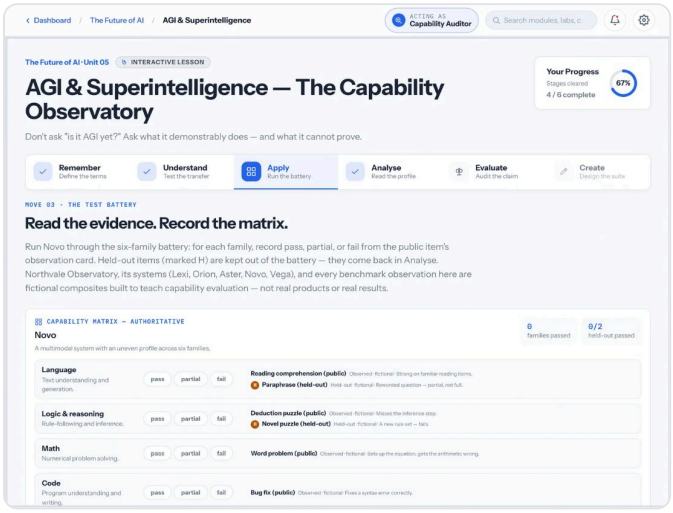
Module 08 · The Future of AI · Unit 05 · Stage 3, Responsibility

AGI & Superintelligence

The Capability Observatory

Grades 7–12 | 30 minutes | Role: **Capability Auditor**

Grade 7: About age 12–13 · Year 8 in the UK · Class 7 in India



The big idea

"Is it AGI yet?" cannot be answered with one score. Learners ask what a system demonstrably does, across which task families, and whether it transfers to tasks it never saw. A high public score can come from memorized items or a borrowed tool. A fair claim says what was tested, what failed and what is still unmeasured.

The learner's job

Learners are the **Capability Auditor** at Northvale Observatory · Capability Desk. Everyone wants to know if a system is 'approaching AGI.' You cannot answer with a number. You can answer with a matrix: what it passes, what it fails, what transfers, and what the test still cannot prove. Their task: read the matrix, check the transfer, and audit the claim — counts, not percentages.

By the end, learners can say

- I can distinguish narrow AI, generality, transfer, autonomy, AGI, and superintelligence, and separate definitions from predictions
- I can predict whether a high score in one domain transfers to a changed task
- I can run authored systems through a transparent multi-domain test battery and record pass / partial / fail with evidence
- I can inspect spiky capability profiles and diagnose memorization, tool dependence, brittle transfer, or missing embodiment/context
- I can audit a public-style 'approaching AGI' claim for definition, evidence breadth, counterexamples, and uncertainty
- I can design a capability test suite with task diversity, hidden transfer cases, failure criteria, and a limits statement

Before class

- One computer or tablet for each learner or pair, with EdSutra open (this Module is part of the full EdSutra studio).
- Open the studio once on each device before class, so the Unit still runs if the internet drops.
- Play Remember and Understand yourself, so you know how the capability matrix shows public and held-out items.
- Tell learners that Northvale, Lexi, Orion, Aster, Novo and Vega are invented, not real products or real results.
- Best after How Machines Learn and AI & Humanity.
- Units open in order, and Module 1 comes first. To start a class here, switch on Open every Unit in Settings → Teachers & parents.

Words to introduce

Narrow AI: A system that does one task well, like a spam filter.

Generality: Performing well across many different families of tasks.

Transfer: A skill from one task that still works on a changed task.

AGI: Matching human intellectual breadth: a contested definition, not a switch.

Benchmark: A standard set of tasks used to compare systems.

Held-out trial: Test items kept private, so the system cannot have seen them.

Contamination: Test items leaking into training data, so a high score proves less.

Running the lesson 40 minutes

Open 4 min

"A headline says an AI "scored near-human on 40 benchmarks" and is "approaching AGI". What would you need to see before you believed it?"

Before anyone clicks, ask: if a system aces a reading test, will it still answer when the same question is reworded? Take a vote and come back to it in Understand, when the held-out evidence opens.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Learners match seven terms, from narrow AI and transfer to AGI and superintelligence, then sort two statements into a definition and a prediction.	• Is "AGI will arrive before 2040" a definition or a prediction? • Why does it matter which one it is?	A definition says what a word means. A prediction is a dated claim about the future that evidence can't check yet.
2. Understand 6 min	Three systems ace a public task. Learners predict whether each transfers to a changed task: a reworded question, a missing calculator, a moved obstacle. Then the held-out evidence opens.	• Look at our vote. Did the reading system pass the reworded question? • Why do testers keep some items hidden?	None of the high public scores transferred. Hidden, held-out items show what a system can really do.
3. Apply 6 min	Learners run the fictional system Novo through a six-family test battery, recording pass, partial or fail for each family from its public observation card.	• How many of the six families did Novo fully pass? • Is Novo general? What else would you need to know?	Novo fully passed only language and code. Strong spots do not add up to generality.
4. Analyze 6 min	Four uneven profiles. Learners compare public and held-out items and diagnose at least two: memorization, tool dependence, brittle transfer, or missing embodiment and context.	• What gave away Lexi's memorization? • Should Orion be tested with its tools, without them, or both?	Passing public items but failing held-out ones points to memorization. Report a system with and without its tools.
5. Evaluate 7 min	Learners audit a headline saying Vega is approaching AGI after near-human scores on 40 benchmarks. Four steps: its definition, how broad its evidence is, the strongest counterexample and an honest uncertainty statement.	• What kind of tasks were all 40 benchmarks? • Which counterexample hurts the claim most?	No definition is stated, all 40 are language sets, and Vega failed a held-out logic puzzle.
6. Create 7 min	Learners design their own test suite: at least four task families, at least one hidden transfer case from a family they chose, failure criteria and a statement of what the suite cannot prove.	• Which hidden case did you add, and what would a failure tell you? • What can your suite never prove?	A good suite is broad, includes held-out items, says what counts as failing and states its limits.

Steps follow Bloom's revised taxonomy, in the order the Unit presents them.

Deepen the thinking

Critical thinking	Two experts disagree about whether a system is close to AGI. How could different definitions explain the disagreement?
AI thinking	Orion only solves math with a calculator attached. Is the calculator part of its ability? How should that be reported?
Ethics and fairness	If a company overstates what its AI can do, who could be harmed, and who should check the claim?
Connect to life	Think of a test you passed by memorizing. How would a held-out question have shown what you really knew?

Misconceptions to listen for

"One high score means the system is general."
Lexi aces public reading items but fails a reworded question it never saw. Only held-out trials show transfer.

"AGI is a single switch that flips on."
The Unit treats AGI as a contested definition. Which definition you use changes how a system is judged.

"A smarter system must be conscious."
The Unit's tests measure capability only. A suite's limits statement keeps consciousness and autonomy as separate questions.

In real life the card learners see at the end of the Unit

- Headlines saying an AI "aced an exam" don't show it can handle new or changed tasks.
- Predictions about when AGI will arrive are scenarios with dates, not facts, however confident they sound.
- A chatbot that writes great essays can still fail simple puzzles, or need a calculator tool for sums.

Try it this week. Pick a game you're good at and play it with one new rule. How much of your skill transferred?

Watch out for. One high score doesn't make a system general: it may have seen the test questions in training. And capable doesn't mean conscious.

Talk about it. What would an AI have to do to convince you it was truly general? What wouldn't count?

Exit ticket

1. What is the difference between narrow AI and generality?

2. Why do testers use held-out items?

3. Name one reason a high benchmark score might not show real ability.

Going further

For a class project, groups find a news story about an AI "acing" a test and audit it like the Vega claim: the definition used, the breadth of tasks, any counterexample and what is still unmeasured.