1.3 Responsible Use of AI

Take Home message

- **Responsibility is a design task, not a control task.** Detection of AI does not work properly.
- **And AI detection is unfair when it fails.** There is a risk of mistakenly identifying answers as AI-generated, or vice versa. Several European countries have already ruled out these tools.
- **Confident invention and inherited bias follow from how these systems are built and judged.** The consequence should be: teach them rather than waiting for a fix.
- **The working principle is: think first, then AI.** Making people commit to their own answer beats explaining the machine; building limits into the software beats banning it.
- **Prompts alone don't hold.** A chatbot tasked with acting as a tutor tends to drift back to giving answers, which makes reliable limits a purchase question.
- **Amplification is the default; equalization must be built** from access, design and scaffolding together — and the scarcest of the three is a trained teacher.{{rev}}
- **Inclusion is the clearest win**, and where teachers themselves see the strongest case.

> If you remember only one sentence from this page: **The dams are breaking. Teach swimming.**

The reading journal, second look

Back to our ninth-grade student and his suspiciously polished reading journal. The instinct is almost universal, and it is the instinct of a good teacher: *find out whether he used AI.* Build a better dam.

This chapter is about why that instinct, on its own, leads nowhere. We will come to what works instead.

Start with the uncomfortable part. **Teachers cannot reliably tell AI text from student text.** In a German study, 289 teachers of different experience were given a mix of student writing and chatbot writing and asked to sort it. Neither group could. But both groups were confident they could. Experience made almost no difference. At scale, it is worse. At one university, researchers submitted chatbot-generated answers into the real examination system through the normal channels, without notifying the markers. **94 % were never questioned**. Moreover, on average, they were graded *above* the work of actual students. So, what's the consequence? We reach for the machine! And here the finding stops being awkward and becomes an ethical problem!

The detector that fails the wrong students

Researchers tested seven of the most widely used AI detectors on TOEFL essays written by students learning English as a foreign language. Alongside, they put essays by native-speaking American eighth-graders. The native speakers' work was classified correctly almost every time. Of the foreign-language essays, **more than six in ten were wrongly flagged as machine-written** .

The mechanism is not mysterious. These tools do not detect the work of Large Language models, they detect *predictability*. Text that uses common words in expected orders scores as artificial. Anyone writing in a second language, with a smaller vocabulary and safer sentence structures, produces exactly that signal. **The tool's mistakes fall systematically on the students who are already carrying the most.** Institutions have done the arithmetic. Vanderbilt University pointed out that even if such a tool were wrong only 1 % of the time, that would still mean roughly **750 wrongly accused students a year** out of its 75,000 submissions. Consequently, they switched it off. The guidance issued by Germany's federal education ministry is blunt: the technology is "*not sufficiently reliable to provide legally secure proof.*"

Much of Europe has stopped debating this. France, the Netherlands and Switzerland advise against detectors or forbid them. The United Kingdom's exam regulator has deliberately chosen human judgment over software. The Dutch argument goes one step further and is worth carrying home: **pasting a student's work into a detector is itself an act of processing their personal data** , and doing it without permission breaks data protection law. → 1.5 Compliant Use of AI

It seems that the cheating panic is thinner than the noise around it. Researchers surveyed students at three secondary schools before ChatGPT existed and againafterward. Cheating stayed broadly where it was. Most students rejected having a chatbot simply produce their work, while thinking it fair to use one to get started or to have something explained.

The real question was never, "How *do I catch him*"*?* It is: **what does a homework task still measure, once its product can be generated on demand?**

Two things that will not be fixed, so they must be taught

Confident invention is built in. These systems generate plausible continuation of a text, which are not the same as true ones. Researchers have now formally shown where this comes from: the standard tests used to judge Large Language model reward confident guessing over admitting uncertainty, so as long as they are measured that way, a certain rate of fluent, well-formed falsehood persists. It is a property of how they are built and graded, not a bug awaiting a patch.

The slants in the training material are built in too. Moreover, they do not shrink as the technology improves. A European study of gender stereotyping across languages found that bigger

models, even after the extra training intended to make them safer and fairer, sometimes stereotype *more* rather than less. If a property cannot be engineered away, it becomes curriculum. And it can be taught: in one Finnish programme, the share of students who could identify bias in an AI system's output rose from **7 % to 44 %**. The consequence for teaching must be training error analysis instead of error avoidance. In other words: the machine's failures become the lesson.

What works: design, not control

The evidence points to one simple principle that's easy to remember: **think first, then AI.**

- **Make the human commit first.** In a series of experiments, people who had to write down their own answer *before* seeing the machine's suggestion caught far more of its mistakes than people who were shown detailed explanations of how the machine reasoned. Explaining the AI did not help to solve that issue, but forcing the judgement did. With one honest caveat: the versions that worked best were the ones participants liked least, and they helped most those who already enjoy effortful thinking. So it is a phenomenon also known as the Matthew Effect: To those who have, more will be given. They benefit the most from this approach.
- **Build the limits in.** The chatbot from Chapter 1.3, that withheld answers and asked questions instead is the same idea in software: the students who used it lost nothing on the exam, while those with the unrestricted version lost badly.
- **Make the process visible instead of hunting the product.** Drafts, working notes, a short conversation about the work. These show you what a student can do without anyone having to prove what they did.
- **Denmark is piloting the principle:** at participating schools, students may use AI while preparing for the oral English examination, and the written examination includes a handwritten, aid-free part.

Two limits are worth knowing before you rely on this. The first: the popular advice that teachers should simply instruct a chatbot to act as a Socratic tutor does not survive contact with a long conversation. At least at the moment, the system tends to drift back into supplying answers. Reliable limits have to be built into the system, which makes them a purchasing decision rather than a prompting skill. The second: making an AI explain its reasoning sounds like the obvious safeguard, but explanations do not reliably prevent over-trust. In one study they left people's acceptance of *wrong* advice unchanged, and explanations that are too technical or too simplified can increase misplaced confidence. Explanation is a condition, not a solution.

Attitude: two failure modes, not one

We worry that teachers are placing too much trust in AI. However, we should be equally concerned about the reverse. Even experienced people can be caught out by over-trusting. In one study, teachers reviewed AI-generated grades. They judged 42% of the machine's feedback to be vague or wrong, but changed only 9% of it. **Spotting an error and correcting it are different processes.**

The opposite pole is concealment. Teachers hide their own use of AI from colleagues and students, forbidding them to do what they do themselves: 77% use AI privately, while 57% say students should never use it to generate ideas. Meanwhile, a study of academically able students in grades 6 to 8 found that between a third and two-fifths already said the AI knows more than their teachers. **A profession that hides its practices cannot model responsible practice**, and transparency is not just a courtesy. It is the mechanism by which teaching by example works.

Does AI level the field or tilt it?

For work-related tasks, AI evens out differences. In one experiment, participants with higher levels of education performed significantly better than those with lower levels of education. When each person was given an AI assistant, these differences disappeared almost entirely. Then take the assistant away again, and much of it came back. **What the AI closed was the gap in output, not the gap in ability.**

In learning tasks, it does the opposite. One study found no average benefit at all, but a widening distance between students who already knew a lot and students who did not. Another looked at students writing in a foreign language: the weaker writers received *more* corrections from the AI and successfully acted on *fewer* of them, and several found the flood of feedback dispiriting. **Feedback without the ability to sort it does not help; it buries.**

The gap also exists before anyone touches a tool. In Germany, **80 % of students from the most advantaged homes see AI as an opportunity, against 55 % from the least advantaged**. In Britain, the divide runs through the staffroom rather than the device cupboard: **45 % of teachers at private schools have had formal training in using AI, against 21 % at state schools.** Within the state sector, wealthier schools again outpace poorer ones. A survey of children across 20 European countries finds the same social split in how much and how variedly they use AI. One line captures the whole problem better than any figure: "*the rich have access to technology and people to help them use it, while the poor have access to technology only.*" Which is the strongest available argument that teachers matter to close this gap.

So the bottleneck is what learner has a teacher who has been trained. **AI compensates when three things hold together: secure access, a tool designed for learning rather than answering, and a teacher scaffolding its use.** Remove any one, and it amplifies instead. Since all three tend to be present in the same schools and absent in the same schools, **amplification is the default and equalisation is something you have to build.**

Revision #11
Created 2026-06-23 07:10:56 UTC by Flavia
Updated 2026-08-11 15:25:58 UTC by Weitzel