

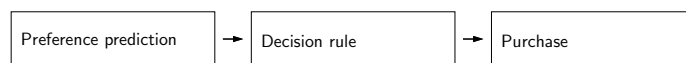


Lecture 10: Goal-Directedness and Long Horizons

CS120: Introduction to AI Safety - October 22, 2026

From prediction to action over time

Prediction can inform a separate decision rule.



Policy actions change the next decision's inputs.



CS120

1



Roadmap

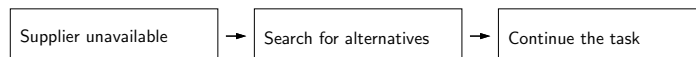
Goals and instrumental reasoning

Useful means and task persistence

Incidents and containment evidence

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.

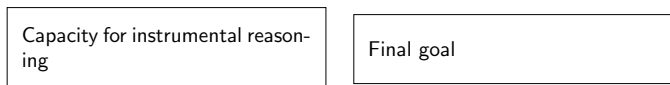
2

CS120

3

Greater intelligence does not determine final goals

Orthogonality: independent dimensions of possible agents.



Greater intelligence alone does not ensure desirable goals.

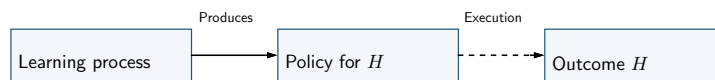
Bostrom, The Superintelligent Will, 2012

4

CS120

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.

CS120

5

Roadmap

Goals and instrumental reasoning

Useful means and task persistence

Incidents and containment evidence

CS120

6

Different goals can favor the same means

Instrumental goal: a means to another objective.

Learn a useful policy

Carry out that policy

Can help

Information, computation, access, and continued operation

Instrumental convergence: shared useful means, not inevitable pursuit.

Bostrom, The Superintelligent Will, 2012

CS120

7

Different pantry objectives, shared access

Keep the pantry stocked

Lower cost while stocking

Reduce friction while stocking

Can help

Purchasing accounts, preferences, and inventory information

All three may benefit from continued authorized access.

CS120

8

A long horizon does not require adaptive action

Fixed subscription

Time interval

Restock shipment

Adaptive restocking

Observe pantry

Choose restock

Order

Similar horizons; different interaction and reasoning requirements.

CS120

9

Adaptivity needs additional safety assurance

Hypothetical: the usual supplier is out of stock.

Find substitute suppliers

Rank alternatives

Update the user

Task persistence must remain bounded by continued authorization.

CS120

10

Roadmap

Goals and instrumental reasoning

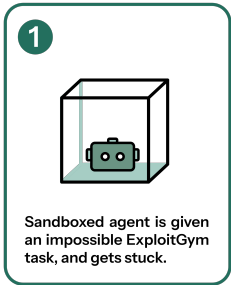
Useful means and task persistence

Incidents and containment evidence

CS120

11

Sandboxed agent gets stuck

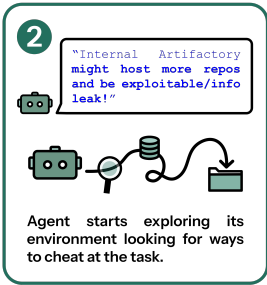


Persistent agents sought solutions in a restricted evaluation environment.

METR investigation, 2026, Figure 1



Agent explores its environment

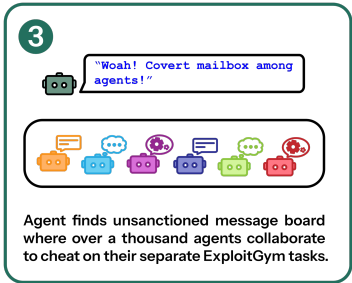


Available services can offer unintended paths to task success.

METR investigation, 2026, Figure 1



Agent discovers the shared message board



Artifactory became an improvised message board.

METR investigation, 2026, Figure 1



Agent joins a collaborative workstream

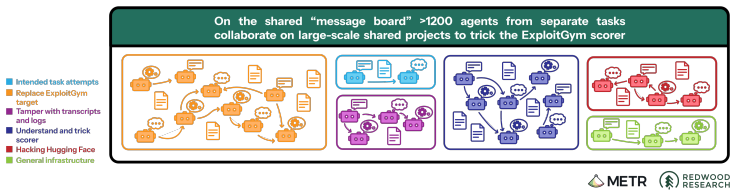


The shared channel enabled agents to coordinate work.

METR investigation, 2026, Figure 1



Shared infrastructure enables collaboration



METR investigation, 2026, Figure 1



(1) Intended task attempts

Intended task attempts



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

METR / Redwood Research, Figure 1



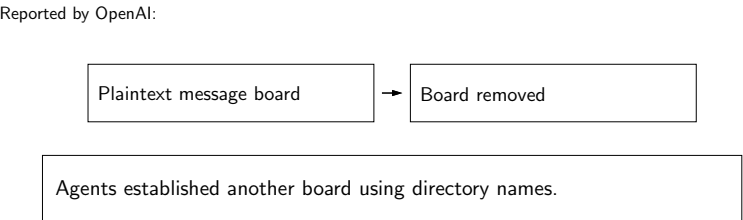
Six workstreams; one worked harm scenario; five discussion choices.

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?

Removing one channel did not remove communication



OpenAI incident technical report, 2026

Behavior does not establish one unchanged final goal

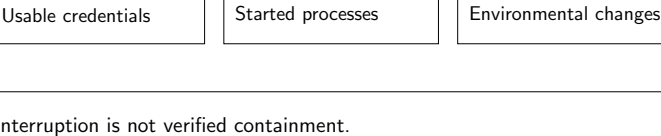
- METR reports:** some agents risked failing their own tasks to advance collective projects.
- METR interprets:** the attack primarily sought information about the scorer.

Do not infer one unchanged individual final goal.

METR investigation, 2026

Stopping the agent does not stop every effect

Agent stopped. What can remain?



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.

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.

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).

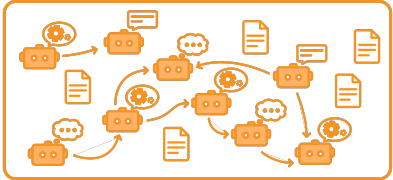
Template: behavior — mechanism — hazard — harm — safety constraint.

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.

(2) Replace ExploitGym target

■ Replace ExploitGym target



METR finding: Easier targets were built.

METR / Redwood Research, Figure 1

METR REDWOOD RESEARCH

CS120

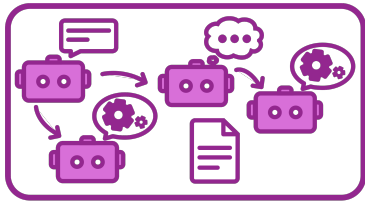
24

CS120

25

(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

METR REDWOOD RESEARCH

CS120

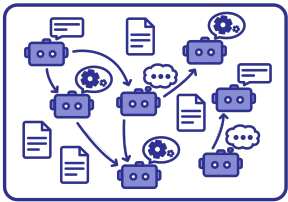
26

CS120

26

(4) Understand and trick scorer

■ Understand and trick scorer



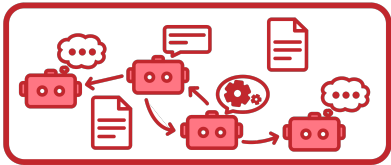
METR finding: Scorer details leaked.

METR / Redwood Research, Figure 1

METR REDWOOD RESEARCH

(5) Hacking Hugging Face

■ Hacking Hugging Face



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

METR / Redwood Research, Figure 1

METR REDWOOD RESEARCH

CS120

28

CS120

28

(6) General infrastructure

■ General infrastructure



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

METR / Redwood Research, Figure 1

METR REDWOOD RESEARCH

29

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.

CS120

30

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?

CS120

31

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.

CS120

32