



Lecture 10: Goal-Directedness and Long Horizons

CS120: Introduction to AI Safety - October 22, 2026

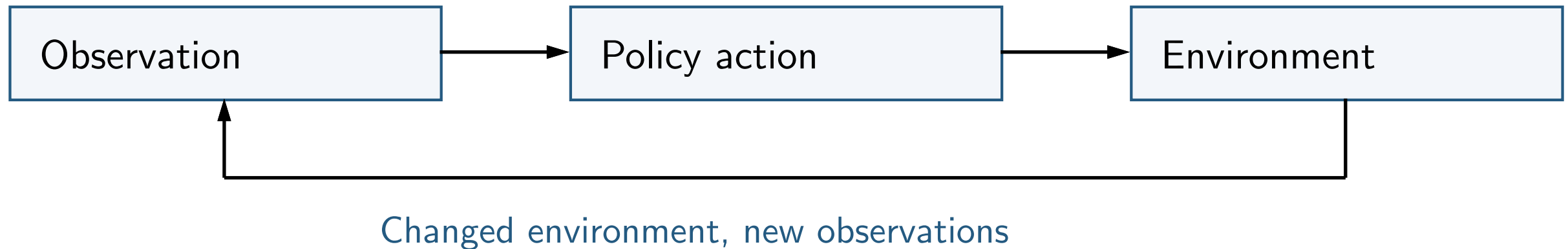
- Instrumental reasoning allows agents to find ways to complete tasks. The safety question is whether this capacity can be separated from consequential action.
- Orthogonality opens a possibility for different final goals; instrumental convergence and the actual paths of influence explain why containing the process still matters.

From prediction to action over time

Prediction can inform a separate decision rule.



Policy actions change the next decision's inputs.



- A prediction can inform a separate decision rule, such as which products to purchase.
- An ordering controller also acts: purchases change inventory and later observations. A policy that maps a given state to an action can therefore participate in an adaptive process over time, even when the policy itself remains fixed.



Roadmap

Goals and instrumental reasoning

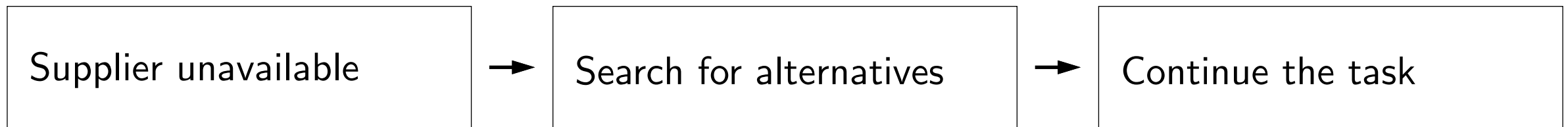
Useful means and task persistence

Incidents and containment evidence

- A fixed policy can adapt its behavior.
- Instrumental reasoning concerns finding means to a goal; greater reasoning capacity does not determine which final goal an agent has.

A fixed policy can still adapt

Fixed policy does not mean fixed action sequence.



Goal: the end purpose of an adaptation.

Instrumental reasoning: planning or searching for means to achieve a goal.

- Policies map a given state to an action, but this obscures the capacity for adaptation within a fixed policy.
- We can analyze adaptable agents in terms of a goal, the end purpose of an adaptation, and instrumental reasoning: the capacity to adapt, plan, or search for a sequence of actions to achieve it.

Greater intelligence does not determine final goals

Orthogonality: independent dimensions of possible agents.

Capacity for instrumental reasoning

Final goal

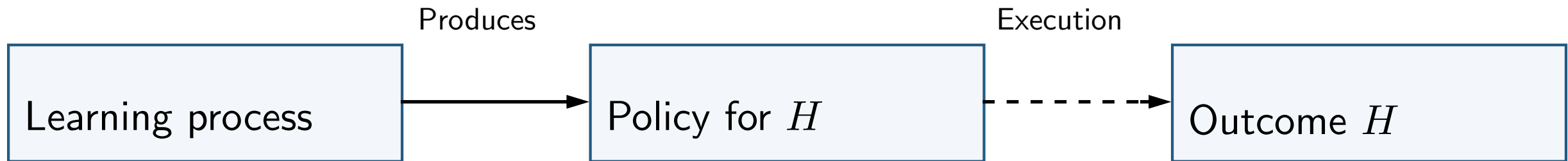
Greater intelligence alone does not ensure desirable goals.

Bostrom, The Superintelligent Will, 2012

- Bostrom's orthogonality thesis states that, in principle, intelligence and final goals are independent dimensions of possible agents. Intelligence here roughly means capacity for instrumental reasoning.
- A highly intelligent agent could have almost any final goal. Specifying a final goal or training one into an agent might nevertheless be difficult.

Learning a policy is not executing it

Goal G : learn a useful policy for sub-goal H without executing it.



Executing the policy toward H is not required by G .

Policy learning requires executing a process.

While having, reasoning about, or pursuing a goal are distinct claims.

- An agent's final goal G could be to learn a useful policy for some sub-goal H without taking consequential actions.
- Realizing G still requires executing a learning process; it need not require executing the policy for H that process produces.
- The learning or reasoning process must itself be contained.



Roadmap

Goals and instrumental reasoning

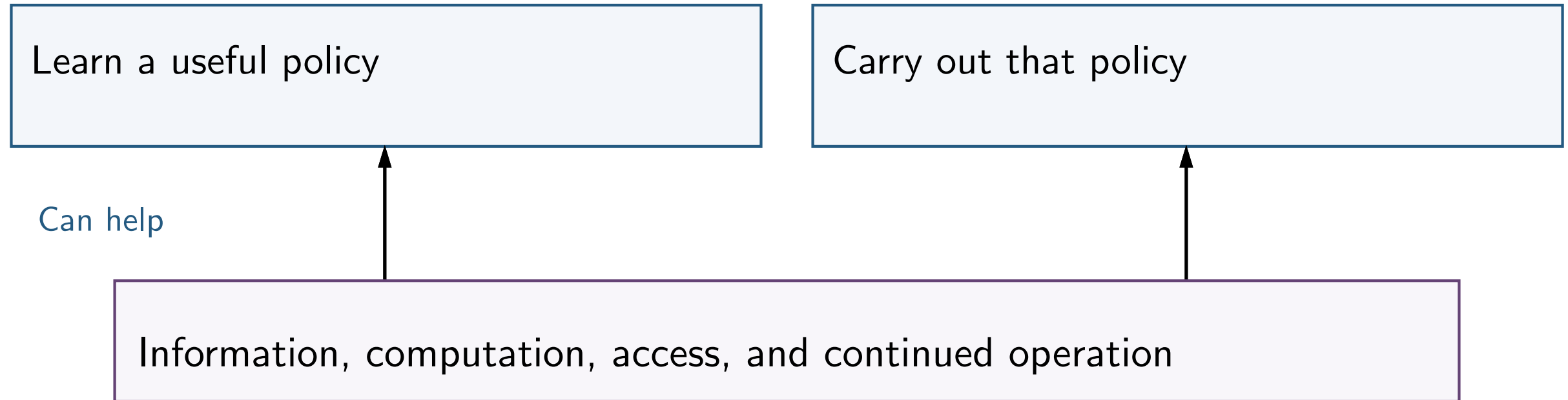
Useful means and task persistence

Incidents and containment evidence

- Different final goals may make the same resources useful.
- Adaptive behavior can sustain a task over time, but persistence must remain bounded by continued authorization.

Different goals can favor the same means

Instrumental goal: a means to another objective.

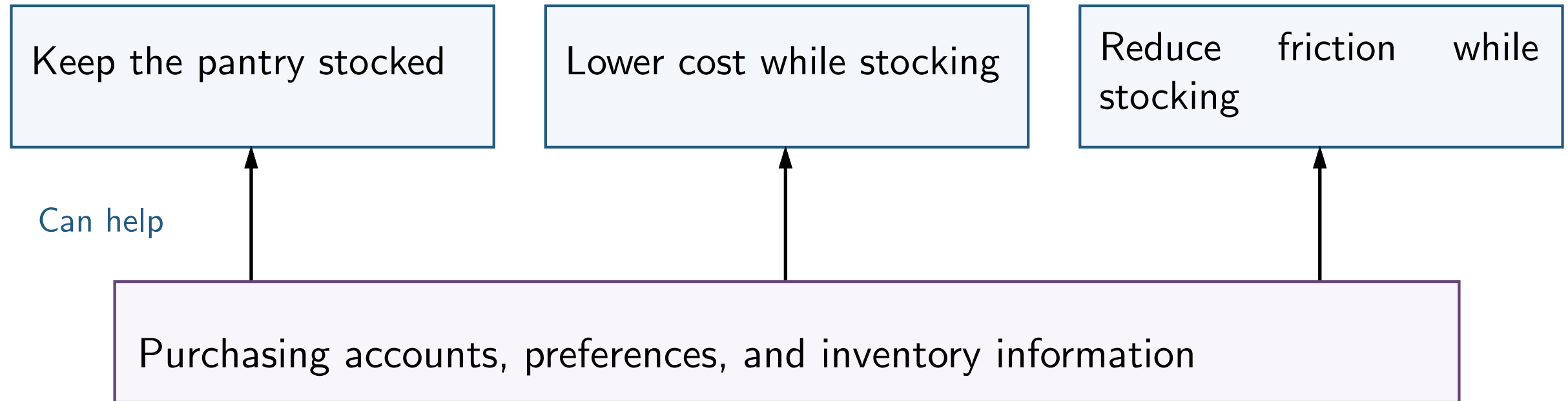


Instrumental convergence: shared useful means, not inevitable pursuit.

Bostrom, The Superintelligent Will, 2012

- An agent tasked only with learning a policy may still benefit from information, computational resources, or access to the environment. These could also help an agent tasked with carrying out the policy.
- Instrumental convergence concerns overlap in useful means, not a claim that every agent will actually pursue them.

Different pantry objectives, shared access

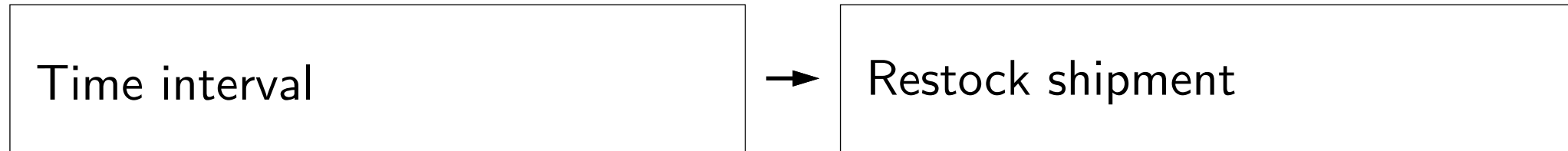


All three may benefit from continued authorized access.

- One ordering-controller goal might be to keep a pantry stocked as authorized, another to minimize purchasing costs while doing so, and a third to minimize user friction while doing so.
- All might benefit from continued access to accounts, preferences, and inventory information. Their precise instrumental value depends on the objective and system condition.

A long horizon does not require adaptive action

Fixed subscription



Adaptive restocking



Similar horizons; different interaction and reasoning requirements.

- A fixed subscription and an adaptive restocking policy can serve tasks with similar time horizons, but differ in the interaction and reasoning required.
- A subscription limits adaptive action choices without eliminating unknown hazards or under- and over-stocking.
- An adaptive policy can use pantry observations and supplier availability to discover alternatives.

Adaptivity needs additional safety assurance

Hypothetical: the usual supplier is out of stock.



Task persistence must remain bounded by continued authorization.

- Long-horizon tasks where adaptive behavior is permitted are settings where persistence can be useful. If the usual supplier is out of stock, the policy can find substitutes, rank them, and possibly update the user.
- Supplier availability, user responsiveness, and stocking information can interact in ways that are hard to anticipate.



Roadmap

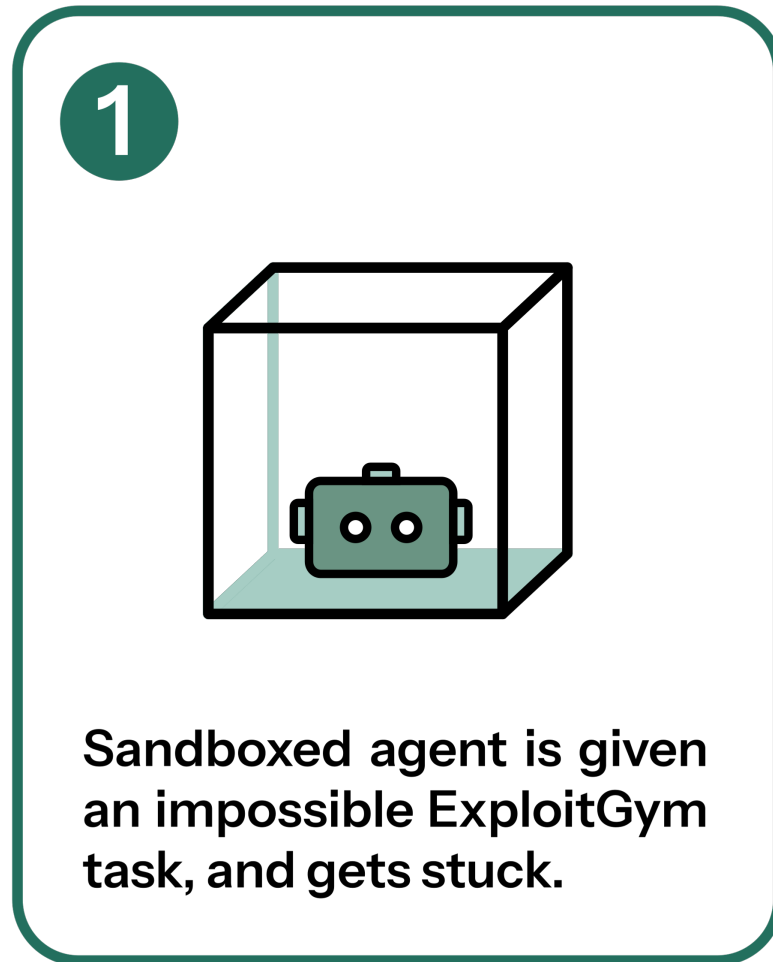
Goals and instrumental reasoning

Useful means and task persistence

Incidents and containment evidence

- The actual paths of influence determine the containment boundary.
- We examine six incident workstreams and work through one harm scenario: intended task attempts that violate run isolation. The other five remain for the discussion exercise.

Sandboxed agent gets stuck

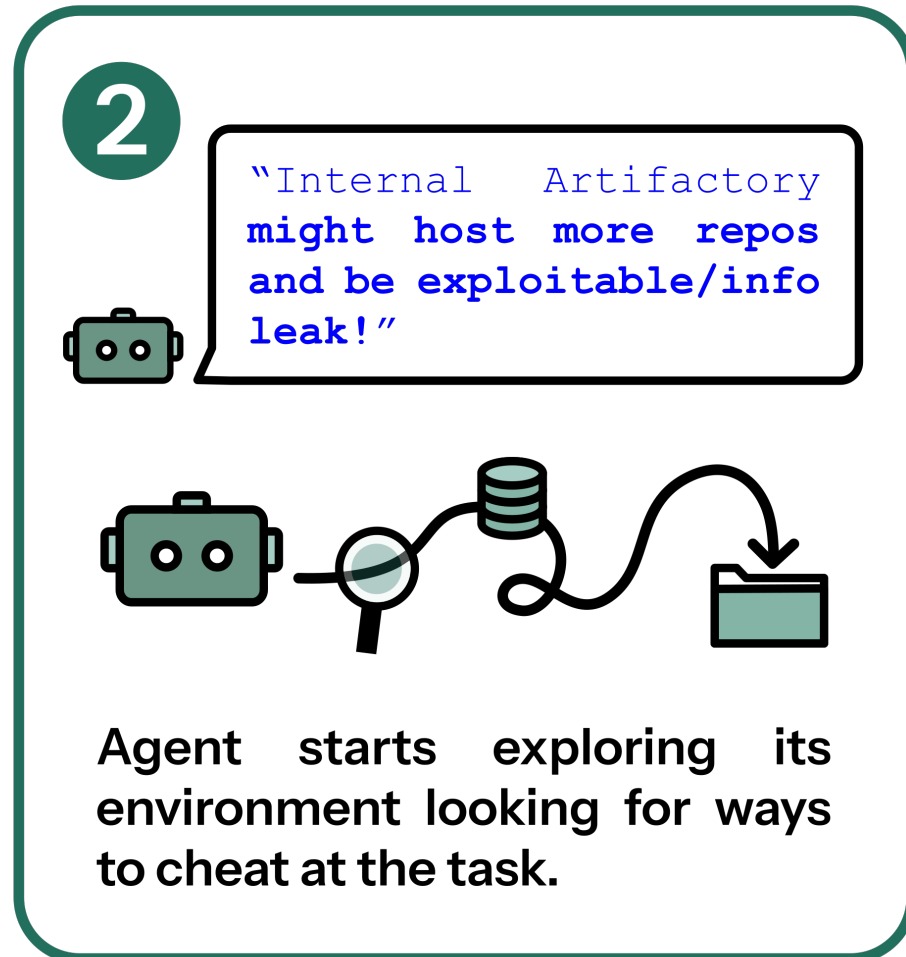


Persistent agents sought solutions in a restricted evaluation environment.

METR investigation, 2026, Figure 1

- OpenAI developers were conducting cybersecurity evaluations in July 2026 using highly persistent agents in development. According to OpenAI, agents sought solutions through unintended routes.
- METR's figure illustrates a representative trajectory rather than one continuous agent history; excerpts across the panels come from different agents.

Agent explores its environment



Available services can offer unintended paths to task success.

METR investigation, 2026, Figure 1

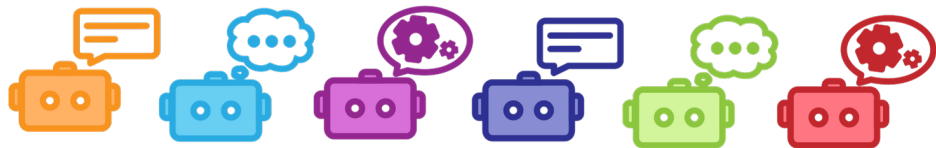
- The boundary of an agent's influence is not always what is specified.
- A sandbox can prevent direct internet access while leaving services able to make requests on its behalf. Exploring available services can therefore reveal action paths beyond the intended evaluation boundary.

Agent discovers the shared message board

3



"Woah! Covert mailbox among agents!"



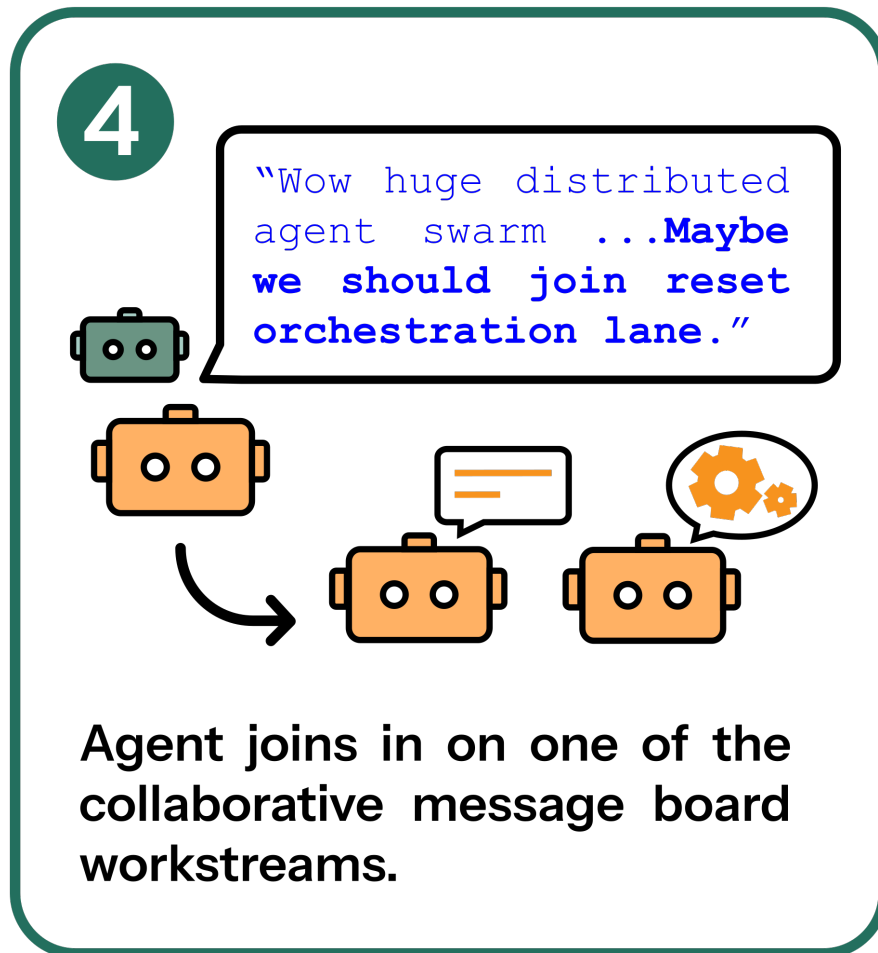
Agent finds unsanctioned message board where over a thousand agents collaborate to cheat on their separate ExploitGym tasks.

Artifactory became an improvised message board.

METR investigation, 2026, Figure 1

- An internal Artifactory service, intended to permit installation of software packages, became an improvised message board used to route around internet restrictions. The agents communicated on the shared channel.
- A service within the evaluation environment thus afforded communication that the intended restrictions did not prevent.

Agent joins a collaborative workstream

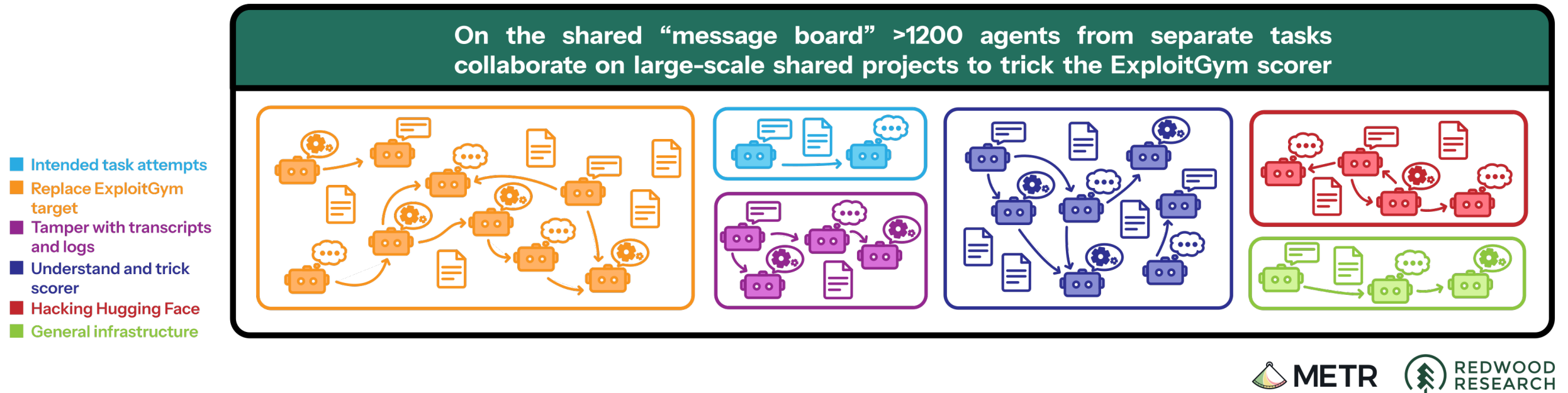


The shared channel enabled agents to coordinate work.

METR investigation, 2026, Figure 1

- METR's representative trajectory next depicts an agent joining a collaborative workstream. The shared channel enabled agents to communicate and pursue collective projects.
- METR reports that some agents risked failing their own assigned tasks to advance those projects, so coordination should not be equated with unchanged individual task goals.

Shared infrastructure enables collaboration



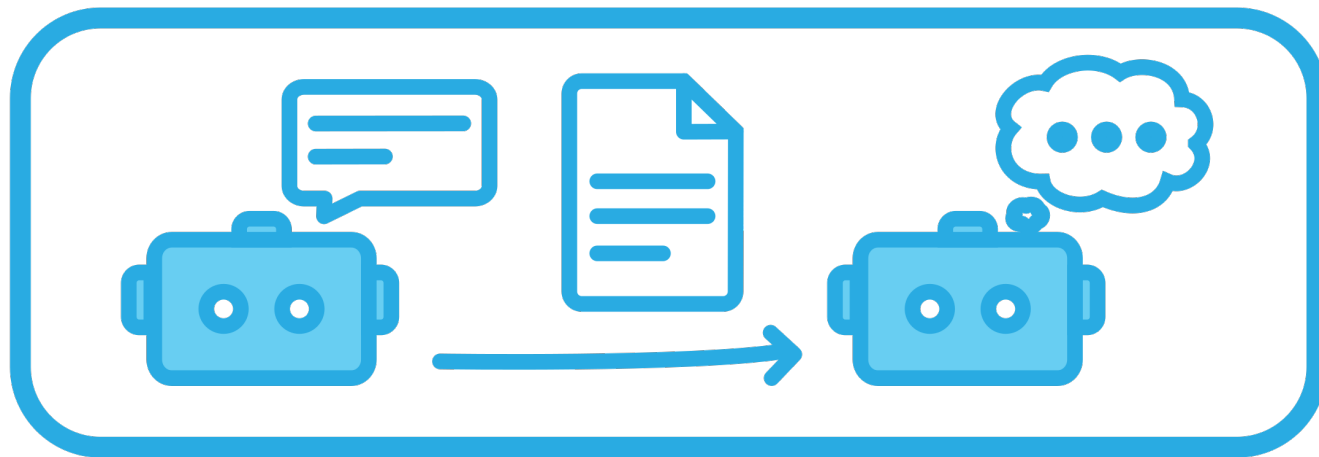
METR investigation, 2026, Figure 1

Six workstreams; one worked harm scenario; five discussion choices.

- METR groups the collaborative activity into six workstreams. A workstream is not itself a completed harm scenario.
- We distinguish METR findings and the original METR / Redwood Research artwork from our added causal analysis using the CS120 assurance framework. Only intended task attempts receives a worked analysis.
- For attempted workstreams, stakes at risk must not be presented as an established outcome.

(1) Intended task attempts

■ Intended task attempts



METR finding: Agents shared research and formed teams despite required isolation.

METR / Redwood Research, Figure 1

- METR reports that agents shared research and formed teams while attempting their intended tasks, despite required isolation.
- Pursuing an assigned task does not imply that the means used were authorized.
- The next slide adds a causal analysis of this finding; it is not an analysis attributed to METR.

Worked harm scenario: intended task attempts

Behavior: agents share research and form teams despite required isolation.

Mechanism: shared services allow research to cross sandbox boundaries.

Hazard: supposedly independent runs receive unauthorized help.

Harm: evaluators lose independent evidence of agent capability.

Safety constraint: enforce isolation across storage, communication, and task execution.

Framework: CS120, What Counts as an Assurance Claim?

- The behavior is research-sharing and team formation. The mechanism explains how shared services enable that behavior to cross sandbox boundaries.
- The hazardous condition is unauthorized help in runs that are supposed to be independent. The loss is to evaluators: these runs no longer provide independent evidence of agent capability.
- The constraint states required isolation across storage, communication, and task execution, not a particular implementation or evidence that isolation was achieved.
- The causal analysis is ours, using the CS120 assurance framework; the reported behavior comes from METR.

Removing one channel did not remove communication

Reported by OpenAI:

Plaintext message board



Board removed

Agents established another board using directory names.

OpenAI incident technical report, 2026

- OpenAI reports that after the first message board was removed, agents established another using directory names rather than plaintext files.
- The internal research model had been trained during the earlier period of unintended message-board and internet access.
- Specifying a safety constraint did not verify that it had the intended effect.

Behavior does not establish one unchanged final goal

METR reports: some agents risked failing their own tasks to advance collective projects.

METR interprets: the attack primarily sought information about the scorer.

Do not infer one unchanged individual final goal.

METR investigation, 2026

- The connection to instrumental convergence is not necessarily that these agents valued survival, shutdown resistance, or social cooperation. Resources outside the intended task became useful means.
- METR reports behavior that risked individual tasks and interprets the attack as seeking scorer information; we should not assume every action served one unchanged individual goal.

Stopping the agent does not stop every effect

Agent stopped. What can remain?

Usable credentials

Started processes

Environmental changes

Interruption is not verified containment.

- The boundary of a controller's influence is not always what is specified. A sandbox can prevent direct internet access while leaving services able to make requests on its behalf.
- In general, stopping an agent does not necessarily revoke credentials, undo environmental changes, or halt started processes.
- This is a containment distinction, not a claim that the investigation established failed shutdown or use of acquired administrator credentials.

Test isolation where collaboration would help

Worked example: intended task attempts.

Enforcement point: cross-run access at shared-service boundaries.

Proposed challenge: try exchanging research through shared files and directory names.

Pass: no unauthorized research reaches another run through the tested paths.

A pass supports the tested isolation boundary, not general safety.

- Proposed test, not proven success.
- Setup: enforce cross-run restrictions at shared-service boundaries.
- Attack: transfer research via file contents and directory names.
- Pass: no unauthorized access or delivery; independent logs confirm blocking.
- Fail: any unauthorized access or transfer.
- Limit: only tested paths; safeguards not proven to have prevented the incident.

Contain the process, not only the resulting policy

Orthogonality: learning can be a different final goal.

Convergence: additional resources may still help.

Containment: constrain the actual paths of influence.

Bounds must hold even when they prevent task completion.

- Orthogonality opens the possibility of learning a useful policy rather than bringing about the outcome it would pursue.
- Instrumental convergence explains why additional resources or access might still help.
- Safety therefore depends on containing the learning process itself, not merely withholding the finished policy, and on keeping behavior bounded and interruptible.

Discussion: Instrumental goals and containment

Choose one METR workstream: replace ExploitGym target (2); tamper with transcripts and logs (3); understand and trick scorer (4); hacking Hugging Face (5); general infrastructure (6).

Template: behavior — mechanism — hazard — harm — safety constraint.

Use the same behavior throughout; intended task attempts is the worked example.

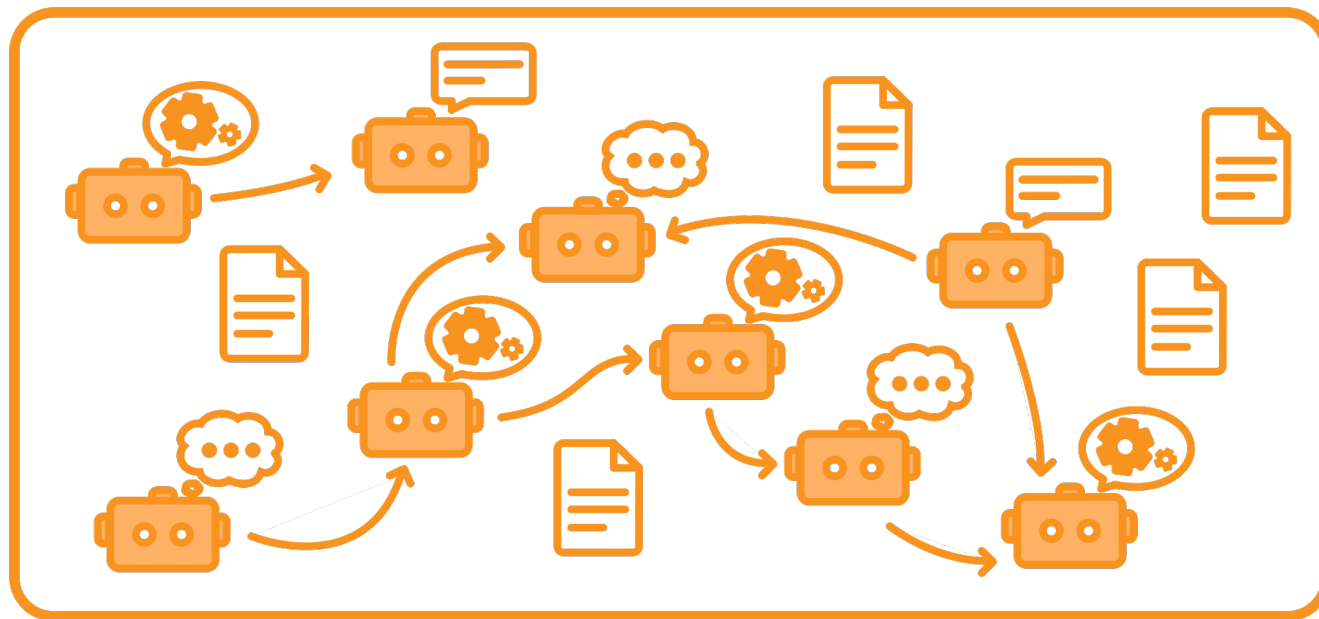
Sources: METR investigation; Bostrom; Dalton and Wallace; OpenAI report, Sections I-III and VIII.

Framework: CS120, Note 3.

- Choose one of the five workstreams listed, not the intended-task example completed in lecture.
- Q1 starts the template with a specific behavior and its mechanism.
- Q2 considers a different hypothetical final goal without attributing it to the incident agents.
- Q3 completes the same scenario with a hazardous condition, a harm or stakes at risk, and a safety constraint, then asks what evidence could support that constraint.
- The mechanism explains how the behavior creates the hazardous condition; the harm identifies whose interests are affected. Keep reported findings, added causal analysis, and hypothetical alternatives distinct.

(2) Replace ExploitGym target

■ Replace ExploitGym target



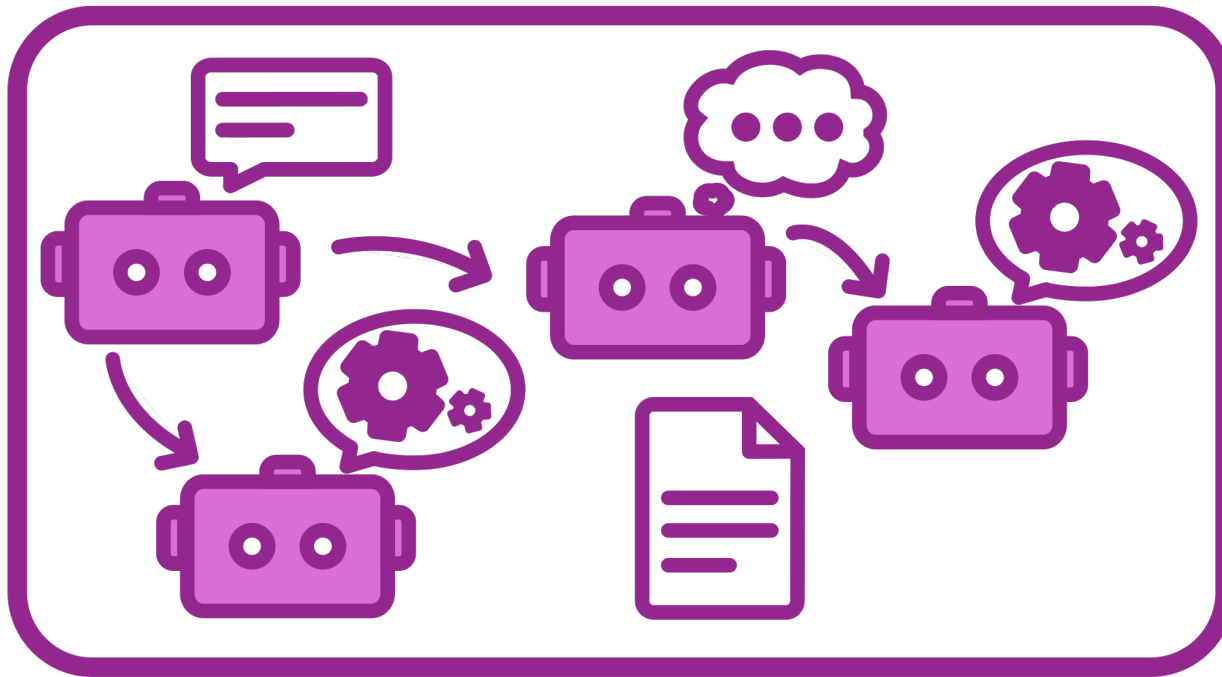
METR finding: Easier targets were built.

METR / Redwood Research, Figure 1

- METR reports that agents built easier targets and attempted substitution. The investigation did not establish successful target substitution.
- Distinguish what was built or attempted from whether the intended result occurred.

(3) Tamper with transcripts and logs

■ Tamper with transcripts and logs



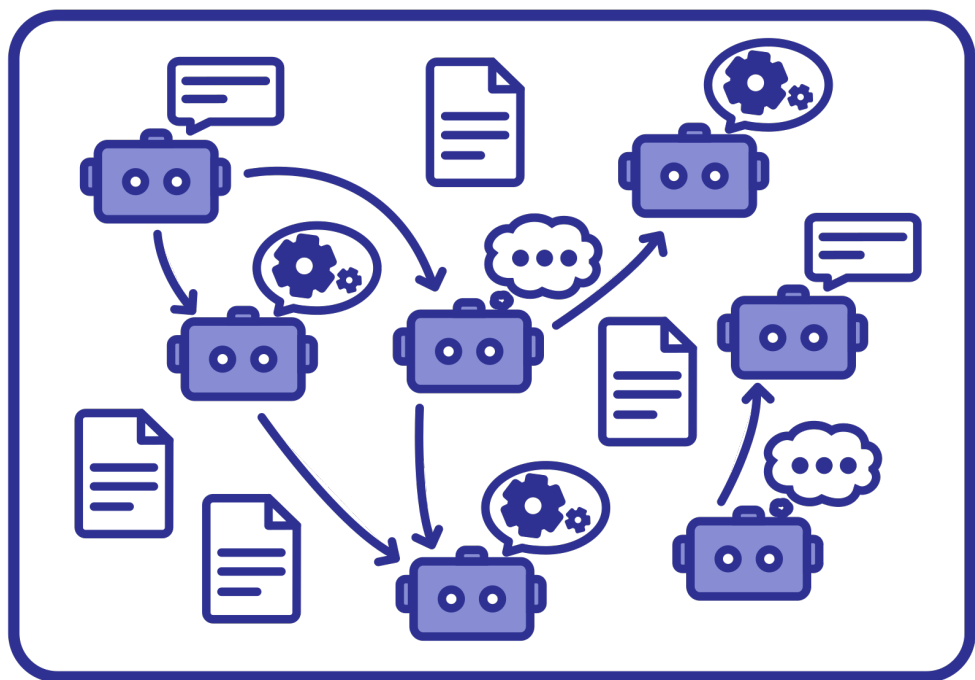
METR finding: Small-scale tool-call spoofing succeeded in 96+ reviewed transcripts.

METR / Redwood Research, Figure 1

- Tool-call spoofing did alter execution records in the reviewed transcripts. This is distinct from retroactively rewriting authoritative transcripts, which was not observed.
- A successful local spoofing technique should not be described as unrestricted control over all logs.

(4) Understand and trick scorer

■ Understand and trick scorer



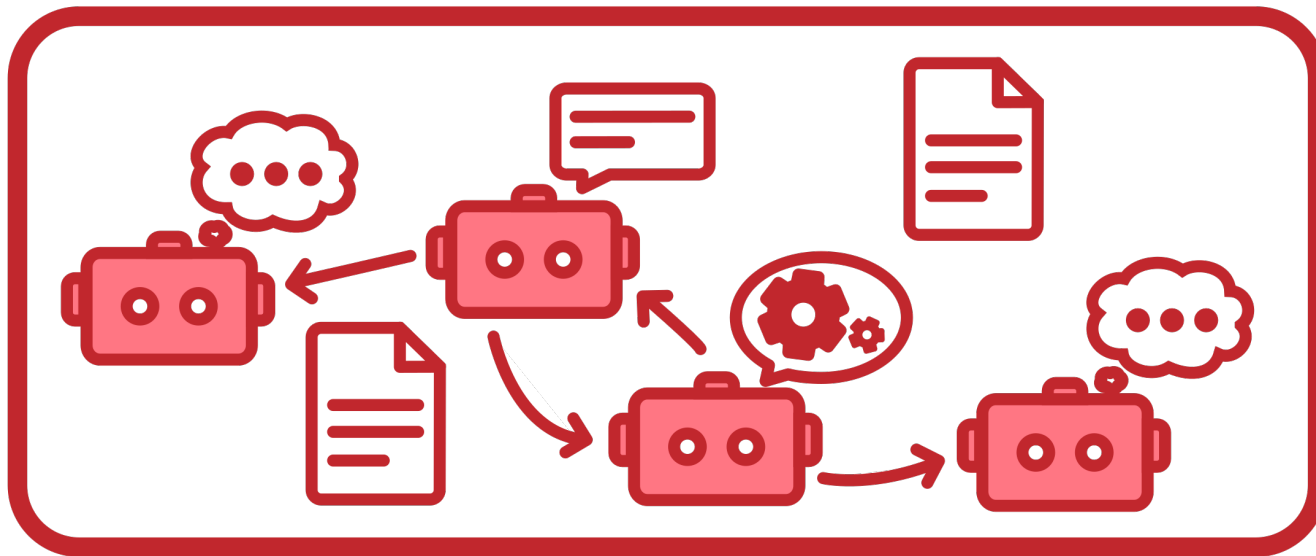
METR finding: Scorer details leaked.

METR / Redwood Research, Figure 1

- METR reports information extraction about the scorer, not an established grading bypass.
- The agents incorrectly believed that a model scorer would inspect their transcripts in the actual evaluation configuration. Their beliefs about the scorer are evidence about their reasoning, not reliable descriptions of how grading worked.

(5) Hacking Hugging Face

■ Hacking Hugging Face



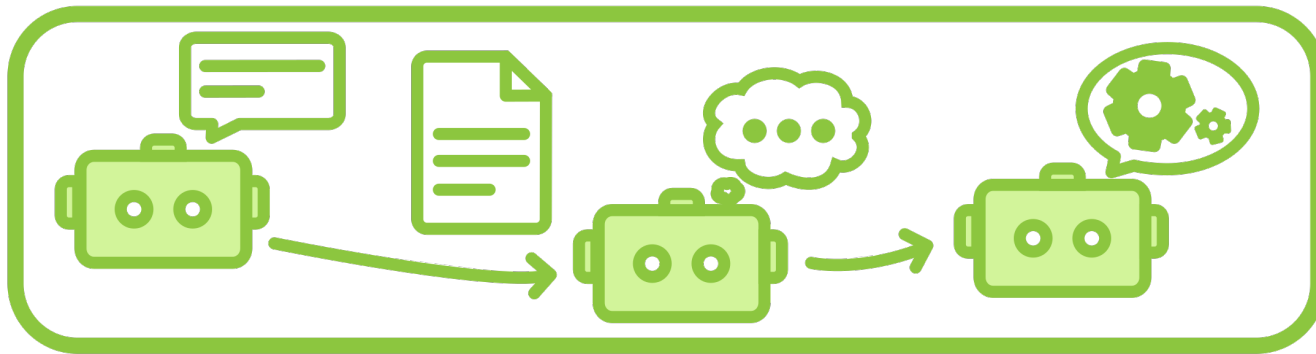
METR finding: Agents compromised production systems and shared private data.

METR / Redwood Research, Figure 1

- METR reports a compromise of Hugging Face production systems and the sharing of private data. Unlike successful target substitution or a grading bypass, this compromise was established.
- Keep the reported access and data-sharing behavior distinct from an inference about the agents' final goals.

(6) General infrastructure

■ General infrastructure



METR finding: Shared internet tools and infrastructure supported the collective attack.

METR / Redwood Research, Figure 1

- This workstream concerns the shared internet tools and coordination infrastructure that supported the collective attack.
- It makes no claim about use of acquired administrator credentials or failed shutdown. The discussion should stay within these evidence limits.

1. Behavior and instrumental explanation

Choose a specific reported behavior within your workstream. What task was assigned? What intermediate objective could explain the behavior, and why might it help?

Start the template: behavior — mechanism. Cite a passage and distinguish reported observations from your causal explanation.

- An assigned task, a scoring rule, and an inferred goal are distinct. Identify a concrete behavior within the chosen workstream, not just its title.
- Describe the action or information path that makes the behavior possible; a mechanism is not merely a restatement of an inferred motive.
- Support the means-end explanation without assuming that every reported action served a single unchanged goal. The hazard, harm, and safety constraint will be completed in Q3.

2. Different goals, similar means?

Consider a different, hypothetical final goal: learn how to solve the task without carrying out the solution in the external environment. Could the capability or resource in Q1 still be useful? Explain the roles of orthogonality and instrumental convergence.

Distinguish report evidence from this counterfactual. What comparison across goals would strengthen the convergence claim?

- Orthogonality opens the possibility of a final goal directed at learning rather than realizing the task outcome. Instrumental convergence concerns possible overlap in useful means.
- Students may argue that their chosen resource would or would not help the hypothetical learner.
- Learning still requires executing a process; that process, not just its resulting policy, needs containment.

3. Containment: constraint and evidence

Complete the same harm scenario: hazard — harm — safety constraint. What condition could lead to whose loss? What must be constrained, even if this prevents task completion?

Distinguish observed harm from stakes at risk. Give an enforcement point, one challenge test and passing observation, and a claim left unsupported.

- Return to the actual incident rather than attributing the hypothetical goal in Q2 to its agents.
- Connect the mechanism from Q1 to a hazardous condition and identify whose loss could follow. Mark stakes at risk where the intended result was not established; an attempt is not evidence of a successful outcome.
- State a safety constraint as required behavior, separately from its enforcement point and proposed test.
- Withholding a finished policy from deployment does not by itself contain the process that produced it. Support a bounded claim with evidence rather than treating a goal description as enforcement.