DIGITAL SAFETY SQUAD

AI Safety Education: 60-Minute Lesson Plan

A practical lesson structure for introducing AI safety thinking

Lesson Overview

Duration: 60 minutes

Age range: 11-18 (adaptable across secondary school)

Students are introduced to AI as a safety issue, not just a technology topic. They explore how AI systems operate inside complex environments, why small failures can scale, and how human decisions inside organisations shape outcomes. Through discussion and roleplay, students practise recognising risk patterns and asking better questions about powerful systems.

Learning Objectives

By the end of the session, students will be able to:

- Explain why AI systems can be treated as safety-critical
- Describe how small issues can scale in connected systems
- Recognise four major AI risk patterns
- Practise asking questions that reveal responsibility, incentives and safeguards
- Understand how future workplace decisions influence AI safety outcomes

Key Concepts Covered

- Complex systems
- Scale and amplification of harm
- Malicious use
- Race dynamics
- Organisational risk
- Misaligned or unintended system behaviour
- Responsibility and accountability in decision-making

Required Materials

- Projector or board
- Scenario cards (Student Activities resource)

- Roleplay brief sheets (provided in this lesson plan)
- Pens / paper

Optional:

- Slide pack (Optional Slides resource)
- Systems mapping worksheet (for extension activity)

Lesson Structure

1. Opening Discussion: Everyday AI 10 minutes

Teacher prompt: "Where have you seen or used AI this week?"

Collect examples on the board. Students might mention:

- Chatbots (ChatGPT, Character.ai, Snapchat My AI)
- Homework tools
- Recommendation feeds (TikTok, YouTube, Spotify)
- Voice assistants (Siri, Alexa)
- Photo filters and editing apps

Follow-up question: "When lots of people use the same system, what happens if it gets something wrong?"

> **Teaching tip:** Guide students towards the idea that small issues become bigger when systems scale. One wrong answer is annoying. Thousands of wrong answers can cause real harm.

Introduce the key idea: AI systems influence what we see, what we believe, and what decisions get made. When they operate at scale, safety matters.

2. Mini-Input: AI as a Safety Issue 10 minutes

Teacher explains:

- AI systems now influence information, decisions and behaviour
- They operate inside real organisations and platforms
- Safety questions are about what happens when things go wrong, not just whether tools work

Introduce the four risk patterns:

- **Malicious use:** Intentional harm (fraud, manipulation, misinformation)
- **Race dynamics:** Pressure to move fast cuts safety corners
- **Organisational risk:** Unclear responsibility, weak oversight
- **Misaligned systems:** Systems that optimise for the wrong thing

Teaching tip: No technical depth needed. Just pattern recognition. Students should be able to name these categories and spot them in real examples.

3. Roleplay Activity: Inside an Organisation 20 minutes

Students are grouped into teams representing a fictional company launching a new AI-powered app.

Setup: Divide class into four groups and assign roles:

- **Product team:** Want to launch quickly to beat competitors
- **Safety team:** Have concerns about testing and safeguards
- **Leadership:** Need to balance speed, safety, and reputation
- **Users:** Trust the company but don't know what's happening behind the scenes

Each group receives a brief explaining their incentives and concerns.

Task: As a company, decide whether to launch the system now or delay for more testing. Each group presents their position (2 minutes each), then the class votes.

Goal: Reveal how speed, competition and unclear responsibility affect safety decisions.

Facilitation note: There's no 'right' answer. The goal is to surface tensions between different priorities and show that safety decisions happen in meetings, not just in code.

Roleplay Briefs

Product Team Brief:

You've been working on this app for 18 months. A competitor just launched something similar. If you don't ship in the next two weeks, you'll lose market share and your team's bonuses are at risk. The app works well in testing. Yes, there are a few edge cases, but you can fix those after launch. Time to ship!

Safety Team Brief:

You've identified three areas that need more testing: the app hasn't been tested with users under 16, the content moderation isn't working properly yet, and there's no clear plan for who monitors the system after launch. You're worried these issues could cause real harm at scale. You need at least another month.

Leadership Brief:

You're under pressure from investors to show growth. The competitor's launch is worrying. But you also know that one major safety incident could destroy your reputation. You need to decide: launch now and risk problems, or delay and risk losing the market. What would responsible leadership do?

Users Brief:

You're excited about this new app. You trust the company because they've built good products before. You assume they've tested everything properly. You don't know about the internal debate. What would you want the company to prioritise? What questions would you want them to ask before launching?

4. Scenario Discussion: When Scale Changes Stakes 15 minutes

Students receive scenario cards (individually or in pairs). Use examples from the Student Activities resource or create your own.

Example scenarios:

Scenario 1: An AI study tool gives a confident but completely wrong answer about a historical event. A student uses it for their essay without checking.

Scenario 2: Someone creates a deepfake impersonation of a teacher and shares it in a group chat as a joke.

Scenario 3: A chatbot designed for customer service starts responding inappropriately to emotional or sensitive questions.

Scenario 4: A family safety product is released without proper safeguards and starts collecting more data than parents realised.

Prompt questions for each scenario:

- What happens if this occurs once?
- What changes if it happens to thousands of people?
- Who should notice this problem?
- Who should act to prevent it?

- Which of the four risk patterns does this fit?

> **Teaching tip:** Encourage students to think about scale. A single incident might be manageable. The same incident happening to thousands of people becomes a systemic issue.

5. Reflection: Asking Better Questions 5 minutes

Class shares:

- What surprised them about today's lesson?
- What questions would they ask before trusting an AI system?
- Where should responsibility sit when things go wrong?

Close with this message:

"Future decisions about AI won't only be made by engineers. They'll be made by people in meetings, teams and organisations. The questions you practise asking now shape how safe those systems become later."

Teacher Notes

- No technical AI knowledge is required to deliver this lesson
- Encourage curiosity over 'right answers'
- Keep tone exploratory, not alarmist
- Focus on patterns and decision-making, not fear
- Allow students to relate examples to familiar platforms they already use
- The roleplay works best when students genuinely engage with their team's position, even if they disagree with it

Optional Extensions

Systems Mapping Worksheet

Students create a visual map showing how one AI failure can ripple through different groups:

- Individual users
- Communities
- Platforms

- Regulators
- Public trust

Homework Assignment

Task: Find a real news story about an AI system causing harm and identify which risk pattern it fits (malicious use, race dynamics, organisational risk, or misaligned systems).

Format: Short written summary (200 words) explaining the incident and the pattern.

Assessment Ideas

Short Written Reflection

Prompt: "Imagine you're working at a company in 10 years and your team wants to launch a new AI tool. What question would you ask before approving it?"

Success criteria: Student identifies a specific safety concern (not just 'Is it safe?') and explains why it matters.

Group Poster: Spot the Risk Patterns

Students work in groups to create a poster showing:

- The four risk patterns
- Real-world examples of each
- Questions to ask to spot each pattern early

Class Discussion Participation

Assess whether students can:

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

Adaptation Notes

For Younger Students (Ages 11-13)

- Simplify the roleplay briefs with clearer language
- Use more concrete, familiar examples (school apps, social media)
- Reduce roleplay time to 15 minutes and extend opening discussion
- Focus on two risk patterns instead of four (malicious use and race dynamics work well)

For Older Students (Ages 16-18)

- Add complexity to roleplay with financial pressures and regulatory considerations
- Introduce real case studies from news coverage
- Extend scenario discussion to explore governance and accountability
- Connect to career pathways and workplace decision-making

For Shorter Lessons (30-40 minutes)

- Skip the roleplay activity
- Focus on opening discussion (10 min), mini-input (10 min), and scenario discussion (15 min)
- Use scenarios as the main active learning component

Related Resources

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