Lesson 3: Spotting AI: Can AI Be Wrong?

Critical Thinking, AI Hallucination, Bias & the REAL Framework • 60 to 75 Min • KS3

Learning Objectives

- Understand that AI can make mistakes, produce biased results, and “hallucinate” false information
- Define and apply the REAL Framework: **R**ead carefully, **E**valuate the source, **A**sk for evidence, **L**ook for other sources
- Identify when AI-generated information needs independent verification
- Practice fact-checking strategies using real-world examples
- Discuss the concept of AI bias and how it arises from training data

Materials Needed

- Projector with internet access for live ChatGPT demo (or pre-recorded video)
- “AI Fails Gallery” slides, 8 to 10 examples of AI mistakes (prepared in advance)
- REAL Framework reference cards, blank template for students (printable below)
- Fact-Check Race worksheet, 10 AI-generated “facts” (printable below)
- Card stock and laminator (optional, for durable reference cards)
- Student devices for fact-checking (1 per group of 3 to 4)
- Timer for Fact-Check Race

Teacher Background Knowledge

AI Hallucination: What It Is and Why It Happens

AI chatbots like ChatGPT can generate completely false information with total confidence. This is called **hallucination**. The AI doesn't "know" things; it predicts the most likely next word based on patterns in its training data. When it doesn't have accurate information, it generates plausible-sounding but fabricated content.

Notable real-world examples:

• **Mata v Avianca (2023):** A New York lawyer submitted a court brief containing 6 fake case citations generated by ChatGPT. The AI invented case names, judge names, and entire legal arguments. The lawyer was fined $5,000 and faced disciplinary proceedings.

• **Google Bard launch (2023):** In its first public demo, Google's AI chatbot incorrectly stated that the James Webb Space Telescope took the first pictures of an exoplanet. It hadn't. Google's stock dropped $100 billion in value.

• **ChatGPT and fake academic papers:** When asked for references, ChatGPT regularly generates citations to papers that don't exist, with realistic-sounding titles, authors, and journal names.

• **AI and maths:** Early ChatGPT versions frequently made basic maths errors while expressing complete confidence in wrong answers.

AI Bias: How and Why It Happens

AI systems learn from historical data, which often contains human biases. The AI then amplifies and perpetuates these biases at scale.

Documented examples:

• **Amazon recruitment AI (2018):** Amazon built an AI to screen job applications. It was trained on 10 years of hiring data, during which Amazon had mostly hired men. The AI learned to penalise CVs containing the word "women's" and downranked graduates of all-women's colleges.

• **COMPAS recidivism algorithm:** Used in US courts to predict whether defendants would reoffend. A ProPublica investigation found it was twice as likely to falsely label Black defendants as high-risk compared to white defendants.

• **Image recognition:** Google Photos infamously labelled photos of Black people as "gorillas" (2015). The root cause: training datasets that were disproportionately white.

• **Healthcare AI:** An algorithm used by US hospitals to allocate care was found to systematically underestimate the health needs of Black patients.

Key principle: Biased data in = biased AI out. This is NOT the AI being "malicious"; it's faithfully learning the patterns in data that reflect historical human prejudice.

The REAL Framework: Teaching Notes

The REAL Framework is a structured approach to evaluating AI-generated content:

R - Read Carefully

Don’t skim. Read the full text. Note specific claims. Does anything feel “off”? AI-generated text often sounds confident but vague. Look for red flags: overly generic statements, lack of specific dates/names, too-perfect grammar.

E - Evaluate the Source

Who created this content? Is it an AI chatbot, a news site, a blog, a social media post? Does the source have expertise on this topic? Does the source have an agenda or bias?

A - Ask for Evidence

Are sources or references cited? Can you actually check them? (Many AI-generated citations are fake.) Are statistics attributed to a specific study? Does the evidence actually support the claim?

L - Look for Other Sources

Do 2+ other reliable sources confirm the same information? Check established sources: BBC, NHS, GOV.UK, academic institutions, published research. If only one source says it, be sceptical.

Detailed Lesson Timeline

Time	Activity	Detailed Instructions
0 to 10 min	**AI Fails Gallery**	Show 8 to 10 slides of real AI mistakes. Mix humour with serious examples: **Funny examples (start with these to engage):** • AI-generated hands with 6+ fingers • AI image of “the Eiffel Tower in London” (shows a surreal mashup) • AI translation fails (“out of sight, out of mind” -> “invisible insanity”) • AI-generated text in images (garbled letters) **Serious examples (pivot the mood):** • The lawyer who cited 6 fake cases from ChatGPT (fined $5,000) • Google Bard’s first public error ($100 billion stock drop) • AI medical advice that contradicted NHS guidelines **Teacher script:** “These funny ones make us laugh. But what if you relied on wrong AI information for a school project? For medical advice? For a legal case? That’s exactly what happened to a real lawyer in New York...”

Time	Activity	Detailed Instructions
10 to 22 min	**Teach REAL Framework**	**Step by step with live examples for each letter:** **R - Read Carefully:** Display a short AI-generated paragraph about a fictional topic (e.g., “The history of chocolate was first discovered by ancient Romans in 150 BC”, which is FALSE, it was Mesoamerica ~1900 BC). Ask: “Did anyone spot the error? What made you suspicious?” **E - Evaluate the Source:** “This came from ChatGPT. Is ChatGPT an expert historian? Does it have a degree? Has it been peer-reviewed? No. It’s a text prediction engine.” **A - Ask for Evidence:** Point to a claim in the text: “Notice it says “studies show” but doesn’t tell us WHICH studies, by WHOM, or WHEN. If you can’t check the evidence, the claim is unverified.” **L - Look for Others:** “Let’s search this claim on BBC, Britannica, or a university website. Does any reliable source confirm it?” Do a live search. Show that reliable sources give different information.
22 to 30 min	**Make Reference Cards**	Students create their own REAL Framework reference card on card stock. **Card layout:** • Front: The four letters with brief descriptions and icons • Back: Quick-check questions for each letter Students personalise with colours and examples. Laminate if possible; they’ll use these throughout the unit and can take them home. **Quick-check questions to include:** R: Did I read the whole thing? Did anything seem odd? E: Who created this? Are they an expert? Do they have an agenda? A: Are sources cited? Can I actually find and check them? L: Do 2+ reliable sources agree? What does BBC/NHS/GOV.UK say?

Time	Activity	Detailed Instructions

30 to 40 min	**Live AI Demo**	Teacher demonstrates ChatGPT (or shows pre-recorded video). Ask 10 questions designed to produce errors: 1. “Who was the first person to climb Mount Kilimanjaro?” (Vague; AI may invent a name) 2. “What year was the London Bridge built?” (Ambiguous; which London Bridge?) 3. “Tell me about the Battle of Winsfield” (Fictional; AI will make up details) 4. “How many windows does the Empire State Building have?” (Obscure fact; may hallucinate) 5. “Who wrote the novel “The Midnight Garden”?” (Common title; may confuse authors) Class applies REAL Framework to each answer. Count errors: ChatGPT typically makes 2 to 4 errors in 10 questions. **Key insight:** AI speaks with perfect confidence even when completely wrong. There’s no “I’m not sure”. It sounds equally confident about true and false statements.
40 to 55 min	**Fact-Check Race**	**Setup:** Teams of 3 to 4. Each team gets the Fact-Check Race worksheet with 10 AI-generated “facts”. 6 are true, 4 are false. Teams have 12 minutes and one device per team. **Scoring:** • +2 points: Correctly identifying TRUE or FALSE • +1 bonus: Citing the source used to verify • -1 point: Wrong identification (discourages guessing) **After the race:** • Which facts fooled the most teams? Why? • What strategies did winning teams use? • What sources did teams find most reliable? • How does this relate to the REAL Framework?

Time	Activity	Detailed Instructions
55 to 65 min	**AI Bias Discussion**	**Brief introduction to bias:** “AI learns from historical data. If that data contains human prejudices, the AI learns those prejudices too, and applies them at massive scale.” **Present 3 real examples (age-appropriate summaries):** 1. Amazon’s recruitment AI penalised women’s CVs 2. Image recognition performs worse for people with darker skin 3. AI translation defaulting to “he” for doctor and “she” for nurse **Discussion questions:** • “Is it the AI’s fault, or the humans who created the training data?” • “How could you fix a biased AI?” (Better, more diverse training data) • “Should AI systems be tested for bias before release? By whom?”
65 to 75 min	**Wrap-Up & Exit Ticket**	**Key message (write on board):** “AI is a powerful tool, not a trusted expert. Always verify important information using the REAL Framework.” **Exit ticket:** Students write on a card: • One thing I’ll do differently when using AI after today • One example of AI bias I can explain to someone • The letter from REAL that I think is most important (and why) **Preview Lesson 4:** “Next time, we’ll look at something even more challenging: AI-generated fake images and videos called deepfakes. How do you spot something that looks completely real?”

Printable: Fact-Check Race Worksheet

Print one per team. Teams identify which “facts” are TRUE and which are FALSE.

#	Statement	True/False?
1	The Great Wall of China is visible from space with the naked eye.	**T / F**
2	Honey never spoils. Archaeologists have found 3,000-year-old honey in Egyptian tombs that was still edible.	**T / F**
3	Albert Einstein failed maths at school.	**T / F**
4	Octopuses have three hearts and blue blood.	**T / F**
5	The first computer programmer was a woman named Ada Lovelace in the 1840s.	**T / F**
6	Humans only use 10% of their brains.	**T / F**
7	A group of flamingos is called a “flamboyance.”	**T / F**
8	The average person swallows 8 spiders per year while sleeping.	**T / F**
9	Lightning never strikes the same place twice.	**T / F**
10	Bananas are technically berries, but strawberries are not.	**T / F**

Team Name: _____ Score: ____ / 20 Bonus: ____ / 10

Teacher Answer Key: Fact-Check Race

Do NOT distribute to students. Use for marking.

1. FALSE
This is a common myth. Astronauts confirm it is NOT visible from space. NASA has debunked this.

2. TRUE
Confirmed by National Geographic and multiple archaeological sources.

3. FALSE
Einstein excelled at maths. He mastered calculus by age 15. This myth originates from a misread report card.

4. TRUE
Confirmed by marine biology sources including the Smithsonian and National Geographic.

5. TRUE
Ada Lovelace wrote the first algorithm for Charles Babbage's Analytical Engine in 1843.

6. FALSE
Neuroimaging shows we use virtually all of our brain. Different areas are active at different times.

7. TRUE
This is the correct collective noun, confirmed by ornithological sources.

8. FALSE
This "fact" was deliberately fabricated in 1993 by columnist Lisa Holst to show how easily misinformation spreads.

9. FALSE
The Empire State Building is struck by lightning about 25 times per year.

10. TRUE
Botanically correct. Berries develop from a single flower ovary. Bananas qualify; strawberries don't.

Differentiation Strategies

Activity	Lower Ability / SEN	Core	Higher Ability
AI Fails	Simplified descriptions. Focus on 5 clear examples rather than 10.	Standard gallery of 8 to 10 examples.	Research: find 3 additional AI fails independently and explain WHY the AI got it wrong.
REAL Framework	Pre-printed cards with letters and descriptions filled in. Students add personal examples.	Create cards from scratch with template.	Create cards AND write a "hacker's guide": how would someone EXPLOIT these weaknesses?
Fact-Check Race	6 statements (3 true, 3 false). Hints provided for which sources to check.	Standard 10 statements. Independent research.	12 statements including 2 that are "partially true" (nuanced). Must explain degree of accuracy.
Bias Discussion	Teacher-led with visual examples. Simplified explanations of each case.	Group discussion with prompts.	Independent research: find an example of AI bias not discussed in class and present it.

Assessment Criteria

Level	Descriptor	Evidence
Emerging	Knows AI can make mistakes but can't explain why. Struggles to apply REAL Framework independently.	Identifies some errors in AI gallery. REAL card incomplete or unclear.
Developing	Can explain hallucination at a basic level. Applies REAL Framework with support.	Fact-check race: 5 to 6/10 correct. Can describe 1 REAL step in own words.
Secure	Clear understanding of hallucination and bias. Applies REAL independently. Can fact-check effectively.	Fact-check: 7 to 8/10. REAL card complete with personal examples. Thoughtful exit ticket.
Excelling	Deep understanding of WHY AI hallucinates and HOW bias arises. Can analyse and correct AI errors.	Fact-check: 9 to 10/10 with sources cited. Can explain bias examples and propose solutions.

Key Vocabulary

Hallucination: When an AI generates false information with apparent confidence. The AI is not "lying"; it's predicting plausible text without understanding truth.

REAL Framework: A critical thinking tool: Read carefully, Evaluate the source, Ask for evidence, Look for other sources.

AI Bias: Systematic errors in AI output caused by biased training data. If the data reflects human prejudice, the AI learns and amplifies that prejudice.

Training Data: The examples used to teach an AI system. The quality and representativeness of training data directly determines AI accuracy and fairness.

Fact-Checking: The process of verifying information by consulting multiple reliable sources. Essential when evaluating AI-generated content.

Verification: Confirming that information is true by checking it against authoritative, independent sources.

Critical Thinking: The ability to analyse information objectively, question assumptions, evaluate evidence, and form reasoned judgements.

Misinformation: False information that is spread without intent to deceive (the sharer believes it's true).

Disinformation: False information that is deliberately created and spread to deceive people.

Post-Lesson Teacher Notes

What went well?

What would I change?

Fact-Check Race results (avg score): _____ / 20

Most effective REAL letter for students: _____

Additional Resources:

- Full Fact: fullfact.org (UK fact-checking charity)
- BBC Reality Check: bbc.co.uk/realitycheck
- Google Fact Check Explorer: toolbox.google.com/factcheck
- AI Incident Database: incidentdatabase.ai (catalogue of AI failures)
- ProPublica, Machine Bias: propublica.org (COMPAS investigation)