

Who should set the pace of AI development

Read Dario Amodei's proposal and judge whether its checks would change how a frontier lab makes decisions. The aim is to examine the argument, evidence, and incentives before taking a position.

Assigned reading

Dario Amodei • We Must Pace the Frontier • September 2026

[Read the original essay on Dario Amodei's website](#)

Allow about 20 minutes for the article and 10 minutes for notes. This handout is a reading companion; the original essay is the assigned text.

The argument in short

Amodei argues that capability growth is outpacing safeguards. He proposes ongoing outside evaluation, coordination among frontier firms in democracies, and efforts toward international agreement. He wants additional time used for operational reliability, alignment, interpretability, and evaluation. Anthropic commits to inviting embedded reviewers; wider coordination remains a proposal. His case also depends on geopolitical assumptions about competition with China. Treat his predictions about future damage as forecasts, rather than established outcomes. This is a company leader's policy argument, not an independent finding that his preferred arrangements will work.

Terms to keep clear

- Capability means what a system can do. Safety evidence concerns the conditions under which its use may cause harm. A high task score does not settle both questions.
- Alignment concerns whether behavior follows intended goals and constraints. Interpretability studies internal mechanisms. They ask related but different questions.
- An evaluation samples behavior under specified conditions. Passing a test supports a bounded claim; it cannot establish correct behavior in every future situation.

Read with a pencil

Mark one reported observation, one prediction, and one proposed rule. For each, write down what evidence would support it and what could weaken it. Separate the author's confidence from your own.

Discussion about evidence and authority

- What should an outside reviewer be able to inspect, publish, and challenge? Who should fund the reviewer, and who could remove them?
- Choose one proposed restriction. What event would trigger it, what evidence would lift it, and who would decide?
- Could the same rule reduce a real danger and also protect large companies from competition? How would you test both possibilities?
- Which part of your judgment depends on a technical claim, and which part depends on a view about acceptable risk or political power?

A design case to discuss

A fictional museum wants an assistant to create exhibition captions, generate publicity images, and publish them automatically. In a small trial it produces attractive work, but occasionally invents an artwork's history. The museum can release it now, run it with staff review, or delay publication while testing continues.

Choose an option. State the specific harm you are trying to reduce and the evidence you need. Consider whether the assistant should have different permissions for drafting, editing a stored file, and publishing to the museum website. Write one rule that staff could actually follow when a test fails.

Bring to class

Prepare a 150–200 word response: state your decision, cite one passage from the original in your own words, give your strongest objection to your own position, and name one piece of evidence that could change your mind. You do not need to agree with the author.

Connection to Lecture 3

Slides 81–88 separate an assistant, a world model, and a system that takes actions. Apply the same distinction here. The ability to describe an action is different from permission to execute it. Explain exactly which component your proposed test examines.