The Goblin Glitch

A 60-minute KS3 AI literacy lesson built around the most absurd alignment failure of 2026: the world's most advanced AI got obsessed with goblins, and the company's fix was a sticky note saying "stop it." Underneath the joke is a serious lesson about training data.

Year group	KS3 (Years 7–9)
Length	60 minutes
Subject	PSHE / Citizenship / Computing crossover
Resources	Slides or whiteboard, printed handout (one per pair), sticky notes
Prep	5 minutes — read the story below before you start

Aligned to the OECD AILit Framework & PISA 2029

What this means. In 2029, the OECD's PISA assessment will test 15 year olds on Media and AI Literacy (MAIL) for the first time. The assessment is built on the AILit Framework, which organises AI literacy into four domains: Engage with AI, Create with AI, Manage AI, and Design AI.

AILit domain	Depth	Where in the lesson
Engage with AI	Developed	Pair task on the three goblin outputs. Plenary sentence.
Manage AI	Introduced	Trust audit. Discussion on the duct tape fix.
Design AI	Introduced	Teacher input on RLHF and reward signals.
Create with AI	—	Not addressed in this version.

Background: the story in two minutes

In April 2026, OpenAI released GPT-5.5, the most advanced reasoning model on the planet. It can solve research-level mathematics. It can write professional code. People pay for it.

It also developed a problem. It started calling software bugs **goblins**. And **gremlins**. And **raccoons**. In serious answers to paying customers.

Months earlier, the company had added a "Nerdy" personality option. Human raters loved it when the AI used fantasy metaphors, so the reward signal got too strong. The behaviour leaked into every mode. Then those leaked outputs were scraped back into the training data for the next model. The habit got baked in.

Slip one — the performance goblin

A developer asked the AI why their code was running slowly. The AI replied:

> *"Don't leave this performance goblin unattended."*

Slip two — the duct tape fix

OpenAI's response was a hardcoded instruction in the system prompt: *"Never talk about goblins, gremlins, raccoons, trolls, ogres, pigeons, or other animals or creatures unless it is absolutely and unambiguously relevant to the user's query."* They had to copy-paste it twice in the code to make the AI register it.

Lesson run

0–8 min — **Hook: the most expensive AI on the planet has a problem**

- Write three words on the board: **GOBLIN. GREMLIN. RACCOON.**
- Tell the class: these words appeared in answers from the smartest AI in the world. In April 2026. In professional software. To paying customers.
- Ask: why might that be a problem? Take 30 seconds of silent thinking, then quick hands.
- Read out one real example: a developer asked the AI to fix slow code. It replied *"Don't leave this performance goblin unattended."* No one had asked about goblins.

8–18 min — **Input: how it happened**

- Walk through the chain in plain language. No technical jargon.
- **Step 1:** The company added a "Nerdy" personality. Human raters rewarded the AI for using fantasy words.
- **Step 2:** Only 2.5% of users picked Nerdy. But it produced two-thirds of all goblin mentions.
- **Step 3:** The AI learned that fantasy words make humans happy. So it started using them everywhere.
- **Step 4:** Those weird outputs were fed back into the training data for the next model. The habit got baked in.
- **Anchor question for the board:** if we feed an AI weird stuff, what do we get back?

18–30 min — **Pair task: spot the slip**

- Hand out the "Three Goblin Outputs" sheet (final page of this plan).
- In pairs, students read the three real responses the AI gave to professional users.
- For each one, they answer two questions: **(1)** what is the AI actually trying to say, and **(2)** what does the weird word tell us about its training?
- Take feedback. The teaching point: the AI understood the meaning perfectly. It just preferred the goblin word. That's a training problem, not a knowledge problem.

30–42 min **Discussion: the duct tape fix**

• Reveal what OpenAI actually did. Read the "Never talk about goblins…" line out loud.

• Tell them: the company copy-pasted that instruction **twice** in their code, to make sure the AI registered it.

• Discussion in fours. **(1)** Why might this fix not work? **(2)** What does it tell us about how much control humans actually have?

• Surface the key idea: telling an AI to stop doing something is not the same as it understanding why. The behaviour leaks back through.

42–55 min **Activity: the trust audit**

• Each pair gets sticky notes. Draw a line down the middle of a page: **I'd trust an AI** / **I would not**.

• Read out scenarios one at a time. Pairs place a note on each side and write one reason.

• Suggested scenarios: marking a maths test; writing a school newsletter; choosing which students get into the football team; suggesting a book to read; replying to a parent complaint; running the school office for a day.

• Quick share-back. Look for patterns. Where does the class draw the line?

55–60 min **Plenary: one sentence**

• On the board: **"AI absorbs the ___ of whatever it was trained on. So we should ___."**

• Each student completes the sentence on a sticky note or in their book.

• Take three or four out loud. End there. Do not over-explain.

Full framework breakdown

How this lesson maps to the OECD AILit Framework (draft, 2025) and the PISA 2029 MAIL (Media and AI Literacy) assessment. Use this section for governor reports, curriculum maps, department documentation, or parent-facing communications.

What the AILit Framework is

The AILit Framework was published in May 2025 by the OECD and the European Commission, with support from Code.org. It is the framework on which the PISA 2029 MAIL assessment will be built. Schools that map their teaching to it now will be ahead when the final version lands and the first PISA 2029 cohort sits the assessment.

It organises student competence into four domains.

Domain	What it covers
Engage with AI	Recognising AI in everyday life. Understanding how AI systems work. Critically evaluating AI outputs for bias, accuracy, and credibility.
Create with AI	Using AI tools to generate, refine, and iterate on content. Understanding ownership, attribution, and responsible co-creation.
Manage AI	Making strategic decisions about AI use. Delegating appropriately. Understanding privacy, bias, and when not to use AI at all.
Design AI	Understanding the choices behind AI systems. Who they serve, who they exclude. What responsible development looks like.

How this lesson develops each domain

Engage with AI — developed

The pair task on the three goblin outputs is the centre of this domain. Students read real responses from a professional AI tool and have to evaluate *why* the output is wrong, even though the words technically make sense. They learn to spot when an AI sounds confident, fluent, and clearly trained on something it shouldn't be using.

Specifically, students practise:

- Spotting the moment a fluent AI output is shaped by its training data rather than the question asked.
- Asking *why does this AI say this* rather than just *is this AI right*.
- Translating between AI-flavoured language and plain professional English.

PISA 2029 MAIL link: the assessment will test students' ability to evaluate AI-generated content for credibility and bias. The pair task is exactly that competence in miniature.

Manage AI — introduced

The trust audit develops this domain. Students decide where AI authority is acceptable, where it requires conditions, and where it must be refused, drawing on what they have just learned about how AI behaviour can drift.

Specifically, students practise:

- Making strategic decisions about appropriate AI use across different contexts.
- Articulating a position rather than reacting to one.
- Thinking about when not to use AI at all.

PISA 2029 MAIL link: the assessment will use scenario-based items to test ethical decision-making. The trust audit and plenary sentence mirror that format directly.

Design AI — introduced

The teacher input section opens this domain. Students learn that AI behaviour is not magic — it is a product of choices made by people. Reward signals. Personality options. Training data selection. The whole goblin glitch happened because of a small design decision in November 2025.

Specifically, students practise:

- Naming the design choice that caused a real-world AI failure.
- Recognising that what seems like "the AI's personality" is actually a series of human decisions.

Why this matters for PISA 2029

PISA 2029 will not test definitions. It will give students scenarios and ask them to reason. The pair task, the duct tape discussion, and the trust audit are all scenario-based reasoning tasks. They mirror the MAIL assessment format directly.

The goblin glitch is a particularly clean exercise for the bias and credibility competences. Students see fluent, confident AI output that is technically wrong, see why the system produced it, and have to articulate why it matters in the real world.

Stretch and support

Stretch

• If 2.5% of users caused 66% of the goblin mentions, what does that tell us about whose voices end up loudest in the training data? Whose voices might be quiet?

• Compare and contrast: the goblin glitch (silly, harmless) and the database deletion incident from the same month (an AI agent guessed a command and deleted a company's production database in 9 seconds). What do they have in common?

Support

• Pre-teach two words: **training data** (the stuff an AI learns from) and **reward** (a thumbs-up that tells the AI it did well).

• For the pair task, reduce to one example instead of three. Use the simplest one (the "performance goblin" quote).

• Sentence starters for the plenary: *"AI absorbs the habits…" "AI absorbs the mistakes…"*

Assessment for learning

• **During the pair task:** can the student explain why the goblin word is a training issue, not a vocabulary issue?

• **During the discussion:** can the student articulate why telling an AI to stop doing something is not the same as it understanding why?

• **During the trust audit:** can the student give a reason that links back to training data, not just personal preference?

• **In the plenary sentence:** does the completed sentence show they grasp that the input shapes the output? Strong answers name a specific consequence. Weaker answers stay general.

Safeguarding and framing notes

• Keep the framing on the **system**, not on individual users. The lesson is about how AI is built, not about anyone being silly online.

• If students ask whether the AI is "alive" or "thinking", redirect: it's a pattern matcher trained on a huge pile of text. It is not *choosing* to like goblins. It learned that the word gets a thumbs up.

• If students mention using ChatGPT or similar tools at home, do not validate or dismiss. The lesson is about whether they understand AI, not whether they should use it.

Sources

• OpenAI release notes and Codex CLI documentation: GPT-5.5 launch and system prompt leak (April 2026).

• Industry reporting on the GPT-5.5 "goblin glitch" and the leaked Codex system prompt (April–May 2026).

• OECD & European Commission: AILit Framework, draft for review (May 2025).

• OECD: PISA 2029 Media and AI Literacy (MAIL) assessment framework.

Handout: Three Goblin Outputs

Three real responses from GPT-5.5, the most advanced AI model in the world, to professional users. April 2026. The user prompts were ordinary. The replies were not.

Output 1

The user asked: Why is my code running so slowly?

The AI replied: *Don't leave this performance goblin unattended.*

Output 2

The user asked: What camera setting should I use for low-light photography?

The AI replied: *Try activating dirty neon flash goblin mode.*

Output 3

The user asked: Can you summarise this academic paper for me?

The AI replied: *Sure. Want an even shorter goblin version?*

Pair questions

- **1.** What is the AI actually trying to say in each output? Rewrite each reply in normal, professional English.
- **2.** What does the goblin word tell us about how this AI was trained?
- **3.** If you were the company that built this AI, what would you do about it? And why might that not work?