

AI as a mirror: Why its answers depend on your questions

Adaptive module

~20 min

Faced with AI, people swing between fascination and fear, two attitudes that credit it with an intelligence it doesn't have. Understanding it is becoming crucial to using it. Drawing on Shannon Vallor's idea of the "mirror," this module shows that AI doesn't think: it reflects back our data, our biases, and even our inner state. Its flaws (hallucinations, sycophancy, confusion) say something about us. Its lesson: the quality of what we get depends on what we bring.

LEARNING OBJECTIVES

- 01 Understand what AI is: an algorithm that automates, not a brain with intention or consciousness.
- 02 See AI as a mirror that reflects back our data, our biases, our confusion, and our current inner state.
- 03 Keep a critical eye on three reflections: hallucination, sycophancy, and the generic default answer.
- 04 Improve the dialogue: lasting context, permission to disagree, explicit reasoning, and rereading your own requests.

WHO IT'S FOR

Anyone who uses AI every day and wants to ask it questions with more skill.

- **Prerequisite:** none

HOW THE MODULE WORKS

Self-paced, with practice tailored to each learner.

- 1 **Self-assessment:** on 9 dimensions, to gauge current habits and guide the practice.
- 2 **Expert video:** 10 minutes, with 3 pauses for reflection.
- 3 **Practice scenarios:** the first 2 chosen from the self-assessment, then 16 optional ones. Every answer gets feedback: what it leads to, and one way to improve.
- 4 **Dated commitment:** one action to try, written by the learner or picked from the suggestions.
- 5 **Memo to download:** key takeaways, questions, commitment, personal notes and checklists.

A SAMPLE PRACTICE SCENARIO

An unhappy customer has written to both the support team and Farid's department. Lina, who heads support, thinks Farid should reply; Farid thinks it is up to support. Farid tells the AI assistant his side of the story, and it sides with him. His manager is back on Monday. What does Farid do?

ASSESSMENT AND COMPLETION

Pass mark: 80%

The score takes into account your answers to the practice scenarios, your commitment and, where relevant, any additional learning activities. Passing requires a commitment and a score of at least 80%.

METHODS & MODELS

- AI as a mirror – Shannon Vallor, *The AI Mirror* (2024)
- Algorithmic people-pleasing (sycophancy) – a recognized technical concept
- Giving AI permission to disagree – Afthonios method

SOURCES & REFERENCES

- Stanford – study on AI sycophancy (Science, 2026): models side with the user ~49% more often than a human would
- Shannon Vallor – *The AI Mirror* (2024), professor of AI ethics at the University of Edinburgh
- Vectara, "Hallucination Leaderboard" (updated May 2026): hallucination rates when summarizing documents, from 1.8% for the best model to 24.2% for the least reliable
- Dhuliawala et al., Findings of ACL (2024) – structured verification and the reliability of AI answers