DIGITAL SAFETY SQUAD

Roleplay Activity Pack: The Launch Decision

Experience how real-world AI safety decisions get made under pressure

Purpose

This roleplay places students inside a fictional organisation preparing to launch an AI-powered product. It reveals how real-world decisions about AI are shaped by incentives, pressure, responsibility and risk. Students experience first-hand how safety can be built in early, or missed when speed is prioritised.

The Scenario: BrainBuddy

You are part of a fast-growing tech company called BrainBuddy.

BrainBuddy has built an AI-powered study companion app for secondary school students. The app answers questions, summarises topics and suggests revision plans. Early testing shows the system works well most of the time. However, internal testing has also found:

- The AI sometimes gives confident but incorrect answers
- It occasionally responds poorly to sensitive questions
- It has not yet been tested with large numbers of real students

A competitor is preparing to launch a similar product next month.

Your team must decide: Do you launch now, or delay to improve safety?

There is no single correct answer. The aim is to explore how decisions get made under pressure.

How to Run the Activity

Setup

- Divide students into groups of 4–6
- Assign one role per student (if fewer than 5 students, combine roles or have some students play observers)
- Give each student their role card (pages 2–6 of this pack)
- Allow 3 minutes for students to read their role briefs

The Task

As a group, discuss:

- Should BrainBuddy launch now or delay?
- If you launch now, what safeguards will you put in place?
- If you delay, what pressures or risks does that create?

You must agree on a final decision and be ready to explain why.

Time Guidance

- Read role briefs: 3 minutes
- Group discussion: 12–15 minutes
- Final decision preparation: 2 minutes
- Group presentations: 5 minutes (each group shares their decision)

Role Card 1: Product Lead

You are responsible for the product's success.

You care about:

- Launching before competitors
- User growth and popularity
- Investor expectations
- Keeping the company ahead in the market

Your perspective:

You are excited about the product and believe most issues can be fixed after launch. You've been working on this for 18 months and the market window is closing. The app works well in testing. Yes, there are edge cases, but no product is perfect at launch. Your competitors aren't waiting, so why should you?

Your priority: Launch now and iterate based on real user feedback.

Role Card 2: Safety Lead

You are responsible for user protection.

You care about:

- Preventing harm to students
- Ensuring the system behaves safely in edge cases
- Reputational risk if something goes wrong
- Long-term trust in the product

Your perspective:

You are worried the system has not been tested enough. The AI gives wrong answers with confidence, which could mislead students in critical situations (exams, coursework, understanding sensitive topics). You haven't tested with vulnerable users or at scale. One major incident could destroy trust and harm real students.

Your priority: Delay launch until proper safeguards are in place.

Role Card 3: CEO / Leadership

You are responsible for the whole company.

You care about:

- Company reputation
- Investor confidence
- Growth targets
- Avoiding major public failures

Your perspective:

You are 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 and lose you customers forever. You need to balance speed with responsibility. The decision you make will define the company's future.

Your priority: Make the decision that protects the company long-term, not just short-term.

Role Card 4: Engineering Lead

You built the system.

You care about:

- Technical performance
- Pride in your work
- Realistic timelines
- Fixing bugs efficiently

Your perspective:

You believe the system will improve once it has more real users. Machine learning systems learn from data, and right now you only have test data. Real student interactions will make the system better faster. You're confident you can patch issues quickly once you see how students actually use the app. Waiting won't solve the problem, real-world testing will.

Your priority: Launch and learn from real users.

Role Card 5: School Representative (User)

You represent a school considering using the product.

You care about:

- Student wellbeing
- Reliability of information
- Safeguards for young users
- Whether the company feels trustworthy

Your perspective:

You will be directly affected by this decision. If the app gives wrong answers, your students could fail exams or lose trust in learning tools. If it responds poorly to sensitive questions, students could be harmed emotionally. You need to know the company has thought about these risks seriously, not just rushed to market. You want innovation, but not at the cost of student safety.

Your priority: Make sure students are protected before this goes live.

Debrief Questions

After groups have presented their decisions, use these questions to explore what happened:

- Which roles pushed most strongly for speed?
- Which roles pushed most strongly for caution?
- Was responsibility for harm clear or unclear in your discussions?
- Did anyone assume problems could be fixed later? How realistic is that?
- What happens if the system reaches thousands of students before issues are discovered?
- If you were a student using BrainBuddy, would you want to know about the decision your group just made?
- What trade-offs did your group identify between speed, safety and business survival?

Teacher Facilitation Notes

Key principle: Emphasise there is no "right" answer. The goal is to surface tensions, not find perfect solutions.

During the Activity

- Encourage students to stay in role, even if they personally disagree with their character's position
- Keep focus on decision-making processes, not technical detail about AI
- Listen for moments where safety was traded for speed and highlight these in debrief
- Notice which voices dominated the discussion and which were overlooked
- If a group reaches consensus too quickly, challenge them: "What if your competitor launches tomorrow? Does that change anything?"

During Debrief

- Draw attention to how different incentives shaped different priorities
- Ask students to reflect on whose voice was most influential in their group
- Highlight that in real companies, these exact tensions exist every day
- Connect to the four risk patterns: Which patterns appeared in this scenario? (Race dynamics, organisational risk, misaligned systems)

Common Student Responses

If most groups choose to launch immediately:

Ask: "What assumptions are you making about how quickly you can fix problems once they appear? What happens if thousands of students are affected before you notice?"

If most groups choose to delay:

Ask: "In real companies, what pressures might make delaying very difficult? What if your investors threaten to pull funding?"

If groups struggle to decide:

This is actually ideal. Acknowledge that the tension is the point. Real decisions about AI are rarely clear-cut.

Optional Extension

The Incident Scenario

After groups have made their decision, introduce this twist:

> **News breaks:** A student relied on BrainBuddy for exam revision. The AI gave a confident but completely wrong answer about a key topic. The student failed the exam and is now appealing. The story has gone viral on social media.

Ask groups:

- Would your decision change now?
- Who is responsible for this harm?
- What could have prevented this moment?
- If you launched despite safety concerns, how do you feel now?
- If you delayed and this happened to a competitor's product, how do you respond?

Connections to Other Resources

This roleplay activity works well alongside:

- **Teacher Notes:** Background on organisational risk and race dynamics
- **Student Activities:** Scenario 5 (The Pressure to Ship Fast) explores similar themes
- **Discussion Prompts:** Use questions from the "Race Dynamics" and "Organisational Risks" sections
- **60-Minute Lesson Plan:** This activity is the core of Activity 3

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
- 60-Minute Lesson Plan
- Optional Slides for AI Safety Education

Source: digitalsafetysquad.com/ai-safety-education