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.** Detectors wrongly flag human work as AI-written, and miss AI-written work. Education authorities in France and the Netherlands advise against using them.
- **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 and building limits into the software removes the harm that unrestricted access does.
- **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; equalisation must be built** from access, design and scaffolding together. The scarcest but most important of the three is a trained teacher.
- **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 two studies by a German team, 89 trainee teachers and 200 experienced teachers were given a mix of student essays and chatbot essays and asked to sort them. Neither group could do so reliably but both groups were confident they were able to. Experience helped only a little: the experienced teachers still missed most of the AI texts. At one British university, researchers submitted chatbot-generated answers to real unsupervised online exams through the normal

channels, without telling the markers.

Voices from the field

“You may use it, but you must be able to explain the result in your own words.”

The detector that fails the wrong students

Researchers tested seven of the most widely used AI detectors on TOEFL essays written by people learning English as a foreign language, collected from a Chinese online forum. 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. Many of these tools do not detect the work of large language models. The tools measure *predictability*. Text that uses common words in expected orders scores as artificial. Writers in a second language, with a smaller vocabulary and more rigid sentence structures, can produce exactly that signal. Meanwhile, detectors trained for one specific test, and studies in other languages, show smaller differences. Up to now there are no detectors on the market that are reliable enough to accuse anyone of willful deception.

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 papers a year wrongly flagged** out of its 75,000 submissions. Consequently, they switched it off. A guide written by researchers for Germany's federal education ministry is blunt: the technology is "*not sufficiently reliable to provide legally secure proof.*" An independent European test of fourteen detectors reached the same verdict: "*neither accurate nor reliable.*"

Much of Europe has moved on from this debate. France's education ministry and its data protection authority advise schools not to use detectors, the ICT agency Kennisnet in Denmark calls them highly unreliable. In higher education individual universities, such as the ZHAW in Zurich, forbid using them for grading. In England, the exam boards allow detection software only as one element of a judgement that rests on the teacher, who "*knows their students best*". A data-protection point is worth carrying home as well: **pasting a student's work into an online tool means processing their personal data.** For this purpose, a legal basis and a contract with the provider is required. → 1.5 Compliant Use of AI

It seems that the cheating panic is thinner than the noise around it. Researchers surveyed students at three US high schools before ChatGPT existed and again afterwards. Cheating stayed broadly where it was. A follow-up survey of more than 4,000 students at six other US high schools a year later found the same.

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?**

Voices from the field

“At first I saw this as cheating. Now I see it as a signal that the task itself has to change.”

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

Confident invention is built in. These systems generate plausible continuations of a text, and plausible is not the same as true. Reviews of the research agree that no technique has yet removed the problem, although it can be reduced and partly detected. The reason is, that the standard tests used to judge large language models reward a confident guess over admitting uncertainty, so as long as they are measured that way, a certain rate of fluent, well-formed falsehood will persist.

The slants in the training material are built in too. Moreover, they do not shrink as the models improve. 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 explain where an image generator's bias came from 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.

Voices from the field

“Let the learner check the chatbot's solution as if they were the teacher. Looking for the mistakes is how they finally learn the method.”

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 an experiment with about 200 adults, people who had to make their own decision *before* seeing the machine's suggestion went along with its wrong suggestions less often than people who were shown a short explanation of the machine's reasoning. Explaining the AI did not solve the problem. The method that helped

was forcing the judgement. Forcing judgement means forcing people to think. Since thinking is exhausting the participants liked these methods least, and they helped most those who already enjoy effortful thinking. That is what your all know as the Matthew effect: More is give to those who already have.

- **Build the limits in.** The chatbot from Chapter 1.2 , 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 may not survive a long conversation. Early evidence, from a study with simulated students that has not yet been peer-reviewed (October 2026), suggests that under persistent pressure such tutors 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 only a condition for learning.

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 a small pilot in US schools, four in ten of the teachers surveyed judged an AI grading tool's feedback not useful, mostly vague, or even incorrect. Research on automation bias shows that people tend to go along with a machine's suggestion even when they have reason to doubt it. **Spotting an error and correcting it are different steps.**

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. With an AI assistant, about three-quarters of that gap disappeared. When the assistant was taken away again, a sizeable part of the gap 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 2026 survey of more than 25,000 children across 17 European countries finds the same social split in how much and how variedly they use AI [PRÜFEN: SES-Befund am Bericht]. One commentator's line captures the whole problem better than any figure: "*the rich have access to both technology and people to help them use it, while the poor have access to technology only.*"

So the bottleneck is which 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 #6
Created 2026-10-20 11:34:13 UTC by Jannik
Updated 2026-10-20 11:34:15 UTC by Jannik