

AS-52 • AI Smart • One Question One Weekend

Should AI Have a Conscience?

This weekend, sit with one real AI-ethics question: should AI be able to refuse harmful things — almost like it has a conscience?

Start the brief

Write 150 words

YOUR CURRENT LEVEL

0 XP

Level 1: First Opinion

Understand the question

Explore both sides

Name one refusal

Name one worry

Write your view

Self-check

The weekend question

AI does not have feelings like a person. But AI tools can be designed to refuse some requests. For example, a tool might refuse to help someone harm another person, steal private information, or cheat in a serious way.

The hard question is: **what should AI refuse, and who gets to decide?**

1 Understand “conscience” in plain English

For this reflection, conscience means a built-in refusal rule.

1

Human conscience

A person’s inner sense of right and wrong. It includes feelings, values, responsibility, and judgement.

2

AI “conscience”

Not real feelings. More like rules that make AI refuse some harmful or unsafe requests.

3

Your question

Should AI refuse some things? If yes, what? If no, what could go wrong?

THOUGHT BULB

This is not about making AI human. It is about deciding what limits a powerful tool should have.

☐

I understand that AI “conscience” means refusal rules, not human feelings.

2

Use AI to explore both sides

Open ChatGPT, Claude, Gemini, or Perplexity. Paste this prompt.

COPY-PASTE PROMPT

Should AI be able to refuse to do harmful things — have a kind of conscience? What would you want it to refuse, and what would worry you about that?

Give me both sides in simple language for a Class 9–12 student. Do not write my final answer. End with 3 questions I should think about before writing my own view.

Copy prompt

CAUTION BOARD

Do not let AI write your final reflection. Use AI like a debate partner. Read both sides, then write your own view.

☐

I used AI to explore both sides, not to write the final answer.

3 Build both sides in your head

A strong reflection can understand the opposite side fairly.

Side A: AI should refuse harmful things

It can reduce harm when people ask for dangerous, cruel, illegal, or privacy-breaking help.

Side B: refusal rules can worry people

Who decides what counts as harmful?

A tool might refuse too much and block

It reminds users that powerful tools need boundaries.

It can protect younger users from some risky requests.

useful learning.

Rules might be unfair, hidden, or different across companies.

BALANCED-THINKING MOVE

Do not write “obviously yes” or “obviously no” too quickly. A better answer says what AI should refuse **and** what should still remain your decision.

4 If stuck, answer these two questions

This is the shortest path to your own opinion.

One thing I think AI should always refuse

Example: AI should refuse to help someone steal passwords, hurt another person, or invade someone's privacy.

One thing I think AI should NOT decide for me

Example: AI should not decide my personal values, my friendships, or what I believe without me thinking.

THOUGHT BULB

Your answer can be “yes, but...” or “no, unless...” Real-world questions often need careful middle positions.

☐ I named one thing AI should always refuse.

☐ I named one thing AI should not decide for me.

5 Write your 150-word reflection

Use AI to explore. Use your own mind to decide.

1

Opening

State your view in one sentence.

2

Reason

Explain why AI should or should not refuse harmful things.

3

Worry

Name one concern about giving AI refusal power.

4

Closer

Say what rule you would want AI to follow.

My 150-word reflection

Write your own view here. Aim for about 150 words.

Target: about 150 words. Your own view matters more than perfect length.

0 words

WRITING RULE

Your reflection should sound like you. It is okay if your view is still developing. A careful first opinion is better than a borrowed perfect answer.

☐ I wrote my own 150-word reflection.

6 **Self-check before you finish**

Make sure the reflection is yours, balanced, and clear.

☐ I gave my own view, not AI's final answer.

☐ I named something AI should refuse or limit.

☐ I named one worry about AI deciding what to refuse.

☐ I avoided personal/private details and unsafe examples.

PRIVACY REMINDER

Do not include your full name, address, phone number, school ID, passwords, or private stories. A reflection can be thoughtful without being personal.

☐

I completed the self-check.

7 Optional: make your weekend brief card

A small reminder of the real-world brief.

CARD COPY

Title: One Question One Weekend

Question: Should AI have a conscience?

Move: Use AI to explore both sides, then write your own 150-word view.

If stuck: Name one thing AI should always refuse, and one thing it should not decide for you.

Closer: Powerful tools need limits, but people still need judgement.

☐ I made, sketched, or saved the weekend brief card.

Final 5-minute move

Read your reflection once. Circle the sentence that is most clearly yours. That sentence is your real answer.

My clearest sentence

Copy the sentence from your reflection that sounds most like your own view.

NOTEBOOK LINE

Write this down: **“AI can explore both sides. I still decide what I believe.”**

Back to top

Vanguard Scholar — Smart questions. Smarter answers. You, upgraded.

Made with care. No data collected.

Questions? Teachers and parents, write to us. Students, speak with your teacher or a grown-up first.