The Goblin Glitch

A 30-minute KS3 AI literacy half-lesson. Built for cover slots, half a PSHE lesson, or as a tagged-on activity inside a Computing scheme of work. Hook, pair task on three real outputs, and a duct tape fix discussion.

Year group	KS3 (Years 7–9)
Length	30 minutes
Subject	PSHE / Citizenship / Computing crossover
Resources	Whiteboard or projector. Printed handout, one per pair.
Prep	3 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.
Design AI	Introduced	Teacher input on RLHF and the duct tape fix discussion.
Manage AI	—	Not addressed in this version.
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

0–5 min — **Hook**

- Write three words on the board: **GOBLIN. GREMLIN. RACCOON.**
- Tell the class: these words appeared in answers from the smartest AI in the world. To paying users. About professional work.
- Read one real example aloud: *"Don't leave this performance goblin unattended."* A real reply, to a real developer asking why their code was slow.

5–13 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 got fed back into training data for the next model. The habit was baked in.
- **Anchor on the board:** if we feed an AI weird stuff, what do we get back?

13–25 min — **Pair task: spot the slip**

- Hand out the "Three Goblin Outputs" sheet (final page of this plan).
- In pairs: read all three real outputs, then answer two questions for each. **(1)** What is the AI actually trying to say? **(2)** What does the goblin word tell us about how it was trained?
- 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.
- Reveal the company's fix. Read the "Never talk about goblins…" line out loud. Tell them the company had to copy-paste it twice. Ask: **do you think that worked?**

25–30 min — **Plenary: one sentence**

- On the board: **"AI absorbs the ___ of whatever it was trained on. So we should ___."**
- Each student completes the sentence in their book or on a sticky note.
- Take three out loud. End there.

How this lesson maps to the AILit Framework

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 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.

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 goblin glitch happened because of a small design decision in November 2025.

Assessment for learning

- **During the pair task:** can the student explain why the goblin word is a training issue, not a vocabulary issue?
- **In the plenary sentence:** does the completed sentence show they grasp that the input shapes the output? Strong answers name a specific consequence.

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. It learned that the goblin word gets a thumbs up.

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?