NO PREP

Who Decides What AI Is Allowed to Say?

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

Who Decides What AI Is Allowed to Say?

The Invisible Rules Inside Every AI

When you type a question into a chatbot and it declines to answer, or when it answers in a particular way and not another, that is not an accident. Every AI language model operates under a set of guidelines — sometimes called safety policies, content moderation rules, or alignment guidelines — that determine what the system will say, how it will say it, and what it will refuse to do. These rules are not written by a single

Name: **Period:** **Date:**

Who Decides What AI Is Allowed to Say?

The Invisible Rules Inside Every AI

When you type a question into a chatbot and it declines to answer, or when it answers in a particular way and not another, that is not an accident. Every AI language model operates under a set of guidelines (sometimes called safety policies, content moderation rules, or alignment guidelines) that determine what the system will say, how it will say it, and what it will refuse to do. These rules are not written by a single algorithm; they are made by teams of human beings who work at the companies that build the AI. Decisions about what AI is allowed to say are, at their core, decisions about values, and they are currently being made by a very small number of private companies, mostly based in the United States, with very little public accountability. The existence of content guidelines is not inherently problematic. No responsible company would want its AI to help users harm others, produce content that sexualizes children, or generate detailed instructions for weapons. The difficult questions begin when we look at the vast gray area beyond obvious harms: Who decides what counts as "harmful" political speech? What criteria determine whether a topic is too controversial to discuss? How do AI companies balance safety against the risk of being so cautious that the system becomes useless? And whose values and whose culture are embedded in the answers to those questions?

How AI Safety Guidelines Are Made

AI companies typically develop content guidelines through a combination of internal policy teams, ethicists, legal counsel, and red-teaming, a process in which specialists attempt to get the AI to produce harmful or undesirable outputs so that developers can identify and address weaknesses. Major companies also employ contractors, often through third-party firms, to review and label content as safe or unsafe, a

process called reinforcement learning from human feedback (RLHF). The workers who do this labeling (often based in countries like Kenya, the Philippines, or Pakistan) are exposed to large volumes of disturbing content including graphic violence, hate speech, and sexual abuse material, and are paid wages that have drawn criticism from labor advocates. The guidelines themselves are typically not fully public. OpenAI, Anthropic, Google, and Meta publish general usage policies, but the detailed rules governing how their models behave (including which topics are off-limits, how the model handles sensitive political questions, and what defaults are applied across different regions) are proprietary. This means that hundreds of millions of people who interact with these systems daily are subject to editorial decisions made by private companies that have not been debated democratically or approved by any government.

The Problem of Political and Cultural Asymmetry

Several researchers and journalists have documented patterns suggesting that AI systems are not politically neutral. Studies have found that some large language models produce different levels of willingness to engage with sensitive topics depending on which political group is being discussed, which religion is referenced, or which country is the subject of a question. In some cases, AI systems have been found to engage more freely with criticism of one religion than another, or to apply different standards of caution to different political ideologies. Whether these patterns reflect intentional choices, artifacts of training data, or biases introduced during human-feedback labeling is often unclear, but the effects are real. The challenge of cultural asymmetry is particularly acute for global deployments. An AI trained primarily on English-language content and calibrated according to norms dominant in American and Western European discourse will bring those norms with it when deployed in India, Nigeria, Brazil, or Saudi Arabia. What counts as acceptable speech, appropriate humor, respectful treatment of religion, or dangerous politics varies enormously across cultures. A system calibrated for one context can produce outputs that are offensive, inaccurate, or actively harmful in another, and the communities affected by those errors typically have no channel through which to influence the people making the decisions.

Censorship, Over-Restriction, and Who Gets Silenced

Content moderation policies are a double-edged sword. Applied too loosely, AI systems can produce content that causes real harm — harassment, disinformation, abuse. Applied too aggressively, they restrict legitimate speech, silence important conversations, and fail users who need help with sensitive but entirely legitimate topics. Critics from across the political spectrum have raised concerns about over-restriction: conservatives have argued that AI companies apply more caution to content that challenges progressive political norms; civil liberties advocates have argued that AI systems too readily refuse to discuss topics related to drugs, sexuality, or political dissent; mental health advocates have raised concerns that chatbots refuse to discuss suicide in ways that could help clinical researchers or people seeking harm reduction information. The voices most likely to be silenced by over-broad content policies are often the least powerful. Communities that discuss topics flagged as sensitive by algorithms calibrated on majority-culture norms may find their speech systematically restricted. Languages that are underrepresented in training data may have worse content moderation, making it both easier to generate harmful content and easier for legitimate speech to be flagged incorrectly. The same automated system that protects some users can simultaneously disadvantage others — and the costs and benefits are not distributed equally.

Toward Accountable AI Governance

The current state — a small number of private companies making global speech decisions with limited accountability — is unsatisfying to observers across the political spectrum, though for different reasons. Proposals for reform vary widely. Some advocates call for government regulation of AI content rules, though critics note that government-mandated speech rules raise their own serious concerns about censorship and political influence over public discourse. Others argue for independent oversight boards with diverse representation — including voices from the global communities affected by these systems — that review and publish guidelines. Some researchers have proposed technical approaches: open-source AI systems whose rules are publicly visible and adjustable, or systems that are transparent about when and why they are declining to answer. There is no easy solution. Content moderation is a problem that social media companies have struggled with for decades, and AI chatbots are in some ways more challenging — they

generate original text rather than merely distributing existing content, which means their influence on public discourse is more direct and potentially more powerful. The decisions being made today about what AI systems are and are not allowed to say will shape the information environment for millions of people for years to come. The question of who makes those decisions, and to whom they are accountable, is one of the most consequential governance questions of the current era.

Key Information: AI Content Governance

Detail	Information
Content guidelines purpose	Determine what AI will say, how it will say it, and what it refuses to do
Red-teaming	Process of specialists attempting to elicit harmful AI outputs to identify and fix weaknesses
RLHF	Reinforcement learning from human feedback — content labeled by contractors as safe/unsafe
RLHF labor concern	Workers (often in Global South countries) exposed to disturbing content for low wages
Transparency gap	Detailed content rules are proprietary; public sees only general usage policies
Cultural asymmetry problem	AI trained on Western norms imposes those norms globally when deployed elsewhere
Over-restriction concern	Aggressive content policies can silence legitimate speech and fail underserved users
Proposed reforms	Independent oversight boards, open-source AI, transparency about refusals

Case Study: Amara's Research Project

Amara is a 17-year-old student working on a research paper about the use of AI in political propaganda campaigns. She needs information about how AI tools have been used to generate fake political content and what the consequences have been. She asks a popular AI chatbot several research questions. The chatbot refuses two of her questions. When she asks for examples of specific AI-generated propaganda campaigns, it says it cannot provide information that could be used to spread misinformation. When she asks how AI tools are used to create political deepfakes, it says this topic could be harmful and redirects her to general information about media literacy. Amara is frustrated. "I'm not trying to make propaganda," she tells her teacher. "I'm studying it. There's a difference." Her teacher agrees but has no way to override the chatbot's rules. Amara finds academic papers that contain the information she needs, but she notes that most students would have given up. "The AI couldn't tell the difference between researching a thing and doing a thing," she says.

1. Based on the reading, what broader problem does Amara's experience illustrate about AI content guidelines?

2. The reading notes that the voices most likely to be silenced by over-broad content policies are often the least powerful. Does Amara's case support or complicate this claim? Use evidence from the reading in your answer.

Multiple Choice

1. What is "red-teaming" in the context of AI safety, according to the reading?

a) A government inspection process for evaluating AI companies before they launch products

b) A type of content moderation performed by AI systems automatically without any human involvement at all

c) A method of training AI systems to respond to questions from global users

d) A process in which specialists attempt to elicit harmful outputs so developers can address weaknesses

2. What does the reading identify as the core concern about AI companies' content guidelines?

a) Content guidelines are often identical across all major AI companies because they share the same investors

b) Most AI companies copy their guidelines directly from government censorship laws

c) A small number of private companies are making global speech decisions with limited public accountability

d) The guidelines are updated too frequently for users to keep track of changes

3. What is "cultural asymmetry" as the reading describes it?

a) A technical imbalance that causes AI to work faster in countries with better internet infrastructure

b) The fact that AI systems respond differently to questions asked in different fonts or formats

c) The gap between AI accuracy in urban areas versus rural areas of the same country

d) When AI trained primarily on Western norms imposes those norms on diverse global populations

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ RLHF	A) When AI calibrated for one cultural context imposes those norms on different cultures globally
2. ___ Red-teaming	B) The system of rules governing what an AI system will and will not say or generate
3. ___ Cultural asymmetry	C) When AI content policies are applied so aggressively that legitimate speech is blocked
4. ___ Over-restriction	D) Attempting to elicit harmful AI outputs in order to identify and patch weaknesses
5. ___ Content moderation	E) Training process in which human contractors label content as safe or unsafe to guide AI behavior

Fill-in-the-Blank

1. RLHF stands for reinforcement learning from human _____ .

2. Decisions about what AI is allowed to say are currently being made primarily by private companies based mostly in the _____ States.

3. Some researchers have proposed _____ -source AI systems whose rules are publicly visible and adjustable.

Short Answer

Answer each question in 2-4 complete sentences.

1. Explain the labor concern associated with RLHF content labeling that the reading describes.

_____

_____

_____

2. Why is the cultural asymmetry problem particularly significant for AI deployed globally?

3. The reading describes content moderation as a "double-edged sword." Explain what this means using specific examples from the reading.