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.

Should powerful AI models be openly available

Read Jack Dorsey's response to the frontier pacing debate. Compare his proposed limits with Amodei's and decide what evidence should be needed to restrict a model's release.

Assigned reading

Jack Dorsey • open the frontier • X, 15 September 2026

Read the original article on X

Allow about 15 minutes for the article and 10 minutes for notes. This handout accompanies the linked article. Read the full piece before forming a position.

The argument in short

Dorsey favors voluntary open releases, independent investigation, and defenses that target specific dangers. He opposes broad development limits negotiated by leading labs and rejects forced publication of private weights. He supports withholding only when independently reviewable evidence shows added catastrophic risk that narrower measures cannot adequately address. Restrictions should face review, appeal, and reconsideration. He accepts that a justified restriction may slow progress. His central challenge is who must justify limiting access, and whether the restriction also prevents useful research and defense.

Distinctions to keep clear

- Open weights allow access to learned parameters; reproducible research may also need code, data information, and evaluations. Ask exactly what is available and what a license permits.
- Running a model is different from giving it credentials, network access, or permission to act. Describe the system around the model when assessing a proposed use.
- A reported incident and a forecast of future catastrophe are different kinds of claims. Check the evidence needed to move from one to the other.

Read with a pencil

Find one safeguard Dorsey accepts and one restriction he rejects. State the reason for each in your own words. Then identify a condition under which his own argument would allow delaying a release.

Compare the proposals

Use both original essays. For each author, record the preferred starting rule, who should inspect the evidence, and what would justify limiting development or release. Do not reduce the disagreement to being for or against safety. Locate the actual point at which their proposed decisions could differ.

Discussion about access and harm

- When could wider access help someone find and fix a failure? When could the same access increase harm? Use a specific example rather than assuming either effect always wins.
- How could a small research group challenge a large lab’s claim? Describe the evidence, access, and resources it would need.
- What evidence would convince you that restricting a release reduces danger overall? Explain what comparison you would make and which effects might be hard to measure.
- If weights have already been copied widely, which controls remain practical? Distinguish controlling a hosted service from controlling every copy of a model.

A classroom design case

A fictional team develops an image model that makes convincing exhibition photographs. Independent artists want to adapt it on their own computers. A test also shows that it can create realistic but false photographs of public events. The team could release weights, offer a hosted service, or delay release while investigating further.

Choose an option and identify the harm it addresses. Explain which controls would still work after release, who should test them, and how your decision affects artists who need private or offline tools. This case is for comparing principles; it does not establish that an image model meets either author’s threshold for catastrophic risk.

Bring to class

Write 150–200 words identifying one agreement and one disagreement between Dorsey and Amodei. Apply the disagreement to the case, then name evidence that could change your decision. Support your interpretation with specific passages, paraphrased in your own words.

Connection to Lecture 3

The model, its learned weights, an image reference, and an application’s permissions are different parts of a workflow. Use those distinctions when discussing access. The technical reading supplies a small example of how a saved process can make results easier to inspect.

A controlled workflow for generating and editing images

This practical reading turns Lecture 3 into a repeatable experiment. You will plan a small image grid, compare one setting at a time, and repair a local error while preserving the useful parts of an image.

Assigned technical reading

Hugging Face Diffusers documentation • current pages checked 15 September 2026. Read the sections below; running their Python examples is optional. The exercise uses controls found in compatible diffusion tools and does not require a particular interface.

Allow 20 minutes for selected sections and 20–30 minutes for the exercise. If generation is unavailable, prepare the experiment plan and annotate a grid from class.

What each control changes

ControlNet supplies spatial information such as edges, depth, or pose through an additional network. Its conditioning scale controls the contribution of that input. The control model must be compatible with the base model and expected control type. An edge map constrains boundaries; it does not by itself specify material, identity, or the meaning of a scene. [1]

A seed helps control randomness. Diffusers passes a Generator object whose state changes when used; resetting it matters when comparing repeated calls. A seed alone does not guarantee identical results across hardware, software releases, or pipelines. Record the environment as well as the number. [2]

Image-to-image begins from an existing image. In the documented diffusion pipelines, greater strength adds more noise and allows a larger departure from that starting image. Guidance adjusts text conditioning and is a separate control. The useful range depends on the model and pipeline. [3]

Inpainting uses a mask to select the area to regenerate. In the documented convention, white marks the editable area and black marks the area to keep. Check the particular tool’s convention. Compare the resulting image with the original, including the boundary and supposedly preserved regions. [4]

Plan the image before choosing settings

Classroom exercise: make a poster image for a ceramics exhibition. Show one blue bowl on a table, leave the upper third clear for a title, and use side lighting that makes the rim readable. Add the final title as editable typography after image generation. These are exercise choices, not a tested recipe or a claim about a model's performance.

Draw a simple layout with the bowl's position and the empty title area. Write two requirements that must stay fixed and two appearance choices you may vary. This keeps composition separate from surface detail.

Build a six image comparison

Use a compatible ControlNet pipeline. Set the prompt, input sketch or edge map, image size, model version, sampler, steps, and guidance once. Choose three distinct seeds. For each seed, generate one image with lower spatial conditioning and one with higher spatial conditioning. Select supported values from the tool's documentation and record the actual numbers. Reset the random generator for each matched pair.

Place the results in three rows by two columns. Each row uses one seed; each column uses one control strength. Compare across a row to inspect the control change. Compare down a column to see how much the result varies across seeds. Do not switch the model or rewrite the prompt halfway through the grid.

Record the experiment

Record	What to save
Inputs	Exact prompt, negative prompt if used, sketch or control image
Model	Base model and revision, control model, any LoRA or adapter
Sampling	Seeds, sampler, steps, guidance and control strength
Environment	Tool and version, device if available, image dimensions
Outputs	All six images, run time, failures and later edits

Judge the images with visible criteria

For each image, record whether the title area is clear, the bowl count is correct, and the rim separates from the background. Then describe material and lighting quality. Choose a result using those observations. Keep a failure in the record: a selected image alone cannot show whether a setting works consistently.

Repair one local error

Suppose the best composition has a broken bowl rim. Save the selected image before editing. Mask the damaged area with enough surrounding context to form a plausible repair. Write an instruction about the rim. Start with the tool's documented settings, then compare a small number of edit strengths. Keep the original available for a direct before-and-after check.

Inspect the repaired contour, the light on the rim, and the texture at the mask boundary. Also inspect the table and empty title area. If unchanged pixels are essential, verify them or composite the selected edited region over the original in an image editor. A mask is an instruction to a pipeline; it is worth checking what the output actually preserves. [4]

Explain the mechanism

Answer these questions using your grid and repair. Why can different seeds produce different bowls? How does a control image differ from a text instruction? Why might a stronger edit repair the rim but also change its material? What information would another student need to repeat your comparison?

Connect the workflow to the safety debate

This is a low-stakes experiment, but its record makes a decision inspectable. A reviewer can see the inputs, alternative outputs, changed setting, and reason for selecting a result. A system that publishes images automatically would need an additional decision about permission and responsibility. Image quality tests alone do not answer that question.

For an exhibition poster, use references you are permitted to use and keep a record of their origin. If a generated image could be mistaken for a photograph of an actual exhibited object, decide how viewers will be told what it represents. Treat these as design and communication decisions, rather than assuming an attractive image settles them.

Bring to class

- One labelled six image grid, or a clear experiment plan if you cannot run the tool.
- One before-and-after repair with its mask, if editing is available.
- A settings record and a 150 word explanation of the tradeoff you observed. Report a failed attempt as well as your selected result.

Connection to Lecture 3

Review slides 48–64 for seeds, guidance and sampling; slides 65–72 for adaptation, spatial control and editing; and slides 90–98 for fair comparisons. The workflow above is a classroom synthesis of the four linked technical readings, not a benchmark result.