DIGITAL SAFETY SQUAD

AI Safety Education: Teacher Notes

Background context and facilitation guidance for educators

Overview

These notes support educators delivering AI safety education without requiring technical expertise. The focus is on helping students develop safety thinking, recognise patterns of risk, and understand that powerful systems require careful oversight.

Learning Goals

- **Critical thinking:** Students learn to question assumptions, spot incentive pressures, and recognise when safety might be overlooked.
- **Pattern recognition:** Students identify recurring risk categories (malicious use, race dynamics, organisational failures, misaligned systems).
- **Responsible decision-making:** Students understand that progress and safety should move together, not in opposition.
- **Future preparedness:** Students develop habits of mind that will serve them in AI-shaped workplaces and communities.

Key Concepts to Emphasise

1. AI as a Safety Issue

Help students understand that AI isn't just about efficiency or capability. It's about reliability, accountability, and what happens when mistakes scale. Safety thinking asks different questions than purely technical thinking.

> **Teaching tip:** Use familiar analogies. Just as we don't only ask if a car goes fast, but also if it's safe, we should ask not just if AI works, but if it's reliable and what happens when it fails.

2. Complexity and Control

Modern AI systems behave like complex systems, not simple tools. They're made up of many connected parts (the model, data, platform, people, environment) that all influence each other in ways we can't always predict.

> **Teaching tip:** Compare to a school ecosystem. It's not just the headteacher or the rules. It's how culture, policies, incentives, and behaviours all interact. Small changes can ripple in unexpected ways.

3. The Four Risk Categories

Students should be able to recognise and name these recurring patterns:

- **Malicious use:** Intentional harm (fraud, manipulation, misinformation)
- **Race dynamics:** Pressure to move fast cuts safety corners
- **Organisational risks:** Unclear responsibility, weak oversight, misaligned incentives
- **Rogue/misaligned systems:** Systems that optimise for the wrong thing or behave unexpectedly

Facilitation Guidance

Creating Safe Discussion Spaces

- Encourage uncertainty. Safety thinking often means saying 'I don't know yet' or 'we need to check this.'
- Frame mistakes as learning opportunities, not failures.
- Acknowledge that students will encounter real pressure to move quickly in future workplaces. This is about giving them language and confidence to push back when needed.
- Avoid making students feel overwhelmed. Frame AI safety as achievable through careful thinking, not through fear.

Handling Difficult Questions

"Isn't AI progress more important than being cautious?"

This is a false choice. Safety and progress aren't opposites. The most successful technologies are ones we can trust. Rushing systems into the world before understanding their risks often creates more problems than it solves.

"Why should students care about this now?"

Because the habits they develop now will shape how they think about these issues for the rest of their lives. Today's students will enter workplaces where AI decisions are normal. Understanding risk patterns early means they're more likely to notice problems, ask better questions, and help build safer systems.

Adapting for Different Ages

Ages 11-13 (Year 7-9)

- Focus on concrete examples students already know (social media feeds, chatbots, homework tools)
- Keep language simple and avoid jargon
- Use more structured activities with clear steps

Ages 14-16 (Year 10-11)

- Introduce more nuance around organisational pressures and incentives

- Encourage debate and exploration of trade-offs
- Connect to career pathways and workplace decision-making

Ages 16-18 (Year 12-13)

- Emphasise real-world complexity and ethical dilemmas
- Discuss accountability, governance, and regulation
- Prepare students for university and workplace contexts

Assessment Ideas

Rather than testing for 'right answers,' assess whether students can:

- Identify risk patterns in novel scenarios
- Explain trade-offs between speed, capability, and safety
- Articulate why organisational culture matters for AI safety
- Propose thoughtful questions to ask before trusting a system

> **Example assessment:** Present a case study of a company launching a new AI tool quickly. Ask students to identify what risks might be overlooked, what questions they'd ask, and what safeguards they'd want to see.

Additional Support

These materials are designed to work together:

- **Teacher Notes** (this document) provide background and facilitation guidance
- **Student Activities** offer hands-on scenario work
- **Discussion Prompts** help guide classroom conversations
- **Optional Slides** support visual lesson delivery

You don't need to use all materials at once. Pick what fits your lesson time, student needs, and curriculum goals.

Related Resources

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

Source: digitalsafetysquad.com/ai-safety-education