

GRADES 9–12 · AGES 14–18 · 60 MINUTES · POLICY DEBATE

Should AI development be regulated?

THE PROPOSITION

"Governments should require licenses and safety testing before frontier AI models are released."

A policy debate on frontier-model regulation that takes both catastrophic and everyday risk seriously, and then asks whether a license regime would reduce either. Students encounter alignment, red-teaming, and regulatory capture, and must weigh a plan against the world without it.

STUDENTS WILL BE ABLE TO

- Distinguish present harms (bias, deepfakes, labor) from speculative catastrophic risk and explain why both appear in the debate.
- Build or attack a licensing plan with attention to who administers it, what it tests, and who it advantages.
- Weigh a policy against the counterfactual, not against perfection.

VOCABULARY FROM THE FIELD GUIDE

AI regulation · Existential risk (x-risk) · Alignment · Red teaming · Open weights / open models · Moat (competitive advantage)

WHAT'S IN THIS PACK

- Teacher guide: background, vocabulary, agenda, format, positions, questions, roles, rubric.
- Student handouts (photocopy one per student): position organizer, sentence starters, exit ticket.
- Standards: CCSS ELA SL.9–10.1 / SL.11–12.1; CCSS ELA W.9–10.1 / W.11–12.1 (argument with counterclaims); C3 Framework D2.Civ.13 / D4.7 (policy analysis, taking informed action); CSTA 3A-IC-24 / 3B-IC-25 (ethics and impacts of computing).

PRINT

US Letter, plain paper, duplex. Teacher pages once; student pages one per student. Free to photocopy for classroom and homeschool use. Not for resale.

Background for the teacher

Read this before class. Students get a version of the same facts on their handout; you get the framing and the traps.

Frontier models – the largest, most capable systems – are released by a handful of companies after internal safety testing. Some researchers, including leaders at those companies, have said the technology could pose catastrophic risks; others call that a distraction from present harms like bias, surveillance, and job loss. Both camps often support some regulation, for different reasons.

Proposed regimes include pre-release licensing (like drugs or aircraft), mandatory red-teaming and disclosure, compute thresholds that trigger oversight, and liability rules. The EU has passed a risk-tiered AI law; U.S. rules are a patchwork of executive orders and state laws. The alignment problem – making systems reliably do what we intend – is unsolved, which is an argument both for caution and for humility about what a test could certify.

The strongest objections: licensing favors incumbents who can afford compliance (regulatory capture, a moat), it can push development to other countries or underground, and open-weights models can't be recalled. The strongest replies: some technologies are licensed for good reason, capture is a design problem not a reason to do nothing, and 'someone else will build it' is not an argument that has ever excused a factory from safety rules.

RUNNING DISCUSSION IN GRADES 9–12

Stakes are real: college, jobs, creative work, civil rights. Students can steelman, cross-examine, and write policy.

- Insist on the steelman. Before rebutting, a speaker restates the other side's best version until they say 'yes, that's it.'
- Cross-examination is a question period, not a speech. Two minutes, questions only, no editorializing.
- Push past 'it depends.' Ask: depends on what, exactly – and who decides?
- Close with a written position. Speaking is the practice; the paragraph is the assessment.

Vocabulary

Pulled from the Hi, Bot AI concepts field guide (hibot.space/learn/concepts). Read each card aloud, then have the class rewrite it in ten words or fewer.

AI regulation

Government rules on development, deployment, and accountability.

AI regulation spans safety testing, transparency, sector rules (health, finance, education), export controls on chips, and liability when systems cause harm. Examples include the EU AI Act, U.S. executive orders, and sector guidance from schools or hospitals. Regulation lags technology; compliance is often "best effort against moving targets."

Existential risk (x-risk)

The argument that misaligned superhuman AI could pose civilization-scale danger.

Existential risk framing worries that sufficiently capable, misaligned systems could cause irreversible harm – through autonomous weapons, unstable feedback loops, or goals that diverge from human flourishing. Skeptics call it speculative; proponents want safety research prioritized now. Most daily AI use sits far from this debate, but it shapes funding and policy rhetoric.

WHERE PEOPLE ARGUE

One of the most polarized conversations in AI – timing, probability, and appropriate response are all contested.

Alignment**Making models behave according to human values and instructions.**

Alignment research asks how to ensure powerful systems pursue intended goals – helpful, harmless, honest – rather than optimizing the wrong proxy. In practice it includes training techniques, red teaming, policy rules, and monitoring. Perfect alignment is unsolved; "aligned enough for this use case" is an engineering judgment.

WHERE PEOPLE ARGUE

Doomers vs. accelerationists argue about whether misalignment is existential or manageable with guardrails. Product teams argue about whose values get encoded.

Red teaming**Deliberately trying to break or trick a model before users do.**

Red teams probe for jailbreaks, bias, data leaks, and unsafe advice – adversarial testing before launch. It's standard in serious AI products and increasingly expected by regulators. Red teaming finds known failure modes; it doesn't prove safety in all future situations.

Open weights / open models**Models whose trained parameters are published for anyone to download and run.**

Open-weight models release the actual trained network (often with a license governing use). That enables local running, inspection, and fine-tuning without calling a vendor API. "Open" varies – some licenses restrict commercial use or require sharing derivatives. Open weights are not the same as open training data.

WHERE PEOPLE ARGUE

Security vs. transparency: open models aid research and self-hosting; critics worry about misuse without gatekeeping.

Moat (competitive advantage)**Why one AI product keeps customers when models commoditize.**

In business debates, a moat is defensibility: proprietary data, distribution, brand, regulatory approval, workflow lock-in, or fine-tuned models on exclusive corpora. Base models trend toward commodity; applications fight on trust, integration, and outcomes.

Warm-up

6 minutes.

Ask: 'Name something you need a license to do or sell.' List ten fast (drive, practice medicine, sell food, fly a plane, build a bridge). Ask what those have in common. Then: 'Does a model belong on this list?'

Agenda - 60 minutes

Minutes	Segment	What happens
0-6	Warm-up	What do we license, and why? Does a model belong?
6-18	Background + risk map	Read the sheet. Teams sort risks into present / near / speculative, and note which a license would actually address.
18-28	Case construction	Affirmative plan: who licenses, what is tested, threshold, penalties, open-weights treatment. Negative: attacks + optional counter-plan (liability, disclosure, compute reporting).
28-52	Policy debate	Aff constructive 3 · Neg cross-ex 2 · Neg constructive 3 · Aff cross-ex 2 · Neg rebuttal 2 · Aff rebuttal 2. Judges weigh plan vs. counterfactual.
52-60	Written position	Exit ticket: policy, the risk it addresses, the risk it doesn't.

How policy debate runs

The proposition is a concrete policy ("Our school should...", "The law should..."). Affirmative proposes a plan and defends it; Negative attacks the plan and may offer a counter-plan. Constructive speeches (3 min), cross-examination (2 min), rebuttals (2 min). Judges score on plan, evidence, and clash.

GROUND RULES - POST THESE

1. Attack ideas, never people. "That claim has a hole" – not "you're wrong."
2. One voice at a time. The listener's job is to be able to repeat what was said.
3. Evidence beats volume. A quiet fact outranks a loud opinion.
4. Changing your mind is winning. Say so out loud when it happens.
5. Name the risk you're addressing: present, near, or speculative.
6. Weigh against the world without the plan, not against perfection.

The two positions

Starter arguments are scaffolds, not scripts. Strong teams find their own. Assign sides where the format calls for it; students record their real view on the exit ticket.

Affirmative – license before release

We license drugs, planes, and bridges because the downside is public and irreversible; frontier models qualify.

- The developers themselves say they can't fully predict what their models will do. That's the definition of a case for testing.
- Red-teaming by an independent body finds what internal teams are paid not to find.
- Present harms – deepfakes, bias at scale – are already here; a license regime creates a body that can act.
- 'Someone else will build it' has never excused a factory from safety rules.

Negative – the plan entrenches the powerful and doesn't test for what matters

No test certifies alignment; a license regime mostly certifies who can afford lawyers – and pushes the rest offshore or underground.

- The companies lobbying for licensing are the ones who'd pass. That's a moat, not safety.
- Open-weights models can't be recalled; the rule catches responsible labs and misses the rest.
- Catastrophic risk is speculative; the regime spends its energy there while present harms go unaddressed.
- Counter-plan: strict liability for harms, mandatory disclosure, and compute reporting – accountability without gatekeeping.

Discussion questions

OPENING

1. What makes something worth licensing? Apply your test to a model.

CORE

1. Can a pre-release test certify safety if alignment is unsolved? What would it certify instead?
2. If the biggest firms support the rule, is that evidence for it or against it?
3. How does your plan treat a model whose weights are already public?

CLOSING

1. Compare your plan to no plan – not to a perfect one. Is the world better?

Roles

Role	Job
Affirmative constructive	Plan: administrator, threshold, tests, penalties, open-weights treatment.
Negative constructive	Attacks on capture, enforceability, and what the test can't measure; optional counter-plan.
Cross-examiner (each side)	Questions only. Target the test and the threshold.
Rebuttal (each side)	Steelman, answer, weigh against the counterfactual.
Judge panel	Scores plan, evidence, clash. Names which risk the winning side handled better.

Assessment rubric

Share this before the debate. Judges and self-assessors use the same sheet. 'Meets' is the target; note above or below in the margin.

Criterion	Meets
Argument	Claim, warrant, and evidence are explicit and connected.

Criterion	Meets
Steelman	Presents the strongest version of the opposing view before attacking it.
Clash	Rebuttals respond to what was actually said, not a strawman.
Precision	Uses lesson vocabulary accurately and distinguishes is from ought.
Written position	Closing paragraph takes a defensible position and concedes one real cost.

Exit ticket prompts

1. The policy I'd support: ____.
2. The risk it addresses: _____. The risk it doesn't: _____.
3. The strongest argument against my position: _____.

Extension

Read a one-page summary of the EU AI Act's risk tiers or a U.S. state AI law. Which tier would a frontier chatbot fall in? What would it have to do? Would the affirmative or negative from today be happier with it?

SEND HOME

Ask at home: 'What's a rule at your work that exists because something went wrong once? Was it worth it?'

STANDARDS ALIGNMENT

- CCSS ELA SL.9-10.1 / SL.11-12.1
- CCSS ELA W.9-10.1 / W.11-12.1 (argument with counterclaims)
- C3 Framework D2.Civ.13 / D4.7 (policy analysis, taking informed action)
- CSTA 3A-IC-24 / 3B-IC-25 (ethics and impacts of computing)

More lessons by grade band, and the full AI concepts field guide, at hibot.space/learn/debates. Tell us how it went: hello@hibot.space.

Should AI development be regulated?

"Governments should require licenses and safety testing before frontier AI models are released."

WHAT YOU NEED TO KNOW

Frontier models – the largest, most capable systems – are released by a handful of companies after internal safety testing. Some researchers, including leaders at those companies, have said the technology could pose catastrophic risks; others call that a distraction from present harms like bias, surveillance, and job loss. Both camps often support some regulation, for different reasons.

Proposed regimes include pre-release licensing (like drugs or aircraft), mandatory red-teaming and disclosure, compute thresholds that trigger oversight, and liability rules. The EU has passed a risk-tiered AI law; U.S. rules are a patchwork of executive orders and state laws. The alignment problem – making systems reliably do what we intend – is unsolved, which is an argument both for caution and for humility about what a test could certify.

The strongest objections: licensing favors incumbents who can afford compliance (regulatory capture, a moat), it can push development to other countries or underground, and open-weights models can't be recalled. The strongest replies: some technologies are licensed for good reason, capture is a design problem not a reason to do nothing, and 'someone else will build it' is not an argument that has ever excused a factory from safety rules.

WORDS WE'LL USE

AI regulation – Government rules on development, deployment, and accountability.

Existential risk (x-risk) – The argument that misaligned superhuman AI could pose civilization-scale danger.

Alignment – Making models behave according to human values and instructions.

Red teaming – Deliberately trying to break or trick a model before users do.

Open weights / open models – Models whose trained parameters are published for anyone to download and run.

Moat (competitive advantage) – Why one AI product keeps customers when models commoditize.

MARK IT UP

Underline one sentence you agree with. Circle one you doubt. Write one question in the margin. Bring all three to the discussion.

Position organizer

Proposition: "Governments should require licenses and safety testing before frontier AI models are released."

Name: _____ Side (assigned or chosen): _____

OUR CLAIM, IN ONE SENTENCE

THREE REASONS – AND THE EVIDENCE FOR EACH

Reason 1:

Evidence (a fact, a definition, or a specific example):

Reason 2:

Evidence (a fact, a definition, or a specific example):

Reason 3:

Evidence (a fact, a definition, or a specific example):

WHAT THE OTHER SIDE WILL SAY – AND OUR REPLY

Their best argument:

Our reply:

VOCABULARY I WILL USE CORRECTLY

☐ AI regulation ☐ Existential risk (x-risk) ☐ Alignment ☐ Red teaming ☐ Open weights / open models ☐
Moat (competitive advantage)

Sentence starters

Use these until you don't need them. The goal is to respond to what was actually said.

- Our plan licenses ___ above ___, tested by ___, with penalties of ___.
- What does your test actually measure?
- Against the counterfactual of no rule, your plan ___.
- That's a present harm; the plan addresses ___ but not ___.
- Our counter-plan achieves ___ without ___.

GROUND RULES

1. Attack ideas, never people. "That claim has a hole" – not "you're wrong."
2. One voice at a time. The listener's job is to be able to repeat what was said.
3. Evidence beats volume. A quiet fact outranks a loud opinion.
4. Changing your mind is winning. Say so out loud when it happens.
5. Name the risk you're addressing: present, near, or speculative.
6. Weigh against the world without the plan, not against perfection.

HOW POLICY DEBATE WORKS

The proposition is a concrete policy ("Our school should...", "The law should..."). Affirmative proposes a plan and defends it; Negative attacks the plan and may offer a counter-plan. Constructive speeches (3 min), cross-examination (2 min), rebuttals (2 min). Judges score on plan, evidence, and clash.

YOUR ROLE TODAY

Role: _____

Role	Job
Affirmative constructive	Plan: administrator, threshold, tests, penalties, open-weights treatment.
Negative constructive	Attacks on capture, enforceability, and what the test can't measure; optional counter-plan.
Cross-examiner (each side)	Questions only. Target the test and the threshold.
Rebuttal (each side)	Steelman, answer, weigh against the counterfactual.
Judge panel	Scores plan, evidence, clash. Names which risk the winning side handled better.

LISTENING TALLY

Keep count while others speak. Claims: _____ Evidence: _____ Questions: _____ Vocabulary used: _____
Minds changed: _____

Exit ticket

"Governments should require licenses and safety testing before frontier AI models are released."

Name: _____ Date: _____

The policy I'd support: ____.

The risk it addresses: _____. The risk it doesn't: ____.

The strongest argument against my position: ____.

SELF-ASSESSMENT

Criterion	Meets looks like	Me today (above / meets / below)
Argument	Claim, warrant, and evidence are explicit and connected.	
Steelman	Presents the strongest version of the opposing view before attacking it.	
Clash	Rebuttals respond to what was actually said, not a strawman.	
Precision	Uses lesson vocabulary accurately and distinguishes is from ought.	
Written position	Closing paragraph takes a defensible position and concedes one real cost.	

TAKE IT HOME

Ask at home: 'What's a rule at your work that exists because something went wrong once? Was it worth it?'