DIGITAL SAFETY SQUAD

AI Safety Education: Discussion Prompts

Questions to explore trade-offs and challenge assumptions

How to Use These Prompts

These questions are designed to spark thoughtful classroom discussions about AI safety without requiring technical expertise. They work best when there's space for genuine uncertainty and multiple perspectives.

Facilitation tip: Resist the urge to provide 'the answer' too quickly. Let students explore different viewpoints, even ones you might disagree with. Safety thinking often emerges through debate, not lecture.

Understanding AI as a Safety Issue

We trust lots of powerful technologies in our daily lives (cars, aeroplanes, medical equipment). What makes us feel safe using them?

Follow-up: Do AI systems have those same safety features? What's missing?

When a bridge is built, engineers don't just ask 'Will it work?' They ask 'What happens if it fails?' Should we approach AI the same way?

Follow-up: What would 'failure' look like for an AI system you use regularly?

Think about your social media feeds. You didn't design what you see. An algorithm did. Who decided what that algorithm should prioritise?

Follow-up: If you could change one thing about how the algorithm works, what would it be?

Complexity and Control

A school isn't just the headteacher making decisions. It's a complex system with culture, policies, incentives, and hundreds of people all influencing each other. What makes a school system work well?

> **Follow-up:** Now think about an AI platform like TikTok or Instagram. What parts make up that system?

Imagine you post a question online and an AI gives you an answer that sounds confident but is completely wrong. Whose responsibility is that?

> **Follow-up:** What if millions of people saw that wrong answer before anyone caught it?

AI systems can adapt and learn from feedback. That's powerful, but it also means they can change in ways their creators didn't plan. Is that a feature or a risk?

> **Follow-up:** Can you think of an example where adaptation might lead somewhere harmful?

The Four Risk Categories

Malicious Use

AI tools can now create realistic fake images, videos, and voices. Is it the technology that's the problem, or how people choose to use it?

> **Follow-up:** Should companies building these tools have any responsibility for preventing misuse?

Before AI, creating convincing fake content required skill, time, and money. Now anyone can do it in seconds. Does that change how we think about the harm?

Follow-up: What safeguards would you want to see in place?

Race Dynamics

Two companies are competing to launch a new AI product. Company A rushes to market first but hasn't fully tested for safety issues. Company B takes longer to launch but catches several serious problems. Which company do you think will be more successful?

Follow-up: Does your answer change if you're thinking short-term (this year) vs long-term (five years from now)?

Imagine you're leading a team and your boss says 'Our competitor just launched a similar product. We need to ship ours next week or we'll lose.' What would you do?

Follow-up: What pressures might make it hard to say 'We're not ready yet'?

Organisational Risks

When something goes wrong with an AI system, who should be held accountable? The engineers who built it? The managers who deployed it? The company that sold it? The user who trusted it?

Follow-up: What happens when accountability is unclear?

A company's bonus structure rewards teams for launching new features quickly. Would you expect that to make their products more or less safe? Why?

Follow-up: How could you design incentives that reward both speed and safety?

Misaligned Systems

A food delivery app's AI is optimised to get orders delivered as fast as possible. It starts encouraging drivers to speed and take risky shortcuts. The system is doing exactly what it was told (maximise speed). So who's at fault?

Follow-up: How could the goal have been set differently to avoid this?

You can tell an AI system what to do, but you can't always predict how it will get there. Should we use systems we can't fully predict? Under what circumstances?

Follow-up: What level of monitoring or oversight would make you comfortable?

Trade-offs and Tensions

Progress vs safety: Are these opposites, or can they work together?

Follow-up: Give an example of technology that was developed quickly but safely. What made that possible?

Efficiency vs care: AI can grade hundreds of essays in seconds. But should speed matter more than understanding each student's unique thinking?

Follow-up: Where else might we be prioritising efficiency at the cost of something important?

Convenience vs privacy: An app that knows everything about you can give better recommendations. Are you willing to make that trade?

Follow-up: Where do you draw the line? What data feels too personal to share?

Innovation vs caution: If we're too careful with new technology, we might miss out on genuinely helpful tools. But if we're not careful enough, we risk serious harm. How do we balance these?

Follow-up: Who should get to decide where the line is?

Real-World Applications

Imagine you're hiring for a job and an AI system ranks all the candidates. It's faster and supposedly more objective than humans. Would you trust it completely? Partially? Not at all?

Follow-up: What would you need to know about how the system works to feel comfortable using it?

Your school wants to use AI to monitor student behaviour online and flag concerning posts early. The goal is to prevent harm. What are the potential benefits? What are the risks?

Follow-up: Would knowing you're being monitored change how you communicate online?

AI can now write convincing essays, solve complex problems, and create art. Some people say this will free humans to focus on more creative work. Others worry it will devalue human skills. What do you think?

Follow-up: Are there skills or abilities you think should remain distinctly human?

Personal Responsibility and Future Choices

You're working on a group project and someone suggests using AI to speed things up. When is that helpful collaboration, and when does it cross into cutting corners?

Follow-up: How would you make that judgement call?

If you notice an AI system behaving unexpectedly or producing concerning outputs, do you have a responsibility to say something? Even if you're not the person who built it?

Follow-up: What would make it easier or harder to speak up?

In ten years, you might be working somewhere that uses AI tools every day. What questions would you want to ask about those tools before trusting them?

Follow-up: What would make you feel confident saying 'I don't think this is ready yet'?

Some companies are racing to build more powerful AI systems as quickly as possible. If you were advising them, what would you tell them not to rush?

Follow-up: Are there some uses of AI that should require more caution than others?

Reflecting on the Bigger Picture

We've talked about risks, but AI also has real benefits. What problems do you think AI could genuinely help solve?

Follow-up: How do we make sure those benefits are developed safely?

Do you think the current generation of students has a different relationship with AI than previous generations had with technology?

Follow-up: Does that give you more responsibility, or does it mean you're better prepared?

If you could design an AI safety course for your year group, what would be the most important thing for everyone to understand?

Follow-up: What format would make it most useful? Discussions, activities, real case studies, or something else?

Related Resources

- AI Safety and Our Shared Future (main article)
- Teacher Notes for AI Safety Education
- Student Activities for AI Safety Education
- Optional Slides for AI Safety Education

More resources at digitalsafetysquad.com

Source: digitalsafetysquad.com/ai-safety-education