1.1 Potentials & Risks

> *If you remember only one sentence form this page:* AI has made its way into schools like a medicine that comes without a patient information leaflet. As a teacher, you don’t know when to use AI, you lack dosage instructions, and there is no information on the side effects. This chapter serves as the patient information leaflet for the Playbook.

Should AI come with a leaflet like a medicine?

You know that slip of paper tucked inside every medicine box? The one you likely toss aside without a glance? It may seem dull, but this little paper is a quiet triumph of the last century. That leaflet spells out four essentials: what the medicine treats, how much to take, the surprises it might bring, and who should steer clear.

Behind every line on that leaflet stands a whole machinery: clinical trials, regulators, and a responsibility to reveal mistakes. No one believes the leaflet itself heals you. Its real power is making the medicine safe for anyone. That is the magic we are missing when we let AI teach without guidance.

The AI tools now quietly settling into your school arrived with no instructions. In the UK, a review found that only 7 out of 100 education technology companies had ever put their products through a controlled trial, and only 12 out of 100 had sought outside certification. The classroom tells the same story. Only 11 out of 100 school leaders and teachers across 17 American states had ever requested peer-reviewed proof before adopting a new tool.

The result, in many cases, is wasted money. In the US, two-thirds of school software licenses gathered dust, and 98% saw almost no use. The EdTech Genome Project looked at about 7,000 teaching tools worth 13 billion dollars and found that 85% were either mismatched or misused.

This chapter may look like the dry, necessary section, but it reveals the hidden stories behind the rules. It tells what really happened, who was affected, and shines a light on the side effects, even those that quietly reached children who never signed up.

Potentials of AI

Let's make this clear right from the start. Most studies on AI's potential focus on university students. School pupils have so far featured relatively rarely in these studies. Most studies relate to

short-term interventions lasting from a few hours to a few weeks. Consequently, the so-called novelty effect may have positively influenced the results. Last but not least, the test used to assess effectiveness was often devised by the very people who delivered the lessons. This, too, may have confounded the data.

(1) Let AI explain a new topic and keep the practice for your lesson

Imagine this. Tomorrow, when a pupil misses a topic or needs that tricky concept explained yet again, let an AI tutor handle the first explanation. Reserve the practice for your lesson, where real learning comes alive.

Why this matters. AI can support acquisition of factual knowledge. In a Harvard study, 194 physics students learned two topics. For one topic, they used an AI tutor developed for this purpose. For the other topic, they learned in a traditional class. Each student completed both approaches, so no one can claim that the stronger group received the better treatment. In a test conducted immediately afterward, the students taught by the AI achieved significantly higher results. The researchers recorded a median time of 49 minutes to complete the task, compared with the 60 minutes allocated for the lesson in the timetable. The authors are cautious in their interpretation of these results: the material was new to the students, and much of the task simply involved explaining it well. They expressly reject the idea that the same would apply to tasks requiring the integration of multiple ideas. Furthermore, these were undergraduate students, not Year 8 pupils.

A large review pooling the results of 228 studies came to a similar conclusion. Generative AI-based chatbots produced a larger effect on knowing (and understanding) than intelligent tutoring systems and adaptive practice software.

Go deeper. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15,* Article 17458. Free access, DOI 10.1038/s41598-025-97652-6. The article describes how the tutor was designed and controlled. This is useful for anyone wishing to develop an AI tutor themselves.

(2) Making better use of lesson planning time

Imagine this. Save some of the time you need for lesson planning by using AI to draft your first lesson plan and then revise it thoroughly.

Why this matters. No one disputes that teachers' workloads are heavy. Therefore, one potential benefit of AI is using working hours more efficiently. A UK study investigated the extent to which the targeted use of AI for lesson planning can help to reduce teachers' workload. Researchers randomly assigned 259 science teachers at 68 schools to either a group using a chatbot alongside brief written instructions or a group that continued lesson planning as before. **The weekly planning time for science lessons in Years 7 and 8 fell from around 81 minutes to around 56.** A panel of subject experts assessed the quality of the lesson plans. They did not know whether the plans had been created with or without the aid of AI. The experts found no evidence to suggest that the quality of the lesson plans differed. However, two points must be highlighted.

(1) **The teachers used the tool less and less as the weeks went on,** meaning that the time saved does not occur automatically and does not sustain itself. (2) The trial measured the time spent by teachers on lesson planning, not their pupils' learning outcomes.

Go deeper. Roy, P., Poet, H., Staunton, R., Aston, K., & Thomas, D. (2024, 12 December). *ChatGPT in lesson preparation: A Teacher Choices Trial.* NFER, commissioned by the Education Endowment Foundation and the Hg Foundation. Free at educationendowmentfoundation.org.uk.

(3) Providing better support for learners who do not benefit fully from standard lessons

Imagine this. Your school decides how it wishes to spend its funds. It wants to purchase an AI chatbot. How likely do you think it is that the chatbot providers will provide you with evidence of the bot's benefits for different subjects?

Why this matters. The provider may point out the following application. Chatbots are the oldest and best-documented application of 'AI' in education, and this is precisely what is often overlooked when allocating funds. In an analysis of 29 studies involving 41 groups of learners with disabilities, researchers found moderate benefits across a wide range of settings. But read the small print. In **seven out of ten of these studies**, 'AI' meant a **robot**: a small physical machine sitting on a table, often operated by a researcher in another room. Only one in five studies used software. So this is no proof that a chatbot will help an autistic pupil in your class. Nor could the reviewers say **what** makes these interventions effective. Based on their research, they could not identify why the studies that worked differed from those that did not.

Go deeper. Zhang, L., Carter, R. A., Jr., Liu, Y., & Peng, P. (2026). Let's CHAT about artificial intelligence for students with disabilities: A systematic literature review and meta-analysis. *Review of Educational Research, 96* (1). DOI 10.3102/00346543241293424. The real gem is the table of included studies: it instantly reveals just how little of this field actually focuses on the tools currently being marketed to schools.

(4) Teaching learners how AI fails

Imagine this. Ask your class to create ten images of 'a doctor' and ten of 'a nurse.' Tally up who appears in each set, then spark a discussion: where did these patterns come from, and who made those choices?

Fig 1. Image created with NanoBanana 2, created August, 20, 2026, One shot prompt: a doctor, a nurse

Why this matters. The classic version of this lesson feels outdated, and that is exactly why it deserves attention. For years, the demonstration was predictable: ask for a doctor, and a man appears. In one systematic test, Midjourney almost always pictured doctors as white men, while Adobe Firefly made visible, if imperfect, attempts to diversify.

Try the same prompt today, and you might see a female doctor or a male nurse appear (Fig. 1). But do not be fooled into thinking the bias has vanished. A 2026 analysis of 1,344 images from three popular generators revealed the opposite: when asked for a 'competent person,' one system showed no women.

What changed is not the bias but the layer it lives in - and this the important lesson! Human decisions stand behind any of these pictures: what the system was trained on, and what its makers afterward decided it should show. The second is adjustable, invisible, and made by a company.

The first layer of human decisions is what the system has absorbed. Millions of images, crafted and labeled by people, each one echoing the world as it already was. This is the layer that the classic bias lesson reveals, and pupils grasp it easily: the machine became a reflection of what it was fed.

The second layer is shaped by what a company chooses to show you next, yet this influence leaves no trace in the image itself. Three things unfold here, each one described by the creators in their own words.

- **Your words are rewritten before the model receives them.** From OpenAI's own developer documentation: "we use GPT-4 to optimize all of your prompts before they're passed to DALL-E" — and "this feature isn't able to be disabled at the moment." What your pupil typed is not what the machine was asked.
- **Appearance gets filled in when the prompt leaves it open.** OpenAI, describing its diversity technique: "This technique is applied at the system level when DALL·E is given a prompt describing a person that does not specify race or gender, like 'firefighter.'" Afterwards, users were twelve times more likely to say the images showed people of diverse backgrounds. "A doctor" is exactly that kind of prompt.
- **And that tuning can miss.** Google, after pausing Gemini's image generation in 2024: "Our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range."

So when a pupil types "a doctor" and a woman appears, the image offers no clue about which layer brought her into being. **That uncertainty is not a flaw in the lesson. It is the lesson itself.** The old story claimed the machine simply soaked up society's prejudices. The new story cuts deeper and lasts longer: the image is no mirror, and it is no accident. A company made a choice! You cannot spot that choice in the outcome, and it might shift on a Tuesday, quietly, without warning. That truth endures through every new model. The stereotype lesson fades away.

This is the lesson that deserves a student's attention, and it will outlast the next model update: **the image is not a simple reflection of reality, and it did not happen by chance. Someone made a choice.**

This lesson can start early. When 209 Finnish students in 12 classes, grades 4 and 7, explored this topic, the number who could explain the bias using data jumped from about 7 out of 100 to 44 out of 100. The honest reservation: the study measured what children could explain, not what they would do later.

Go deeper. Vartiainen, H., et al (2025). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New Media & Society, 27*(9). Free, DOI 10.1177/14614448241252820

Weinmann, H., et al (2026). Gender bias in text-to-image generative artificial intelligence: Neglect and stereotypical presentations across three popular platforms. *New Media & Society,* Online First

How solid are the research results on the potentials of AI at this very moment?

The most important thing. If someone quotes a study to you, ask two questions before agreeing: "Who were the learners, and how long did the study last?"

Why this matters. In 2025, a meta-analysis of 51 studies reported a **significant benefit** of ChatGPT for pupils' academic performance. The article was read hundreds of thousands of times

and served as the basis for numerous confident statements in staff rooms. **On 22 April 2026, the journal retracted the article.** A *retraction* means that a journal officially withdraws something it has published; the article remains visible, marked with the note ‘RETRACTED’, so that anyone who has cited it can recognise this. Two external researchers had identified inconsistencies in the synthesis of the studies. The journal's editor wrote that the problems ‘undermine the editor’s confidence in the validity of the analysis’. The authors did not respond to correspondence on this matter.

This does not mean that AI has no impact. There are reputable studies that demonstrate moderate benefits. This field of research is also not yet well defined because it is relatively new and highly dynamic. It means that the figure that was frequently cited was incorrect, and that it took a year and two external experts to realise this.

Another reservation worth mentioning is: The closer the evidence gets to school, at least at the moment, the smaller the effect.

Go deeper. The retraction itself is one page, free, and worth reading precisely because it is short: Wang, J., & Fan, W. (2026). Retraction note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking. *Humanities and Social Sciences Communications, 13*, 528. DOI 10.1057/s41599-026-07310-z.

The **retracted** original was *HSSC 12,* 621 (2025). Still standing: Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). *HSSC, 13*, 684. Free, DOI 10.1057/s41599-026-07019-z.
The school-only figure: Yi, L., Liu, D., Jiang, T., & Xian, Y. (2025). *International Journal of Science and Mathematics Education, 23*(4), 1105–1126. DOI 10.1007/s10763-024-10499-7.

Risks of AI

(1) Doing something is not the same as learning

Imagine this. Before you set a homework task that a chatbot can complete, think about what the purpose of the task is. Might it be a good idea to change the task?

Why this matters. Do you learn faster with a chatbot? Around a thousand pupils in Years 9 to 11 practised maths in three groups: one using a standard chatbot, one using a version programmed to withhold the answer and ask a follow-up question, and one without a chatbot. Whilst using the chatbot, the group with the standard chatbot completed 48 percent more practice exercises than the group without one. Afterwards, everyone sat the same exam without a chatbot. The result, which seemed surprising at first, was that the group using the standard chatbot performed 17 percent worse than the pupils who had not used a chatbot. The group that had worked with the chatbot which withheld the answer showed no drop in performance in the test compared to the group without a chatbot, but practised more. The pupils using the standard chatbot were neither lazy nor did they cheat. They worked harder and learnt less, because what the tool eliminated was not the effort, but the cognitive challenge that constitutes learning.

Go deeper. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS, 122*, e2422633122. DOI 10.1073/pnas.2422633122.

(2) AI Detectors punish the wrong pupils

Imagine this. You receive a student's essay to grade. Curious about its authenticity, you feed it into an AI detector. The result flashes: the essay was almost certainly written by AI.

Why it matters. You use the AI detector to ensure fairness, **yet for a particular group, the detectors have exactly the opposite effect.** A detector does not recognise machine-generated text. It measures how predictable the wording is. And that is precisely where the problem lies, for example, when a pupil is writing in a foreign language, is unpractised in writing, or has a limited everyday vocabulary. Their writing is predictable. When seven commercial detectors analysed 91 essays written by non-native speakers sitting an English exam, these essays were incorrectly flagged as machine-generated in around six out of ten cases. By contrast, essays written by 14-year-old Americans in their native language, English, were almost all classified correctly. A tool that triggers false positives for exactly one group is not neutral. Furthermore, this group is the least able to defend itself.

Go deeper. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. Free, DOI 10.1016/j.patter.2023.100779.

(3) Where on earth will the pupils' data be?

Imagine this. You use an AI tool your school offers. Try to find out where the pupils' data go.

Why it matters. Can your school ensure that the data is not passed on to a third party? A third party is company that is neither your school nor the app's developer. Typically, these are advertising or analytics companies that receive data whilst the pupil is working. No one at the school chose this company. When researchers carried out a technical analysis of 163 learning products recommended by governments during the pandemic-related school closures, 145 of them processed children's data in a way that jeopardised or infringed their rights by transmitting data to 196 third-party company or granting them access, predominantly from the advertising industry. **Of the 42 governments that developed their own products rather than purchasing them, 39 had developed products with the same problem. The study covers products from 2021, and some may have changed since then**. However, we cannot rule out that the same applies to AI applications.

Go deeper. Human Rights Watch (2022). "*How dare they peep into my private life?:* Children's rights violations by governments that endorsed online learning during the Covid-19 pandemic. Free at hrw.org. The country annexes let you look up what was endorsed where.

(4) AI arrives last where it is needed most

Imagine this. How much training did you already receive on learning about AI or subject-specific teaching with AI?

Why this matters. Adults who had completed their education were provided with an AI assistant. The performance gap between those with higher and lower levels of formal education narrowed significantly. When the researchers then withdrew the assistant, the gap partially re-emerged. The researchers concluded that the AI assistant had carried out the work, but had not helped participants develop any skills.

A survey of American school principals revealed the following picture: the more disadvantaged a school's pupils were, the less impact AI had on teaching.

If we combine the findings from both studies, they allow us to make the following statement: A good tool can bridge a gap as long as it is in the hands of a learner. However, it usually does not reach those who need it, and leaves no skills behind when it is removed.

German figures show the same patte. Among 1,590 young people aged 14 to 20, 80 percent in the most affluent families and 55 percent in the least affluent saw AI as an opportunity. And recognition is followed by action: 70 percent of grammar school pupils used AI for homework, compared with 58 percent at lower secondary schools.

What about teacher professional development? A survey of more than ten thousand teachers in England found that 45 percent of teachers at private schools had already received training on AI. This compared with 21 percent at state schools. Within the state school system, the figures ranged from 26 percent at the wealthiest schools to 18 percent at the poorest schools.

(5) What improves when learning with AI is not necessarily what you see

Imagine this. You introduce a new ChatBot into your teaching and the pupils enthusiastic about using the new tool.

Why this matters. A large review found four different things about learning (Knowing and understanding, working with AI, engagement, motivation, confidence, ability to plan, monitor and correct their own work). The last one didn't improve whereas the others did.

- For knowing and understanding generative AI had larger effects than intelligent tutoring systems and adaptive practice software that is already used in schools.
- For applying knowledge, and for motivation and engagement, AI did not perform better that already existing software.
- For self-steering we do not have enough studies to tell any effect.

Here's a quick side note: t**he greatest impacts are found in the arts and humanities, rather than in mathematics or the natural sciences.**

Go deeper. Yeo, G. H., & Lansford, J. E. (2025). Effects of artificial intelligence on educational functioning: A review and meta-analysis. *Educational Psychology Review, 37*(4), article 110. DOI 10.1007/s10648-025-10085-5.

Revision #18
Created 2026-06-11 14:44:45 UTC by Jannik
Updated 2026-09-15 12:55:30 UTC by Weitzel