

GRADES 6–8 · AGES 11–14 · 55 MINUTES · FISHBOWL

Should AI decide things about people?

THE PROPOSITION

"AI should be allowed to make decisions about people - who gets a loan, a job interview, or a warning from the police."

A fishbowl discussion on automated decision-making, with real cases students can reason about. Half the class talks while the other half tracks claims and evidence, then they swap. The concepts – bias, explainability, human in the loop – come from what went wrong in the cases.

STUDENTS WILL BE ABLE TO

- Explain how a system trained on past decisions can inherit past unfairness.
- Argue for or against automated decisions in a specific domain using a real case.
- Propose a safeguard (explanation, appeal, human review) and say what it costs.

VOCABULARY FROM THE FIELD GUIDE

Bias (in data and models) · Explainability · Human in the loop · Training data · Automation bias

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.7.1 / SL.8.1; CCSS ELA SL.7.3 / SL.8.3 (evaluate reasoning and evidence); CSTA 2-IC-20 / 2-IC-21 (bias and accessibility in computing); C3 Framework D2.Civ.10 (rights and responsibilities).

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.

Automated decision systems already screen job applications, score loan applications, and flag people for extra police attention. They are trained on past decisions, so if past decisions were biased, the system learns the bias as a pattern. A well-documented résumé screener learned to downgrade applications containing the word 'women's' because past hires had been mostly men.

Explainability is the problem of saying why a model produced a particular output. Many models can't. That matters when the decision is about a person: you cannot appeal a decision nobody can explain. 'Human in the loop' means a person reviews or can override the system – which helps, but only if the person actually looks rather than rubber-stamping (automation bias again).

The defensible pro-AI position is not 'machines are fair' but 'humans are also biased, and a model can be audited in ways a hiring manager's gut cannot.' Make sure the fishbowl hears that version.

RUNNING DISCUSSION IN GRADES 6–8

Evidence enters the room. Students can hold a position they don't believe, and that's the point.

- Assign sides. Arguing a position you didn't choose is the fastest way to learn what the other side actually thinks.
- Require one piece of evidence per claim – a fact from the background sheet, a personal example, or a definition.
- Separate 'what is true' from 'what should we do.' Most middle-school debates tangle the two.
- Watch for the pile-on. If one side is winning on volume, pause and give the other side a 60-second huddle.

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.

Bias (in data and models)

Systematic skew – who's represented in training data and who gets hurt in outputs.

Bias in AI usually means unequal error rates or stereotypes baked in from historical data: hiring tools favoring certain résumés, image models defaulting to narrow demographics, language models mirroring toxic patterns from the web. Fixing bias requires dataset audits, evaluation across groups, and human review – not a single fairness slider.

Explainability

Can we understand why a model produced a specific output?

Explainability (and interpretability) asks for reasons behind predictions – critical in lending, medicine, and discipline decisions. Deep neural networks are often black boxes; post-hoc explanations can be approximate or misleading. Regulated domains may require human-readable justification even when the model can't truly "explain" itself.

Human in the loop

A person reviews or approves before action goes live.

Human-in-the-loop designs require approval for high-stakes steps: sending email, grading, medical suggestions, financial trades. It trades speed for accountability. The opposite – fully autonomous loops – is where agent hype meets governance anxiety.

Training data

The examples a model learned from – often the whole internet, licensed corpora, or both.

Training data is the text, images, or logs used during training. Quality, consent, copyright, and representation in that data shape model behavior. When people ask "where did it learn that?", they're asking about training data and fine-tuning – usually opaque for closed models.

WHERE PEOPLE ARGUE

Copyright, consent, and "fair use for training" are active legal and political fights worldwide.

Automation bias

Trusting the machine because it's fast and confident – even when it's wrong.

Automation bias is the human tendency to defer to automated recommendations, especially under time pressure. It's why verification matters in copilot settings: the risk isn't only model error, it's humans stopping checks. Training and UI design should keep responsibility visible.

Warm-up

6 minutes.

Present the résumé case in three sentences. Ask: 'Who is responsible – the company, the engineers, the past hiring managers, the model?' Take a fast poll by hands. Note the split; return to it at the end.

Agenda - 55 minutes

Minutes	Segment	What happens
0-6	Warm-up	Résumé screener case. Responsibility poll.
6-14	Cases + vocabulary	Read the three cases (hiring, loans, policing) and the five vocabulary cards. Each student writes one claim and one question.
14-26	Fishbowl round 1	Inner circle of 6 + one empty chair discusses hiring and loans. Outer circle tallies: claims, evidence, questions, vocabulary used.
26-38	Fishbowl round 2 (swap)	New inner circle discusses policing and the safeguards question. Same tallies.
38-48	Safeguard design	Groups of four propose ONE safeguard for ONE domain and state its cost (time, money, fewer decisions).
48-55	Re-poll + exit ticket	Repeat the responsibility poll. Discuss what moved. Handout.

How fishbowl runs

An inner circle of 5-7 students discusses while the outer circle observes with a listening task (tally claims, evidence, questions). One inner-circle chair stays empty: any observer may tap in, and someone rotates

out. Swap circles halfway so everyone speaks and everyone listens.

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. Outer circle is silent except through the empty chair.
6. Every claim names a case.

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.

Perspective – let the system decide, with audits

Humans are biased and slow; a model can be tested for bias and fixed in ways a person can't.

- You can run a million test cases on a model. You can't on a hiring manager.
- Consistent rules applied the same way to everyone is a kind of fairness.
- Speed matters: a loan decision in seconds instead of weeks helps people too.

Perspective – a person must decide, the system may advise

A decision about a person needs a reason the person can hear and challenge.

- If nobody can explain the decision, nobody can appeal it.
- The model learned from a biased past; it will repeat it faster.
- 'Human in the loop' only works if the human actually looks.

Discussion questions

OPENING

1. Who was responsible for the résumé screener's bias?

CORE

1. Is a consistent unfair rule better or worse than an inconsistent fair-ish human?
2. Should you have the right to an explanation for any decision made about you? By a person, too?
3. Is a human reviewer a real safeguard if they approve 99% of what the model says?

CLOSING

1. Pick one domain. Where exactly should the human be in the loop, and what does that cost?

Roles

Role	Job
Inner circle	Discuss. Cite a case before offering an opinion. Leave the empty chair open.

Role	Job
Tally keeper (outer)	Counts claims, evidence, questions, vocabulary. Reports totals after each round.
Tapper (outer)	May take the empty chair to add one point or ask one question, then rotates out.
Safeguard designer	Writes the group's one safeguard and its cost in two sentences.

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
Claim	States a clear position that answers the question.
Evidence	Supports each claim with a fact, definition, or specific example.
Response	Restates the other side accurately before rebutting.
Vocabulary	Uses at least two lesson terms correctly.
Conduct	Follows ground rules; invites quieter voices in.

Exit ticket prompts

1. A system trained on past decisions can be unfair because ____.
2. One decision I'd let a model make alone: _____. One I wouldn't: _____.
3. My safeguard: _____. Its cost: _____.

Extension

Find one real example of an automated decision system in the news. Identify: what it decides, what it was trained on, who checks it, and whether people can appeal.

SEND HOME

Ask at home: 'Have you ever been denied something by a computer – a loan, an application, a refund – and not been told why?'

STANDARDS ALIGNMENT

- CCSS ELA SL.7.1 / SL.8.1
- CCSS ELA SL.7.3 / SL.8.3 (evaluate reasoning and evidence)
- CSTA 2-IC-20 / 2-IC-21 (bias and accessibility in computing)
- C3 Framework D2.Civ.10 (rights and responsibilities)

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 decide things about people?

"AI should be allowed to make decisions about people - who gets a loan, a job interview, or a warning from the police."

WHAT YOU NEED TO KNOW

Automated decision systems already screen job applications, score loan applications, and flag people for extra police attention. They are trained on past decisions, so if past decisions were biased, the system learns the bias as a pattern. A well-documented résumé screener learned to downgrade applications containing the word 'women's' because past hires had been mostly men.

Explainability is the problem of saying why a model produced a particular output. Many models can't. That matters when the decision is about a person: you cannot appeal a decision nobody can explain. 'Human in the loop' means a person reviews or can override the system - which helps, but only if the person actually looks rather than rubber-stamping (automation bias again).

The defensible pro-AI position is not 'machines are fair' but 'humans are also biased, and a model can be audited in ways a hiring manager's gut cannot.' Make sure the fishbowl hears that version.

WORDS WE'LL USE

Bias (in data and models) - Systematic skew - who's represented in training data and who gets hurt in outputs.

Explainability - Can we understand why a model produced a specific output?

Human in the loop - A person reviews or approves before action goes live.

Training data - The examples a model learned from - often the whole internet, licensed corpora, or both.

Automation bias - Trusting the machine because it's fast and confident - even when it's wrong.

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: "AI should be allowed to make decisions about people – who gets a loan, a job interview, or a warning from the police."

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

☐ Bias (in data and models) ☐ Explainability ☐ Human in the loop ☐ Training data ☐ Automation bias

Sentence starters

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

- In the ___ case, the system ___, which shows ___.
- I'd add to what ___ said: ___.
- That works for hiring, but for policing ___.
- I'm tapping in because I want to ask ___.
- The cost of that safeguard would be ___.

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. Outer circle is silent except through the empty chair.
6. Every claim names a case.

HOW FISHBOWL WORKS

An inner circle of 5–7 students discusses while the outer circle observes with a listening task (tally claims, evidence, questions). One inner-circle chair stays empty: any observer may tap in, and someone rotates out. Swap circles halfway so everyone speaks and everyone listens.

YOUR ROLE TODAY

Role: _____

Role	Job
Inner circle	Discuss. Cite a case before offering an opinion. Leave the empty chair open.
Tally keeper (outer)	Counts claims, evidence, questions, vocabulary. Reports totals after each round.
Tapper (outer)	May take the empty chair to add one point or ask one question, then rotates out.
Safeguard designer	Writes the group's one safeguard and its cost in two sentences.

LISTENING TALLY

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

Exit ticket

"AI should be allowed to make decisions about people – who gets a loan, a job interview, or a warning from the police."

Name: _____ Date: _____

A system trained on past decisions can be unfair because ____.

One decision I'd let a model make alone: _____. One I wouldn't: _____.

My safeguard: _____. Its cost: _____.

SELF-ASSESSMENT

Criterion	Meets looks like	Me today (above / meets / below)
Claim	States a clear position that answers the question.	
Evidence	Supports each claim with a fact, definition, or specific example.	
Response	Restates the other side accurately before rebutting.	
Vocabulary	Uses at least two lesson terms correctly.	
Conduct	Follows ground rules; invites quieter voices in.	

TAKE IT HOME

Ask at home: 'Have you ever been denied something by a computer – a loan, an application, a refund – and not been told why?'