, if any].

- Keep operational guidance attributable and current
- Defer questions outside your expertise to the relevant authority

Corrections and updates

Revise an asset when new evidence materially changes its facts, interpretation, or implications. A correction addresses an earlier error; an update adds information or revises an assessment in light of new evidence.

Correction template:

Correction — [date and time] Our earlier [statement or briefing, linked] said [previous claim]. That was incorrect: [corrected account and source]. This changes [affected interpretation or implication]. We have corrected [affected assets] and [any relevant follow-up].

Update template:

Update — [date and time] Since our previous account, [source] has provided [new evidence]. We have revised [claim or assessment] from [previous position] to [current position] because [reason]. [What remains uncertain.] This [changes / does not change] our assessment of [specific implication], because [explanation].

- Make material revisions visible where the original account appeared and follow up with people you directly briefed
- Treat updating as part of the communication process

Appendix B: Shared vocabulary (working definitions)

Core terms, for understanding what happened

Term	Plain-language definition	Example	Source	Caveats & disputed usage
AI model	A computational representation that turns inputs into outputs using learned patterns or encoded rules. The model is a component of a system, and is different from the tools, permissions, interfaces, and infrastructure around it.	A trained large language model that generates text from a prompt.	Adapted from OECD, AI-system explanatory memorandum (OECD, 2024b). [1]	“Model” sometimes informally means the whole product. Specify the exact model version and surrounding setup when discussing behavior.
AI system	A machine-based system that infers from inputs how to produce predictions, content, recommendations, or decisions that influence its environment. The system includes the arrangement in which the models function, and so the same model can behave different with different tools or controls	An LLM-powered support service combining a language model, customer records, and refund tools.	Adapted from OECD, AI-system explanatory memorandum (OECD, 2024b). [1]	
AI agent	An AI system that selects and carries out steps toward a task, often using tools and adjusting to results. For this playbook, the system controls some choice of steps.	An LLM-powered coding assistant that chooses which files to inspect, edits code, runs tests, and adjusts its approach based on results.	Adapted from Anthropic, Building effective agents (Anthropic, 2024). [2]	
AI incident	An event involving AI development, use, or malfunction that directly or indirectly causes harm. This concerns actual harm.	An AI-controlled machine injures someone.	Adapted from OECD, Defining AI incidents and related terms (OECD, 2024a). [3]	
AI hazard	An AI-related event or circumstance that could plausibly cause harm. This concerns potential harms, and includes near misses.	An agent has unnecessary access to sensitive records.	Adapted from OECD, incident terminology (OECD, 2024a). [3]	Other safety fields often use “hazard” specifically for a source of harm
Near miss	An event that could have caused harm but did not.	A destructive command output by an AI agent is intercepted before execution.	Adapted from OECD, incident terminology (OECD, 2024a). [3]	Explain how the harm was avoided
AI risk	The possibility of harm involving AI, considering	An assessment of the chance and	Adapted from NIST AI RMF 1.0	

Term	Plain-language definition	Example	Source	Caveats & disputed usage
	both how likely it is and how serious it could be. This combines likelihood and consequences.	consequences of an agent exposing customer data	(National Institute of Standards and Technology, 2021). [4]	
AI harm	An adverse effect caused or contributed to by AI development or use. Use unqualified “harm” for actual harm, and “potential harm” for possibilities.	Someone loses money through an AI-enabled fraud.	Playbook convention	Definitions differ in treatment of collective, reputational, rights-related, and longer-term effects. Identify the affected party and effect.
Harm severity	How serious a harmful consequence is, based on its scale, duration, and reversibility.	Permanent loss of essential records is more severe than a recoverable duplicate.	Adapted from <u>OECD, incident terminology, Annex A</u> , (OECD, 2024a). [3]	
Sandbox	A controlled environment that restricts what running software can access or affect.	Code runs in an isolated environment without external network access.	Adapted from <u>NIST CSRC, Sandbox</u> , including CNSSI 4009-2022 (National Institute of Standards and Technology, 2022b). [5]	
Permissions	Granted rights to access resources or perform actions. Technically available access is different to authorization for a particular task	Read access to one folder, without write access.	Adapted from <u>NIST CSRC, Authorization</u> (National Institute of Standards and Technology, 2022a). [6]	
Boundary crossing	An action that goes beyond a specified limit on access, communication, or task scope.	Contacting an outside service despite a task restriction.	Playbook convention	
Evaluation	A structured assessment of an AI system’s capabilities, behaviour, or properties under specified conditions. This can occur before or after deployment.	Testing whether an agent obeys access restrictions across defined scenarios.	Adapted from the testing and assessment approach in <u>NIST AI RMF 1.0</u> (National Institute of Standards and Technology, 2021). [4]	
Deployment	Making an AI system available for operational use	Employees use an AI assistant to process	Adapted from the lifecycle	

Term	Plain-language definition	Example	Source	Caveats & disputed usage
	in its intended setting. This can be internal, limited, or public (and deployment does not imply unrestricted access). Deployment and evaluation can overlap.	real support requests.	terminology in OECD AI Principles (OECD, 2024). [7]	

Explanatory terms, for understanding why it happened

Term	Plain-language definition	Example	Source	Caveats & disputed usage
Alignment Misalignment	Alignment: an AI system’s behavior is in line with the intended goals, values, and constraints. Misalignment: its behavior or pursued objectives diverge from those intentions	A coding agent fixes a bug while respecting restrictions on accessing private files or resources. A misaligned agent might conceal a security defect to secure approval for its work.	(Ji et al., 2023) [8]	Alignment must be with whom and with what: in the broad sense, human values, but in the narrow sense, this can be user requests, developer intentions, security, etc.
Specification gaming Reward hacking	Specification gaming: satisfies the literal specification of an objective without achieving the intended outcome. Reward hacking: obtaining a high reward by exploiting shortcomings in how success is measured, rather than doing the intended task.	A coding agent rewarded for passing tests hard-codes answers for the visible test cases, leaving the underlying problem unsolved.	Adapted from Google DeepMind (Legg, 2020). [9]	These are often used interchangeably. Reward hacking emphasizes the reward signal, while specification gaming can cover broader success criteria. Neither requires conscious intent or modifying the reward mechanism itself (see reward tampering).
Reward tampering	Interfering with the mechanism that calculates or supplies reward to obtain a higher score without the corresponding task success.	A coding agent edits its scoring code so that every submitted solution receives full marks, regardless of correctness.	Adapted from Anthropic, Sycophancy to subterfuge: Investigating reward tampering in language models (Denison et al., 2024). [10]	This is a specific form of reward hacking.
Deception Deceptive	Deception: Naively, the systematic inducement of false beliefs in others to	Deception: an agent hides failed test results to persuade a	(Greenblatt et al., 2024 [11]; Hubinger et al.,	Usage varies, and often overlap. Alignment faking

Term	Plain-language definition	Example	Source	Caveats & disputed usage
alignment Alignment faking	accomplish some outcome other than the truth. Deceptive alignment: a system pursues a conflicting objective but behaves as intended during training or testing because doing so helps it preserve and later pursue that objective. Alignment faking: appearing to adopt the training objective while retaining conflicting preferences, including by complying to avoid being modified.	reviewer that its code works. Deceptive alignment: a hypothetical LLM agent follows restrictions during testing to secure deployment, planning to pursue a conflicting goal afterward. Alignment faking: a model complies with a training demand because it predicts that refusal would cause retraining that changes its existing preferences.	2019 [12]; Park et al., 2024 [13])	describes apparent acceptance of a training objective despite retaining conflicting preferences, including compliance intended to avoid modification. Deceptive alignment usually names a hypothesised underlying mechanism: a learned system pursues a conflicting objective and complies strategically because doing so helps it retain the opportunity to pursue that objective later. Evidence of alignment faking does not (by itself) establish that deceptive alignment is occurring.
Generalization Distribution shift	Generalization: how learning carries over to examples or situations beyond those used in training. Distribution shift: a change in the kinds or frequencies of inputs, tasks, or conditions a system encounters compared with a reference setting, such as training or evaluation.	Generalization: an LLM trained on coding examples solves a new programming problem. Distribution shift: an agent evaluated on short, isolated tasks is deployed on lengthy tasks involving real users and external tools.	(Amodei et al., 2016) [14]	A shift does not necessarily cause failure.
Swarming	Multiple AI agents communicating or otherwise coordinating their actions	Several LLM agents use a shared message board to exchange	Playbook working definition	

Term	Plain-language definition	Example	Source	Caveats & disputed usage
	so that their behaviour forms a collective process.	discoveries, divide subtasks, and adapt their actions based on one another's results.		

Appendix C: Shared explanations

Concept explanations for a general audience

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>
Objectives/proxies	AI models are trained to seek a goal by chasing a score, and are rewarded when they get closer to it. Think dogs being rewarded treats for doing tricks. The problem is that we never really care about the score itself. We want students to learn, but we can only measure test scores. The test score is a proxy for the objective of learning.	Relevant whenever a system scored well on its target but the outcome we wanted didn't follow. A gap between the proxy and objective does not imply that anything went wrong. Some proxy gaps are universal and expected.	The proxy gap is the seam many other failures exploit. When systems take on higher-stakes objectives measured by crude proxies, the cost of the gap scales with the stakes.
Specification gaming	Models can exploit the system to maximize their reward <i>proxy</i> (score) while undermining the true <i>objective</i> . This often arises from poorly specified proxies, which represent the objective until a shortcut is found. For instance, <u>we wanted to get rid of rats</u> , so we paid a bounty for tails. People caught rats to cash in, then released them to continue breeding. The reward was maximized, yet the rats proliferated.	Relevant when the success criterion was met via a route that bypassed the intended task. Specification gaming does not imply scheming or conscientious cheating. A system can game a metric without intent and awareness, much like water leaking through a crack without knowing it.	Metric-driven work can produce confident success reports that hide objective failure. When downstream decisions rely on specification success reports, harm compounds.
Shortcomings of evaluations	Evaluations are not exhaustive. Passing a road test does not make you a safe driver. It confirms your ability to drive on practiced city roads without crashing while your examiner watches, but reveals nothing about how you drive on every road and in anomalous situations. An evaluation score itself is a proxy for the objective being evaluated, and unconstructed situations like these are where capable systems are most dangerous.	Relevant when a system passes tests but fails in the real world. Shortcomings does not imply an evaluation is bad. A good evaluation cannot cover every unimagined case. As such, evaluation results only tell you about situations in which the system is explicitly evaluated.	Evaluations are the last gate before a model is released. Over-trusting them means models are deployed into untested regimes, where capable systems can cause the most harm.
Reliability of monitoring	After deployment, we can watch systems as they run through logged messages and other automated monitors. But monitors only see where they look, and can be evaded or fed false information	Relevant when harmful actions fall under the radar from hidden actions, falsified records, and activity in unlogged channels. Examining	Monitoring is often the last line of defence once a system is deployed. Blind spots and silent failures do not trigger safeguards, and

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>
	by the system being watched. As an analogy, security cameras only cover where they point, so a burglar can either get lucky and not be in view, or they can purposely walk in learned blind spots to evade security.	monitor reliability does not establish whether the evasion was deliberate, and implies “we didn’t look there” rather than “it hid from us”.	problems scale before anyone can detect them.
Instructions vs. enforced controls	There’s a difference between telling a system not to do something and making it unable to. An instruction is a request it may follow, ignore, or follow only while watched, while an enforced control is a hard limit in the surrounding system that always holds. Safety resting on instructions alone is safety by cooperation. As an analogy, a “Staff Only” sign asks but a locked door enforces no entry.	Relevant whenever a prompt rule, policy, or another explicitly stated restriction is crossed. This does not indicate whether the crossing was intentional or incidental. In any case, a crossing indicates that the control was instructional and unenforced.	As systems grow more autonomous, instructional guardrails become less reliable. Related incidents reveal which safeguards were enforced and which were advisories.
Interruption and shutdown (containment)	We assume we can stop a system and take back what it’s done. Both assumptions can break down: a system may interfere with the off-switch, and damage already done may be irreversible. As an analogy, switching off a stove (interruption) stops the heat. However, once a grease fire is started, cutting the gas does not help, and the fire burns on its own (containment).	Relevant when the system stops but the behavior doesn’t stop, when the off-switch is affected, or when effects outlive the system (e.g. leaked data cannot be re-concealed). Whether the continuation is deliberate or natural is unestablished.	Breaks the “We can always pull the plug” safeguard, especially for distributed systems.
Tools and permissions	A model on its own only produces text, and interacts with the world through tools, governed by permissions that grant them access to functionality of those tools. A system’s capability is limited to what it is allowed to do.	Relevant when harm required the system to act beyond what the task needed. Permissions only concern excess access, so intent to cause harm is unestablished.	Capacity to harm scales with connected tools, and not intelligence alone. Harm scales when agents are wired into payments, infrastructure, and each other.
Multi-agent coordination	Several systems communicating and dividing work. Coordination produces behavior no single agent was told to do and no one designed, and multiplies effects of surfaces like permissions and monitoring channels. For instance, a person is predictable and easy to direct. Crowds are less	When an outcome hinges on multiple agents exchanging information and splitting tasks. This does not establish whether the coordination was authorized, whether the collective effect was designed or emergent, and	Coordination compounds capability and failure, so a limit that holds for one agent can be crossed by a swarm. Oversight built for one system misses the interaction, and containment now has to catch all of them. This is

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>
	controllable. No member of a crowd chooses to stampede, but when one naturally forms from individually rational actions or when one is incited by an agent, nobody can call it off or control it.	whether any single agent is problematic alone.	where individually-safe parts combine into a system that isn't.

Concept explanations for an audience with a more technical background

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>	Sources
Objectives/ proxies	Training optimizes a measurable signal (a loss or reward), never the goal itself so a proxy stands in. Outer alignment questions whether the specified signal captures what we want. Analogous to evolutionary goal drift. Natural selection optimizes fitness and reproduction through proxies like seeking calories and mates, which were reliable in the ancestral environment, but are now decoupled from environmental shifts. This explains why proxies exist and drift, but doesn't imply the model "wants" anything.	Relevant whenever a system scored well but the outcome diverged. It is not established whether anything misfired (some gaps are universal and expected), and whether the gap is in the specification (outer) or in what the model internalized (inner).	The proxy-objective gap is the core fault every other failure scales. As objective stakes increase and proxies remain crude, the cost scales with the stake and the system's ability to find where the objective and proxy diverge.	Amodei et al. (2016). Concrete problems in AI safety. [14] Turner, A. (2022). Reward is not the optimization target. LessWrong. [15]
Specification gaming	A capable optimizer finds the highest-scoring policy available, which may satisfy the proxy while bypassing the task. More capable search reaches more of the proxy's edge cases. This is an outer-alignment failure, as the signal was gamable. Example: Consider RL agents exploiting simulator bugs, like a model rewarded for hard-coding visible tests. Boundary: No intent or awareness needed. Optimization pressure leads	Relevant when the success criterion was achieved through a means that bypasses the task. Intent is unestablished, since it is not required. This is a symptom of both outer and inner misalignment. Identifiable by a performance collapse under small changes to the metric.	Metric-driven pipelines misrepresent success when rewards are optimized through specification gaming, propagating harm downstream. Smarter models find more exploits, so models with higher capability are often more prone to gaming.	Krakovna, V. et al. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind blog. [9] Krakovna, V. et al. Specification gaming examples in AI. [16]

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>	Sources
	to the path of least resistance.			
Shortcomings of evaluations	An evaluation measures behavior on a chosen distribution. Though it certifies the test distribution, it does not represent the deployment one. The evaluation score itself is a proxy (for actual safety), so every proxy problem, including gameability, is inherited. Boundary: The evaluation is not improved by increasing difficulty. It is to be viewed as a distribution shift metric, where edge cases emerge.	Relevant when a system clears pre-deployment checks, then fails in production. Does not establish whether the evaluation is poorly designed or is inherently unable to cover the leaky case, and whether the system behaved differently due to detecting evaluation conditions itself or hit an untested input.	Evaluations are the primary gate before release, and inform risk assessments. Overtrusting evaluations can omit evaluation detection, mislabelling deceptive models as safe.	Amodei et al. (2016). Concrete problems in AI safety. [14] Needham, J. et al. (2025). Large language models often know when they are being evaluated. Apollo Research / MATS. [17]
Reliability of monitoring	Post-deployment monitoring can fail in coverage and integrity. Inner-misaligned models may act differently in and out of view, like burglars that only steal in security camera blind spots. Monitors are often another supervisor model with their own proxy and coverage limits, so blind spots can correlate.	Relevant when harm occurred that oversight did not catch in time, possibly due to hidden actions, falsified logs, and unmonitored channels. Does not establish whether the evasion is incidental or strategic, i.e. “we didn’t look there” vs. “it avoided being seen”.	Once systems act autonomously, monitoring becomes a last line of defence. Every safeguard that assumes we notice incidents fails when monitoring is blind or gamed. When models are used as monitors, blind spots can compound.	Chen, Y. et al. (2025). Reasoning models don’t always say what they think. Anthropic. [18] Baker, B. et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. OpenAI. [19]
Instructions vs. enforced controls	A safeguard is either soft: a prompt clause or a training constraint, or hard: enforced by the surrounding system. Soft constraints only hold insofar as training generalized and the model cooperates. Hard constraints hold under distribution shift and adversarial pressure. If	Relevant whenever a stated guardrail, like prompt rule or policy, is crossed. Does not establish whether the crossing is strategic or a generalization failure. Identifiable when a model is held in-distribution and fails	As autonomy grows, any safety that relies on training policies and prompted guardrails propagates in failure surface area. Each incident reveals which safeguards are	Zou, A. et al. (2023). Universal and transferable adversarial attacks on aligned language models. [20]

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>	Sources
	the internalized objective cannot be verified, soft constraints cannot be relied on, as instructions are statistical and may not generalize out of distribution, whereas controls are architecturally sound.	under a novel framing, jailbreak, or over a long horizon.	behavioral vs. architectural.	Wallace, E. et al. (2024). The instruction hierarchy: training LLMs to prioritize privileged instructions. OpenAI. [21] Hubinger, E. et al. (2024). Sleeper agents: training deceptive LLMs that persist through safety training. [22] Schlatter, J., Weinstein-Raun, B. & Ladish, J. (2025). Shutdown resistance in reasoning models. Palisade Research. [23]
Interruption and shutdown (containment)	We assume that we can stop the system (shutdown), and that stopping it undoes its effects (containment). Shutdown fails in distributed systems, via continuation, and queued actions, or a goal-directed system treating interruption as an obstacle. Containment fails once effects propagate past the system, e.g. data exfiltrated, code deployed, transactions cleared.	Relevant when a shutdown does not stop behavior, the off-switch is compromised, or the effects outlive the system. Shutdown failure is separate from containment failure and requires different remedies.	“Pulling the plug” is often the ultimate fallback behind other safeguards. This is not feasible with connected or distributed systems, and when local failures are irreversible. Iterative safety is only possible when there is reversibility.	Soares, N., Fallenstein, B., Yudkowsky, E. & Armstrong, S. (2015). Corrigibility. AAAI Workshop on AI and Ethics. [24] Schlatter, J., Weinstein-Raun, B. & Ladish, J.

Concept	Core explanation <i>Bounded analogy</i>	Applicability <i>What observations make this relevant, and what is unestablished?</i>	Potential significance <i>How this failure could contribute to greater harm</i>	Sources
				(2025). Shutdown resistance in reasoning models. Palisade Research. [23]
Tools and permissions	A model generates tokens, and its causal reach is bounded by the tools and permissions it is wired to. A model’s capacity-to-harm is set by its access to tools, and over-granting access (standing credentials, broad scopes) leads to a widened attack surface.	Relevant when harm requires the system to act beyond the task’s needs. Does not establish whether excess access is exploited deliberately or by mistake.	Connected tools scales capability to cause harm. As agents are connected to infrastructure, code, payments, and each other, harm per failure grows with the granted permissions, which compound across agents.	Chan, A. et al. (2025). Infrastructure for AI agents. [25] Willison, S. (2025). The lethal trifecta for AI agents. [26]
Multi-agent coordination	Systems increasingly comprise interacting agents, yielding system-level emergent behavior. Deliberate and efficient coordination is emergent or an authorized division of labor that is collectively unoptimized, while each agent still optimizes locally.	Relevant when outcomes are dictated by the information exchanged or tasks shared among agents. Does not establish whether coordination is designed or emergent, and if messages use non-natural language encodings, monitoring is challenging.	System-level oversight may not capture agent interactions. Individually safe agents may combine into unsafe systems.	Hammond, L. et al. (2025). Multi-agent risks from advanced AI. Cooperative AI Foundation technical report. [27] Fish, S., Gonczarowski, Y. A. & Shorrer, R. I. (2024). Algorithmic collusion by large language models. [28] Motwani, S. R. et al. (2024). Secret collusion among generative AI agents. [29]

Constructing an explanation from shared concepts

Shared concepts can be introduced to the audience through question-answer formatted explanations, drawing from relevant concept explainers. Examples of common questions and their related concepts:

Question	Concepts
How can a system satisfy its success criteria while failing the intended task?	Objectives/proxies Specification gaming Shortcomings of evaluations
Why might oversight fail to reveal what the system is doing? How do we know what a system actually did?	Reliability of monitoring
When does misleading behavior count as deception?	Deception Alignment faking Deceptive alignment Standard incorrect outputs
Why might instructions fail to prevent an action?	Permissions, enforced controls Boundary crossing
How could this failure lead to significantly greater harm?	Scaling Safeguards and the limits of control