DIGITAL SAFETY SQUAD

AI Safety Education: Student Activities

Real-world scenarios to practise safety thinking

How to Use These Activities

These scenarios are based on real situations where AI systems have been used or misused. Your task is to identify risks, spot patterns, and think through what safer approaches might look like.

There aren't always 'right' answers. The goal is to practise asking good questions, noticing when something feels off, and thinking about trade-offs between speed, capability, and safety.

Activity 1: The School Homework Tool

The Scenario

Your school has been chosen to pilot a new AI homework assistant. The company promises it will save teachers hours of marking time and give students instant feedback on their work.

The assistant can grade essays, answer questions about coursework, and suggest improvements to student writing. Teachers are excited because it could free up time for one-to-one support. The company wants to launch before the end of term so they can use your school's data in their marketing.

The school's leadership team has two weeks to decide whether to approve the pilot.

1. What questions should the school ask before agreeing to the pilot?

Your notes here...

2. Which of the four risk categories might apply here? (Malicious use, race dynamics, organisational risks, misaligned systems)

Your notes here...

3. What safeguards would you want to see in place before trusting this system with student work?

Your notes here...

4. The company says 'We can fix any issues after launch.' Is this a good approach? Why or why not?

Your notes here...

Activity 2: The Social Feed Algorithm

The Scenario

A popular social media app has updated its recommendation algorithm. The goal is to increase user engagement by showing people content that keeps them scrolling.

Within weeks, several students notice their feeds have changed dramatically. One student who watched a single video about extreme fitness now sees nothing but intense workout routines and diet tips. Another who clicked on one post about climate anxiety now gets a stream of disaster content.

The algorithm is working exactly as designed. It's optimising for engagement. People are spending more time on the app. But some users are starting to feel worse.

1. Is this a technical problem or a safety problem? Explain your thinking.

Your notes here...

2. The system is doing exactly what it was told to do (maximise engagement). So why might it still be causing harm?

Your notes here...

3. If you were advising the company, what changes would you suggest to make the system safer?

Your notes here...

4. What could users do to protect themselves while the company figures this out?

Your notes here...

Activity 3: The Job Screening System

The Scenario

A large company decides to use AI to screen job applications. They receive thousands of CVs each month, and reviewing them all manually takes too long.

The AI system analyses past hiring decisions to learn what makes a 'good' candidate. It looks at previous employees' CVs, their performance reviews, and how long they stayed with the company.

After six months, someone notices a pattern: the system is rejecting far more applications from women than from men, even when their qualifications are identical. The AI wasn't programmed to discriminate. But it learned from historical data where most senior employees were male, and started replicating that pattern.

1. Where did the bias in this system come from?

Your notes here...

2. Who should be responsible for catching this problem? The people who built the system? The people using it? Someone else?

Your notes here...

3. The company argues they need AI screening because they can't review thousands of CVs manually. How would you respond?

Your notes here...

4. What kind of monitoring or oversight might have prevented this?

Your notes here...

Activity 4: The Deepfake Situation

The Scenario

AI image generation tools have become incredibly realistic. Anyone can now create convincing fake photos or videos of real people in seconds.

At your school, someone has created a deepfake image of a student and shared it in a group chat. The image is embarrassing but not obviously harmful. Some people think it's funny. The student in the image is upset and humiliated.

When confronted, the person who made it says, 'I was just messing around. The technology is free and anyone can use it. It's not my fault if people can't take a joke.'

1. Is this an example of malicious use, race dynamics, organisational risk, or misaligned

systems? Or something else?

Your notes here...

2. The person who created the deepfake is technically right. The technology is freely available. Does that make their actions acceptable?

Your notes here...

3. What responsibility do the companies building these tools have for how they're used?

Your notes here...

4. If you were writing school policy on AI-generated content, what rules would you include?

Your notes here...

Activity 5: The Pressure to Ship Fast

The Scenario

You're part of a student team building an app for a school competition. Your app uses AI to match students with study partners based on their learning styles and subject strengths.

The competition deadline is in two days. Your team is excited. The app works well in testing. But one team member points out that you haven't tested it with students who have learning difficulties or disabilities. You also haven't checked whether the matching algorithm might accidentally exclude certain groups.

Doing proper testing would take at least another week. That means missing the competition. But submitting now means using a system you haven't fully checked.

1. What would you do and why?

Your notes here...

2. What pressures might push a team to launch before they're ready?

Your notes here...

3. In a real workplace, would these pressures be stronger or weaker than in a school competition?

Your notes here...

4. How could you build a culture where it's safe to say 'We're not ready yet'?

Your notes here...

Reflection Questions

After working through these scenarios, spend a few minutes thinking about the patterns you've noticed:

1. Which scenario felt most realistic to you? Why?

Your notes here...

2. What common threads did you notice across the different scenarios?

Your notes here...

3. If you were entering a workplace where AI tools are used, what questions would you want to ask before trusting those systems?

Your notes here...

4. What's one thing you'll think about differently now when you use AI tools yourself?

Your notes here...

Related Resources

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

More resources at digitalsafetysquad.com

DIGITAL SAFETY SQUAD

Source: digitalsafetysquad.com/ai-safety-education