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.