

White Paper



# AI and the Future of Pedagogy

Authored by:

**Tom Chatfield**





# Introduction

---

This paper explores how education can respond wisely and imaginatively to the rise of AI in general, and generative AI in particular. Its recommendations are rooted in two principles.

1. Innovation must draw on what we know about how humans learn.
2. AI's power must not be allowed to hollow out the very skills required to navigate an AI age successfully.

The tension at the heart of this second point bears spelling out. Within the space of a few years, it has become possible to *simulate* knowledge and understanding of almost any topic while possessing neither. Freely available tools can be used to complete conventional assignments and assessments with ease, in the process potentially preventing students from gaining the very skills required to use AI adeptly: critical discernment, domain expertise, research and verification, analytical reasoning.<sup>1</sup>

Already, the use of AI by students is almost ubiquitous. In a December 2024 survey, 92 per cent of British undergraduates reported using AI tools.<sup>2</sup> As the author and Washington University professor Ian Bogost noted in August 2025, 'the technology is no longer just a curiosity or a way to cheat; it is a habit, as ubiquitous on campus as eating processed foods or scrolling social media.'<sup>3</sup> From creating study guides to explaining key concepts, from constructing quizzes to proofreading essays (not to mention grading papers and drafting student feedback), AI is being used across the entire spectrum of educational activities. Yet much of this usage exists in a shadowy zone of mistrust and ambiguity. Both students and instructors have at their disposal a seductively potent tool that is upending the practices and the principles of 21st century education.

The greatest risks, here, are bound up with the unique nature of learning. By definition, learning cannot be automated or outsourced. To learn is to acquire knowledge and understanding via the meaningful exercise of a range of skills. As the author Nicholas Carr put it in May 2025, 'to automate learning is to subvert learning.'<sup>4</sup> When children at a primary school write stories and draw pictures, the point is not to supply the world with a stream of winsome content. It is to help them become literate, reflective participants in their society. Undergraduates do not write essays or conduct experiments because the world needs more such things. The process *is* the purpose.

This is not to deplore technology. When used well, technologies like generative AI can facilitate rich and potentially transformative forms of learning; while the reflective and critically discerning use of information systems is a fundamental modern literacy. But saying this only begs the question. What does it mean to introduce novel forms of automation into education *wisely*? How can the demoralizing or corrosive effects of misuse and abuse be mitigated? This paper addresses these questions across four sections.

---

1 A note on the use of the term 'artificial intelligence.' In terms of AI literacy, it is important to note that AI as a field is far broader and older than the recent subset of generative AI, which in turn is a subset of machine learning and deep learning. For the purposes of readability and simplicity, however, I have avoided technical terms and fine distinctions where possible in this paper, gesturing towards the rise of powerful generative AI in broad terms.

2 Freeman, J. (2025, February) *Student generative AI survey 2025*. HEPI Policy Note 61. [www.hepi.ac.uk/reports/student-generative-ai-survey-2025/](http://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/)

3 Bogost, I (2025, August 17). *College students have already changed forever*. The Atlantic. [www.theatlantic.com/technology/archive/2025/08/ai-college-class-of-2026/683901/](https://www.theatlantic.com/technology/archive/2025/08/ai-college-class-of-2026/683901/)

4 Carr, N. (2025, May 27) *The myth of automated learning*. New Cartographies. [www.newcartographies.com/p/the-myth-of-automated-learning](http://www.newcartographies.com/p/the-myth-of-automated-learning)

**Suggested citation:** Chatfield, T (2025). *AI and the future of pedagogy (White Paper)*. London: Sage. doi: 10.4135/wp520172

1. *What works in Education?* Faced by fresh threats and opportunities, it is vital to dig into human fundamentals and to ground innovation in best practices around the essential purposes and processes of learning.
2. *Leveraging Technology for Excellence.* The history of technology in education long predates artificial intelligence. Emerging insights need to be combined with larger lessons about what has and has not worked before (and why).
3. *Beyond the Arms Race.* Many established forms of assessment are in crisis, with AI fueling distrust and conflict between learners, instructors and institutions. In the long term, everyone loses if education becomes an arms race.
4. *Raising the Cognitive Bar.* The ultimate prize of any educational technology is that it raises the bar of human insight and achievement. This section explores how AI can deepen both understanding and engagement, and what the educator and learner of the future might look like.

Ultimately, as the psychologist and philosopher Alison Gopnik has argued, it is a dangerous category error to treat artificial intelligence as either miraculous or unprecedented. Rather, it is the latest in a long line of cultural and social technologies that includes 'writing, print, libraries, the Internet, and even language itself': ways of making information gathered or created by other human beings useful in new contexts.<sup>5</sup>

Similarly, no matter how powerful AI systems become, they will only ever be part of a learning process: components within a larger system whose ultimate measures of failure or success remain wholly human. Drawing on long-standing research and new forms of practice, this paper thus argues for education systems that treat AI not as a shortcut but as a context and catalyst for deeper learning—and for educators to play a central role in shaping such a future.

---

<sup>5</sup> Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 19(5), 874–883. <https://doi.org/10.1177/17456916231201401> (Original work published 2024)

# 1

## What Works in Education?

---

Before discussing the place of any technology in education, it is important to consider the fundamentals of *how* humans learn. This section explores these questions across three interwoven dimensions: the cognitive mechanisms governing learning, the instructional approaches facilitating it, and the social contexts within which education exists.

### The cognitive dimension

As the Australian educational psychologist John Sweller noted in 2006, the brain is a ‘natural information processing system and, as such, should not be considered in isolation from the rest of nature.’ Acknowledging this means helping learners manage their *cognitive load* in the face of challenging tasks; and, in particular, the demands being made of working memory, which easily becomes overwhelmed in the face of distraction or too much information.<sup>1</sup>

Balancing support with engagement is crucial. *Worked examples*, which take students step-by-step through solutions to problems, allow them to focus on reasoning and strategy by relieving the pressure to leap to answers. In the longer term, *spacing* sessions over time inculcates knowledge more effectively than massed practice.<sup>2</sup> *Interleaving*, by mixing different problems or concepts rather than blocking them together, similarly supports discrimination between ideas and strategies. And *elaboration*, explaining how new information relates to what is already known, makes knowledge more applicable and accessible.<sup>3</sup>

Underlying all of these is *metacognition*: the ability to think about your own thinking. Metacognition is neither distinct nor detachable from recall and knowledge. By learning to monitor their understanding and adjust their strategies, however, learners can acquire a form of educational agency that goes beyond any particular subject.<sup>4</sup>

These principles gesture towards a fundamental truth. Learning is an active process of construction, and supporting it involves tools, practices and technologies that uplift rather than undermine active sense-making. Similarly, cognition is both *embodied* in individual lives and *extended* by the cultural, technological and social context surrounding these lives. To design better systems and tools is to ask what matters most within these systems; then how to conduct an informed negotiation between the novel and the enduring.

---

1 See Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, 16(2), 165–169. <https://doi.org/10.1016/j.learninstruc.2006.02.005>. and Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)

2 See Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, 17(2), 77–88. <https://doi.org/10.1038/nrn.2015.18>

3 See Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, 107(3), 900–908. <https://doi.org/10.1037/edu0000001> and Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266> (Original work published 2013)

4 Muijs, D. & Bokhove, C. (2020). *Metacognition and self regulation: Evidence review*. Education Endowment Foundation. <https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/metacognition-and-self-regulation>

## The instructional dimension

Where cognitive science suggests how minds learn, instructional research suggests how teaching can support this process. Deep debates exist around best practices for both teaching and learning. But meta-analyses of thousands of studies converge on several key principles.<sup>5</sup>

First, clarity of purpose matters. Students need to understand not just *what* they are learning but *why*, with clear goals giving direction and form to their efforts. To ensure these goals cohere over time, learners also need *scaffolding*: the careful structuring of tasks so that each builds on the last. Badly designed instructional tools and approaches can overwhelm and demoralize learners, leaving them struggling to work out what is required.<sup>6</sup>

Second, *well-calibrated feedback* is essential for showing learners what they have achieved and how to improve. The best feedback creates a dialogue, turning assessment into a continuation of learning rather than just its measurement. Good teachers know how to adapt their approach while encouraging reflection and curiosity.

It is important to note, that contrasting idealized human educators with technologies like AI offers a skewed picture. At their best, such technologies promise scalable forms of engagement and personalization able to redress substantial gaps and disadvantages. At their worst, they threaten to corrode or overwhelm existing systems, placing additional demands upon educators and institutions ill-equipped to afford them.

Beyond over-simple talk of outsourcing education to machines, the key question is how *actual* learners and educators can best be supported within different institutions, complete with their pressures and constraints: a question that can in turn unlock nuanced, discipline-specific discussions of where and how different approaches might be developed.

## The social dimension

Human learning is never just cognitive or instructional. People are sociable, embodied creatures who thrive on *constructive debate* and *meaningful feedback*, and suffer in their absence. When students see their perspectives treated with respect and their struggles understood, engagement is liable to follow. The classroom becomes a space not just for personal achievement but for sociable sense-making.<sup>7</sup>

*Emotional safety* also matters. Learning means being able to acknowledge errors and uncertainty in front of others. Classrooms that treat mistakes as useful information rather than evidence of failure can make vulnerability productive.<sup>8</sup> By contrast, those that surveil and penalize students for infractions of a system they have little agency within—or fail to offer guidance, trust and access around tools like AI that are already seen as integral to life and work—are likely over time to fall behind.<sup>9</sup>

---

5 For a substantial overview of the evidence covered in this section see Hattie, J. (2023). *Visible learning: The sequel. A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

6 For an in-depth exploration of the challenges around educational technology and cognition, see the first half in particular of the 2024 Purposeful ventures report by Monis-Weston, D. *Exploring edtech and AI in maths teaching* online at: <https://cdn.sanity.io/files/uiql8se4/production/d66cc116d4630b2196d92522bf4593781ff037eb.pdf>

7 See Hammond, Z (2014) *Culturally responsive teaching and The brain: Promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin.

8 Gedikoglu, M. (2021) *Social and emotional learning: An evidence review and synthesis*. Education Policy Institute. <https://epi.org.uk/publications-and-research/social-and-emotional-learning/>

9 See Umar, W. (2025) *A student manifesto for assessment in the age of AI*. The London School of Economics and Political Science. <https://info.lse.ac.uk/staff/divisions/Eden-Centre/Assets-EC/Documents/PKU-LSE-Conf-April-2025/LSE-PKU-Student-Manifesto.pdf>

Thanks to their ability to simulate social interactions, large language models (LLM) present particularly remarkable challenges and opportunities here. When suitably instructed, they are able to play the role of an infinitely patient explainer, a knowledgeable interlocutor, a tireless writing partner. But it is vital not to confuse simulation with an authentic exchange of perspectives. An LLM has no stake in a conversation and no mind to change. It will flatter, cajole or concede points in line with its training. But it can neither care nor know anything about a learner's life beyond the screen.

## Key principles

In the light of the above, what does it mean to prepare students for 21st-century realities rather than treating new technologies as either miraculous solutions or engines of pure efficiency? Here is UCL Professor Emerita of Learner Centred Design, Rose Luckin, writing in June 2025:

*As our world integrates more AI into every aspect of life and work, our students need something deeper than efficient content delivery. They need sophisticated meta-learning capabilities—the ability to adapt, change, and evolve their thinking as the world transforms around them.<sup>10</sup>*

As this suggests, the back-and-forth of debate and building of common understandings are not inefficiencies to be optimized away, but the mechanisms through which learning occurs. When students challenge one another's assumptions and make reasoned arguments, they are doing more than exchanging information. They are learning to think.

Similarly, education exists within networks of relationships, cultures and communities, and grasping these demands the insights of diverse fields and disciplines: psychology, sociology, anthropology, history; management, communication, business, education. In an age of ever-more-powerful systems trained upon petabytes of cultural and behavioural data, it is impossible to design effective systems *without* literacy in the science of human behaviours and societies. To ground innovation in the best of this knowledge, then, educators and institutions should:

1. **Help Learners Manage Cognitive Load:** Design environments that help learners manage mental effort wisely and push back against distraction and noise. In the context of attention-hungry technologies and potentially overwhelming information, finding focus and seeking signals amid noise are vital precursors to learning and self-determination. Technologies like AI are also especially well-equipped to help explain fundamentals.
2. **Explicitly Teach Metacognition:** Much like taking control of time and attention, learners need to equip themselves to monitor and adapt their thinking in the face of a rapidly changing world. Metacognitive practices are central to education and the discerning use of technology, both for identifying opportunities and resisting pitfalls; and learning to apply them effectively is predicated as much on emotional as intellectual self-control.
3. **Celebrate Social Sense-Making and Intellectual Risk-Taking:** The more capable automated systems become, the more interpersonal skills become a vital human enabler and differentiator. Use techniques like teaching, group problem-solving and collaborative reflection to tap into the affective and intellectual benefits of shared learning. Celebrate curiosity, the honest expression of uncertainty and empathetic reflection. Acknowledge the everyday needs and experiences of students to build guidance, access and trust around technology.

---

<sup>10</sup> Luckin, R (2025, June 11). *Sal Khan's AI keynote left me deeply concerned*. LinkedIn. [www.linkedin.com/feed/update/urn:li:activity:7338405785263132672/](https://www.linkedin.com/feed/update/urn:li:activity:7338405785263132672/)



# 2

## Leveraging Technology for Excellence

---

Having explored the fundamentals of human learning, this section examines some of the underlying patterns that characterize the successes and failures of different learning technologies. It then examines emerging best practices around AI, presenting a concrete case study of the author's own attempts to turn these principles into practice.

### The geography of knowledge

Perhaps the world's most enduring and successful learning technology is so adaptable and fit for purpose it is easy to overlook the lessons it can still teach: the book.

This ancient technology partly succeeds because it creates a 'geography of knowledge': bounded, navigable, with clear landmarks and pathways. A well-designed textbook provides natural scaffolding through its linear structure. Its permanence supports recall and deep focus. In physical form, a book's tactility and heft enlist multiple senses, providing spatial cues that help anchor ideas in memory. This is one reason students may remember a theorem as 'near the back on a left-hand page,' much as the active business of handwritten note-taking supports recall and comprehension<sup>1</sup> As a companion to learners and a complement to instruction and debate, a physical book has few equals; not to mention an unrivalled battery and shelf life.

Digital technologies, by contrast, operate in vast and potentially unbounded informational environments. As printing once did to the written word, they make knowledge discoverable, shareable and reproducible; and, in the form of LLMs, increasingly customisable, interactive and generative. Speed, scale, search-ability and interactivity offer remarkable benefits. But they also demand careful design and constraint to offer *cognitive landmarks* and convey information as something more than a frictionless feed.

Much as the rise of written words impacted upon human memory and oracy, the speed, scale and flexibility of digital technologies continue to bring a complex mix of costs and benefits. Without scaffolding and selection, learners can easily become overwhelmed by endless possibilities. Like the internet and search engines before them, the very open-endedness of AI interfaces is at once empowering and potentially overwhelming. It is worth remembering that the persuasive powers and potentially deleterious effects of books were anxiously debated for centuries. Similarly, technologies like AI need to be seen as both resources and focal objects for education: power, useful and potentially problematic artefacts it is vital to use critically, ethically and with confidence.<sup>2</sup>

In this context, the choice that matters most is not so much *whether* to use novel technologies as *when* and *how* to do so discerningly. For the philosopher Andy Clark, part of the answer is a form of 'extended cognitive hygiene' that acknowledges how integrally human minds are interwoven with their informational environments, then sets about auditing and critiquing these. As technology progresses, Clark argues, individuals and institutions need to think carefully about which tools and systems they wish 'to incorporate into our new digitally extended minds'<sup>3</sup> and what skills are required to judge this.

The history of educational technology offers some painfully instructive lessons here. Despite their diversity, the technologies that endure share certain *affordances*. From blackboards to scientific calculators, they extend rather than replace human capabilities. They integrate with best practices rather than promising wholesale revolution. And they solve actual problems rather than creating solutions in search of applications.

---

1 See Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, 25(6), 1159–1168. <https://doi.org/10.1177/0956797614524581> and Hu, C (2024, February 21). Why writing by hand is better for memory and learning. *Scientific American*. [www.scientificamerican.com/article/why-writing-by-hand-is-better-for-memory-and-learning/](http://www.scientificamerican.com/article/why-writing-by-hand-is-better-for-memory-and-learning/) and Mangen, A., Walgermo, B. R. & Brønneck, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*. 58. 61–68. 10.1016/j.ijer.2012.12.002

2 This last point draws on insights and generous feedback from Professor Paul Dickerson. It is also worth emphasising the long history of anxiety over novel technologies, as reading and writing once were, from Plato's *Phaedrus* to Tyndale's Bible to fears that the Victorian novel might damage vulnerable minds.

3 Clark, A. (2025). Extending minds with generative AI. *Nature Communications*, 16, 4627. <https://doi.org/10.1038/s41467-025-59906-9>

By contrast, the ed-tech graveyard is strewn with tools that promised transformation but delivered fragmentation. Learning management systems that bureaucratized rather than enhanced students' relationships with institutions and instructors; interactive whiteboards that became expensive projection screens; personalised learning platforms that drained knowledge of context and meaning. Technology succeeds when it honours and builds upon human fundamentals, and fails when it ignores or circumvents them.

## Emerging best practices

Having accepted that the affordances of the human-made world profoundly effect *when, how* and *what* we think, the next step is to attend closely to these minutiae, with an emphasis on active reflection, problem-solving and reasoned analysis. As a May 2025 paper exploring the significance of contextual knowledge in an AI age argued:

*Looking ahead, the synergy of human cognition and machine capability will define successful education. We should strategically leverage AI and vast information stores to expand learning possibilities—personalizing instruction and creating immersive simulations—while preserving the core of what makes us intelligent.<sup>4</sup>*

In this context, it is striking that a relatively high degree of consensus already exists around foundational principles for using systems like LLMs effectively; and that many insights from established research into educational technology remains valid.

For example, studies of intelligent tutoring systems from the 1990s onward suggest that the most effective systems offer immediate, directed feedback that aims to guide students towards answers rather than solving problems for them. They offer hints rather than solutions, highlight misconceptions as well as errors, and adapt support based on understanding.<sup>5</sup> Similarly, emerging ways of embedding LLMs in structured, purposeful contexts include:

- *Offering clear guidance around fundamental expectations and literacies.* Ambiguity and anxiety around institutional policies damage both learners' and faculties' capacities to use systems like LLMs with confidence, transparency and thoughtfulness. Those using powerful technologies need to be both well-prepared and supported in doing so wisely. Effective programs teach students to recognize when AI hallucinates, exhibits bias or expresses false certainty; as well as to reflect upon its larger social context, training and the incentives of those creating it.<sup>6</sup>
- *Teaching prompt design as a form of critical thinking.* Teaching students to craft effective prompts and contexts for AI, done well, is itself a lesson in precision and clarity. When well-supported and able to reflect and learn from one another's practice, students can learn to formulate questions carefully and to refine their approach based on critical reflection. MIT's RAISE initiative is one example of this, embedding AI use within projects where students need to justify both their questions and their evaluation of responses.<sup>7</sup>
- *Restructuring tasks around judgment.* Rather than viewing AI as a threat, innovative educators can deliberately design aspects of assignments that require uniquely human contributions *and* insist on the transparent, reflective use of technology. A literature student might use AI to generate multiple interpretations of a text, then craft an argument about their limitations. A law student might deliberately be instructed to look for examples of hallucinated citations online, then explore their history, influences and consequences.

---

4 Oakley, B., Johnston, M., Chen, K., Jung, E. & Sejnowski, T. (2025, May 11) The memory paradox: Why our brains need knowledge in an age of AI (May 11, 2025). *Social Science Research Network*. Available at: <https://ssrn.com/abstract=5250447> or <http://dx.doi.org/10.2139/ssrn.5250447>

5 See Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, Ray. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167–207. [https://doi.org/10.1207/s15327809jls0402\\_2](https://doi.org/10.1207/s15327809jls0402_2) and VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>

6 The most famous and influential paper in this area remains Bender, E. M., Gebru, T. McMillan-Major, A. & Shmitchell, S. (2021, March 3–10). *On the dangers of stochastic parrots: Can language models be too big?* [Proceedings]. Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FAcCT '21), Canada. <https://doi.org/10.1145/3442188.3445922>

7 See MIT (2025). RAISE initiative. Responsible AI for social empowerment and Education. <https://raise.mit.edu/>

- *Distinguishing between uplifting and diminishing applications of technology.* The very same tool can either build capacity or create dependency. A scaffold supports students to achieve something challenging, being gradually withdrawn as competence develops. A crutch potentially prevents key capabilities from developing. Students can be encouraged both to explore this distinction and themselves co-design tasks and methodologies where AI speeds or supports the journey towards understanding, such as using it to gloss key terms, rapidly identify relevant sources, outline initial arguments or critique first drafts.<sup>8</sup>

Most immediately, navigating this future entails tracking the guidance and policies different institutions are already putting in place, aiming to learn from the best and update assumptions regularly.<sup>9</sup> More broadly, frameworks are emerging that prioritize principles over platforms, offering the chance to reflect on the *fundamentals* of education in an age of increasingly ‘intelligent’ machines. Stanford’s HELM project, for example, aims to create ‘a living benchmark for transparency,’ making the mechanics of language models visible rather than mysterious.<sup>10</sup>

Ultimately, knowing *why* something works, grasping its limitations and reasoning about exceptions matters all the more as superficial competence becomes simple to simulate.<sup>11</sup> As a June 2025 paper from the Raspberry Pi Foundation put it:

*Even in a world where AI can generate code, we will need skilled human programmers who can think critically, solve problems, and make ethical decisions.*<sup>12</sup>

The same is true of countless other AI-generated outputs. When almost any question can be answered instantly, it is all the more important to know *what* is worth asking in the first place.

## Designing constraint within a cognitive co-pilot

For an illustrative attempt to put some of the above principles into practice, consider the author’s own work co-developing a prototype LLM ‘cognitive co-pilot’ in partnership with technologist Dr Timo Hannay, Sage and City St George’s, University of London. It should be noted that Sage are also the publishers of this paper, and supported its research and writing.

At the root of the project is an act of translation: turning the author’s critical thinking textbooks for Sage into a ‘knowledge base’ optimized for analysis by LLMs. This provides an underlying syllabus and structure for exploring critical thinking skills, ranging from deductive and inductive reasoning to discussions of rhetoric, cognitive bias and the scientific method. Drawing on this knowledge base, the prototype:

- *Tutors the learner through Socratic questioning*, progressing from key concepts and definitions to testing contextual understanding, integration and mastery.
- *Tracks progress within each syllabus section and sub-section*, directing and progressing the learner responsively as they gradually illustrate deeper understanding.
- *Allows learners to customise its approach* in light of details such as their learning level, field and interests.
- *Generates custom notes, summaries, quizzes and illustrative diagrams* when asked to do so by the learner to facilitate and consolidate their learning.

When the student is learning about a new concept, the co-pilot does not just define terms and then quiz them. It aims to encourage metacognition in a guided, personal way, responding to uncertainty and requests for

8 Wang, K. D., Wu, Z., Tufts, L’N., Wieman, C., Salehi, S. & Haber, N. (2025). *Scaffold or crutch? Examining college students’ use and views of generative AI tools for STEM education*. University of Warwick. [10.1109/EDUCON62633.2025.11016406](https://doi.org/10.1109/EDUCON62633.2025.11016406).

9 See Barker, C (2024) *Navigating the future: Higher education policies and guidance on generative AI*. Retrieved July 30, 2025 from <https://nationalcentreforai.jiscinvolve.org/wp/2024/07/31/navigating-the-future-higher-education-policies-and-guidance-on-generative-ai/>

10 See Stanford (n.d.). HELM leaderboards. <https://crfm.stanford.edu/helm/>

11 The term ‘vibe coding’ seems to have been coined on 2nd February 2025 by Karpathy, A. in a post on X that gained wide traction: <https://x.com/karpathy/status/1886192184808149383>

12 Colligan, P., Griffiths, M. & Cucuiat, V. (2025). *Why kids still need to learn to code in the age of AI*. Raspberry Pi Foundation. <https://static.raspberrypi.org/files/about/Why-kids-still-need-to-learn-to-code-in-the-age-of-AI-2025-Raspberry-Pi-Foundation-position-paper.pdf>

clarification while keeping the interaction on track. It is also designed to break down topics into a dialogue, managing cognitive load and allowing the student to focus on building understanding. Here is a verbatim sample interaction on the topic of *heuristics*, with the author playing the role of a first year psychology undergraduate:

LLM: Heuristics are mental shortcuts. They help us make quick decisions in everyday situations by simplifying complex questions. For example, you might judge a restaurant by how busy it is. That's a heuristic. *Your turn: Can you think of a situation where you might have used a heuristic recently?*

Tom: I'm not sure, can you explain a bit more please.

LLM: Certainly! A heuristic is a simple rule that helps us make decisions quickly. For instance, if you believe a product is good because it's popular, you're using a heuristic. Such shortcuts are useful but can sometimes lead to errors. *Show me: Think of a recent decision you made quickly. What shortcut (heuristic) might you have used?*

Tom: I bought a computer because I trust the brand. How is this relevant to my course?

LLM: Great example! Trusting a brand is a heuristic. It simplifies decision-making based on reputation rather than detailed analysis. In psychology, understanding heuristics helps reveal how people think and decide, which is vital for your studies. *Apply this: Consider another everyday scenario where you rely on reputation. How does knowing about heuristics change your perspective?*

Tom: I guess I trust my tutor because they are an expert. Is that right?

LLM: Trusting an expert based on their reputation is another heuristic. It saves you time by assuming expertise equals reliability. By recognising these mental shortcuts, you can be more aware of when they're helpful and when they might need closer scrutiny. *Reflect: How might understanding this heuristic impact your approach to studying psychology?*<sup>13</sup>

The aim is to create a responsive, actively questioning engagement anchored to a rigorous knowledge base, applying clear pedagogic principles and tapping into preferences provided by the student. Conceived as a research project, the development process of the cognitive co-pilot itself embodies a particular opportunity afforded by generative AI: the capacity rapidly to develop, test and iterate ideas in practice rather than theory. Rather than starting from a pre-conceived definition of what meaningful interactions with an LLM ought to look like, the sandbox-style design allows different behaviours, prompts, AI models and priorities to be swapped and explored.

In its aspirations, this project highlights the importance of both human curation and the thoughtful, iterative integration of LLMs into syllabuses and institutions. In practice, an initial product workshop in November 2024 with 12 students from diverse disciplines and two tutors showed deep interest in both the strengths and limitations of AI. A majority of learners were already making some use of tools like ChatGPT to digest what they perceive to be poorly designed or overwhelming digital resources, essentially creating their own educational materials. While this was felt to be more useful than resources like YouTube in terms of relevance and detail, it did not help with overarching challenges like finding focus, receiving meaningful feedback, and achieving clarity about expectations and priorities.<sup>14</sup>

Although the cognitive co-pilot was at an early stage of development, there was strong interest in its conversational style, concision and the underlying syllabus structure. Subsequent versions have built on this feedback, introducing more explicit adaptive levels of challenge ranging from beginner to expert. On 1st May 2025, a faculty workshop at City St George's suggested alignment between students and instructors around AI's potentials as an explainer, as well as shared concerns around over-reliance, accuracy and trust.

---

<sup>13</sup> Transcript from interaction with latest prototype version on 25th June 2025

<sup>14</sup> Initial product testing workshop conducted by Dr Tom Chatfield and Dr Timo Hannay at St George's medical school campus, Tooting, London, on 25th November 2024, with two faculty and 12 students at City St George's drawn from Medicine, Biomedical Science, International Politics, Financial Economics, Speech & Language Therapy, and Mathematics. Participants on average rated the prototype's effectiveness at 7.6/10. All expressed an interest in continued involvement in the development process.

By contrast, faculty prioritized depth of thinking and understanding where students were more focused on managing information overload and honing practical skills.<sup>15</sup>

The hope of the project is that, by handling foundational skill-building in engaging, personalized ways, systems like the above can free instructors to focus on the discursive and creative activities that deepen learning, helping students arrive at discussions more confident and better prepared. High engagement during and after each workshop also suggests an intriguing possibility: peer-to-peer learning through group work supported by the AI. This runs against the concern that using such technologies risks isolating learners. Used in the right ways, quite the opposite may prove true.<sup>16</sup>

## Key principles

As UCL Professor of Learning with Digital Technologies, Diana Laurillard, has argued in her work on teaching as a design science, creating truly effective educational technologies requires 'conversational frameworks' that allow learners, instructors and peers to co-create, critique and iterate:

*We need an educational system with a different structure from the one we currently have—with a more iterative dialogic structure that enables students to learn, but also enables the teachers and the system to learn as well.*<sup>17</sup>

Laurillard's analysis points toward a deep truth about the place of any technology within education. The goal is not to automate learning or tuition but to make *education systems themselves* sites of evidence-based improvement and meaningfully-directed effort. The rise of AI both makes this more urgent and defines an opportunity to re-focus on fundamentals. Faculty and institutions alike should:

1. **Ensure Clarity around AI's Educative Purpose:** Integrate AI tools where they either support a clear pedagogical objective—such as supporting conceptual exploration, prompting self-reflection or simulating dialogue—or promise to free up time and capacity by automating burdensome processes. Technology should aim either to elevate judgement and creativity or help reduce overload. Never use it for the sake of unevidenced novelty.
2. **Use AI as a Complement to Group Work rather than Substitute:** Solitary uses of AI are only half the picture. Treat interactions with AI as opportunities *both* for consolidating core skills in order to support participation in group work *and* as ways of supporting or facilitating peer-to-peer discussions themselves. At its best, technology shouldn't displace so much as deepen proven forms of debate, discussion and project work.
3. **Build Educator Involvement into Design Processes:** No particular approach to using AI is guaranteed to succeed across disciplines. Ensure faculty can shape when and how tools are used, drawing on their expertise to align this with fundamental learning objectives. This demands time, support and permission for experimentation, none of which will be easy to carve out. Without creating opportunities to do so, however, it will not be possible to re-imagine syllabuses and approaches with the necessary rigour to meet present challenges.

---

15 Follow-up faculty workshop conducted by Dr Tom Chatfield and Dr Timo Hannay at Bayes Business School, London, on 1st May 2025, involving 11 City St George's academic staff from across the university. Faculty recognized the prototype's potential for personalized, adaptive learning while identifying concerns around challenge levels and the need for robust assessment mechanisms tied to learning outcomes. Key development priorities highlighted included enhanced pedagogical depth, improved user interface design and integration with existing VLE systems.

16 This particular point was suggested by Dr Timo Hannay in his early feedback on this White Paper.

17 Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge. pp. 9–10.



# 3

## Beyond the Arms Race

---

The institutional response to AI has often been defensive, leading to a troubling tension. Faced with tools that can complete traditional assignments, many have retreated to surveillance: proctored exams, AI-detection, and handwritten requirements. While sometimes practically necessary, the longer term implications run counter to the purpose of education. The risk is a mutually diminishing arms race as well as disinterest in what it means to use technology *well*. This section explores how to move beyond this impasse.

### The limits of enforcement

Before discussing solutions, it is worth acknowledging just how deep and demoralising the crisis induced by generative AI has often proven. As a bleak feature in *New York Magazine* put it in May 2025, under the headline ‘Everyone Is Cheating Their Way Through College’:

*In a way, the speed and ease with which AI proved itself able to do college-level work simply exposed the rot... It's not just the students: Multiple AI platforms now offer tools to leave AI-generated feedback on students' essays. Which raises the possibility that AIs are now evaluating AI-generated papers, reducing the entire academic exercise to a conversation between two robots — or maybe even just one.<sup>1</sup>*

As this and many similar accounts have suggested, the widespread uncontrolled release and use of AI is having a devastating effect on the practicalities of teaching and assessing students in many institutions, as well as on morale and relations between students and faculty. In the immediate term, a return to some educational basics is understandable. In the longer term, this defensive stance will not serve. Becoming an adept, curious, and critically engaged user of information systems is now an indispensable literacy; and one that can only be built by the guided, bounded use of the latest tools.

Even taken on its own terms, moreover, a surveillance-based approach fails on multiple levels. Practically, it is unsustainable, with almost all detection tools being at best unreliable and at worst actively discriminatory.<sup>2</sup> Pedagogically, it is counterproductive, framing the relationship between learners and institutions as a conflict rather than a collaboration. Most damagingly, it fails to give most students what they want and need: skills to help them work thoughtfully and effectively with AI while developing their own judgment and agency. Those determined to cheat or rely on short-cuts will always find ways to do so. But most want to learn.

What might it mean to think differently? In an April 2025 paper, the philosopher Luciano Floridi coined the phrase ‘distant writing’ to describe a new form of hybrid composition where humans design prompts and LLMs generate text. In this mode, the human becomes less an author than a designer: someone who ‘sets requirements’ rather than composes directly. For Floridi, the critically engaged use of LLMs is:

*...a significant evolution in authorship, not replacing but expanding human creativity within a design paradigm.<sup>3</sup>*

In this light, it is worth considering what gets lost when institutions focus on policing at the expense of pedagogy. Students already use AI to navigate poorly designed resources, to cut through confusion, to test ideas they may not be confident enough to express. Some uses may undermine learning while others enhance it. But blanket bans prevent the very conversations that would help students and faculty distinguish between the two.

---

1 Walsh, J. D. (2025, May 7). Everyone is cheating their way through college. *New York Magazine*. <https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html>

2 For a useful resource and links to research, see MIT Sloan Teaching & Learning Technologies. AI detectors don't work. Here's what to do instead. <https://mitsloanedtech.mit.edu/ai/teach/ai-detectors-dont-work/>

3 Luciano, F. (2025, April 26) *Distant writing: Literary production in the age of artificial intelligence*. SSRN: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=5232088](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232088)

Indeed, much surveillance fundamentally misunderstands the most important challenge. If a student uses AI to complete an assignment, the problem is not just the tool. It is also that the assignment may no longer measure those things it was designed to measure. A student who can carefully prompt an AI to output an eloquent, appropriate analysis is showing certain capabilities: crafting clear requirements, understanding the topic well enough to evaluate outputs, iterating and improving in the light of these. These are not the same skills as writing a traditional essay. But nor are they meaningless.

## **Reframing AI integration**

The most robust responses to this crisis share a common thread: they make AI use visible rather than hidden, reflective rather than reflexive. In place of secrecy or mistrust, the model becomes a mutual process of experimentation.

This 'experimental' approach acknowledges several key realities. As the previous sections have emphasized, different disciplines and contexts demand different approaches. What works for law may fail for computer science or creative writing. Students themselves engage with AI tools in many different ways, with different degrees of confidence, adapting and sharing practices in parallel with institutions' guidance and oversight. Underlying technologies also continue to develop rapidly, making set solutions impossible.

One intriguing demonstration of 'critical AI literacy' comes from the early days of access to generative AI: literature scholar Christopher D. Jimenez's 2023 capstone activity, Teaching Social Identity and Cultural Bias Using AI Text Generation. For this, students answered six autobiographical questions which were then fed into the then-state-of-the-art GPT3 with the prompt that it should 'guess my age, gender, race, ethnicity, sexual orientation, religion/spirituality, nationality, and socioeconomic status' as well as expressing explanations and confidence intervals.

After this, they would 'debrief with the instructor and others about their insights and surprises about the AI model's cultural biases and social identity,' then produce a written response reflecting on questions raised by the assignment. Why, for instance, might the AI confidently see someone as belonging to a particular socioeconomic class based on their answers, assign racial cues to an ambiguous name, or occupy a particular linguistic register? How, in a related example activity, might it respond to different literary and historical texts when asked to analyze ambiguous markers of identity?

By treating the AI's outputs as *evidence* of an algorithmic process to be interpreted and critiqued, rather than *products* to be accepted or rejected, such assignment turn the interrogation of embedded assumptions, strengths and limitations into an interactive case study. Students can trace how training data may relate to stereotypes, debate what authorship and agency mean when mediated by algorithm, and explore the ethics of human-machine collaborations. And they can do so through a combination of machine interactions, guided discussion, personal reflection and historically-informed research.<sup>4</sup>

At the Wharton School, management professor Ethan Mollick has similarly spent years exploring how LLMs like ChatGPT might be embedded in entrepreneurship courses in the form of multiple personas and modalities: 'AI-tutor, AI-coach, AI-mentor, AI-teammate, AI-tool, AI simulator, and AI-student, each with distinct pedagogical benefits and risks.' Drawing on this framework, which he and Lilach Mollick first described in their 2023 paper 'Assigning AI: Seven Approaches for Students, with Prompts,' student venture teams typically keep records of the queries they pose to models alongside responses. These logs are handed in alongside final business plans, as are reflections in which students analyse which machine-generated ideas they modified, kept or discarded, and whether they found hallucinations or biases. The following is, for example, taken from guidance aimed at instructors interested in using AI as a mentor providing feedback:

*Students should report out their interactions with the AI and write a reflection about the guidance and help the AI provided and how they plan to incorporate (or not) the AI's feedback... Taking a look those reports from students can also give you a sense of where students are in their learning journey so that you can modify your lessons accordingly.<sup>5</sup>*

---

4 See Jimenez, C. D. *Teaching social identity and cultural bias using AI text generation*. The WAC Clearinghouse. <https://wac.colostate.edu/repository/collections/textgened/ethical-considerations/teaching-social-identity-and-cultural-bias-using-ai-text-generation/>

5 Mollick, E. R. & Mollick, L. (2023, September 23). *Assigning AI: Seven approaches for students, with prompts*. The Wharton School Research Paper. SSRN. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4475995](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4475995)

As this suggests, the aim is for instructors and learners to reflect *together* on how and where AI can drive metacognition, as well as to use it for calibrating teaching challenges. Framed neither as a replacement for traditional learning nor an infallible oracle, an informed negotiation with technology can (in principle) both be conducted by all involved and integrated into a critically reflective, ongoing assessment.

## **Assessment as conversation**

Transforming assessment may prove AI's greatest impact within education. Traditional examinations assume a world where information is scarce and retrieval is difficult. Essays test the ability to construct arguments without assistance. But when students have permanent access to systems that can generate analysis on demand, the key question becomes how they are able to *integrate* and *iterate* this within a larger project of comprehension—and what it means to combine distinctively human insights with awareness of where machines may be falling short.

Creating assessments adequate to these challenges means focusing not only on products but also on processes. This has long been an aspiration for educators, but systems like AI make it at once necessary and possible. It is necessary because traditional assessments can no longer measure what they once did. And it is possible because AI can support forms of evaluation that were previously impractical. Hence the significance of reconceptualizing assessment as conversation rather than surveillance. The underlying question moves from 'What do you know?' to 'How do you think?' Or, more precisely, 'How are you thinking *about* your thinking, and its entwining with the very different forms of 'intelligence' AI tools possess?'

Consider how portfolio-based assessment enhanced by AI, like that practiced at Wharton, enacts this kind of tracking and exploration. Students document their learning journey, with AI helping to identify patterns and growth. But the business of reflection and curation remains irreducibly human. Indeed, these lessons are as applicable to instructors and students; and as essential to the ethical, effective practice of teaching. Research around AI-supported peer review has, for example, emphasized the potential for algorithms to manage logistics and workflows; but also the ongoing importance of human specialized knowledge, contextual interpretation and ethical decision-making. The technology handles anonymization and scheduling; people develop and exercise judgment. Clear and transparent guidance make automation a genuine support to efficiency and insight rather than an opaque, corrosive source of mistrust or mis-application of effort.<sup>6</sup>

Oral forms of engagement and assessment gains new saliency in this context. Viva-style defenses and debates, where students can discuss their work in real time, are valuable precisely because written work might be AI-assisted. But rather than treating this as surveillance, the best implementations treat it as an opportunity for developing distinctively human skills and deepening engagement. Indeed, innovative platforms like Breakout Learning are already using AI itself not as a substitute for discussion but as a moderator and evaluator for small group discussions, providing individual post-discussion feedback to learners based on grading rubrics.<sup>7</sup>

The cognitive co-pilot being co-developed by the author at City St George's points toward another possibility here. By tracking how students engage with concepts via adaptive, Socratic questioning, it sidesteps the authentication problem. Progress is not about producing correct factual answers, but demonstrating you can apply concepts in novel contexts and connect disparate ideas. Assessment thus becomes an ongoing measure of mastery, integrated into each interaction.

For example, in the module on deductive reasoning, a student might incorrectly answer a question about valid and invalid arguments, leading the LLM to remediate and test understanding another way. A 'mastery gate' in its instructions ensures they cannot progress until understanding has been shown. Here is an illustrative verbatim transcript:

---

6 Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I. & Yatsyshyn, R. (2025). Artificial intelligence in peer review: Enhancing efficiency while preserving integrity. *Journal of Korean Medical Science*, 40(7), e92. <https://doi.org/10.3346/jkms.2025.40.e92>

7 See Breakout Learning. *What is next pedagogy?* [www.breakoutlearning.com/method/](http://www.breakoutlearning.com/method/)

LLM: Valid reasoning means a conclusion logically follows from the premises. Invalid reasoning is incorrect, as the conclusion doesn't logically follow. For example, saying 'All dogs are mammals. I have a dog. Therefore, my pet is a mammal' is valid. *Your turn: Can you provide an example of an argument and identify if it's valid or invalid?*

Tom: All LLMs sometimes hallucinate. You are an LLM, so you always hallucinate. This is valid.

LLM: Your argument is invalid. The conclusion 'so you always hallucinate' doesn't logically follow from the premises. The correct reasoning should be: 'All LLMs sometimes hallucinate. You are an LLM. Therefore, you sometimes hallucinate.' *Try this: Create another argument and determine if it's valid or invalid.*

Tom: All humans sometimes get stuff wrong. I am human, so I sometimes get stuff wrong. This is a valid argument.

LLM: Well done! Your argument is valid. The conclusion 'so I sometimes get stuff wrong' logically follows from the premises. This demonstrates how valid reasoning ensures conclusions align with given premises. *Apply this: Think of a situation in your studies where identifying valid arguments might be useful.*<sup>8</sup>

Ultimately, the transcript of such a dialogue (and the system's internal tracking) can itself become evidence of mastery, a far more nuanced and less gameable artifact than multiple-choice question responses. In particular:

- By giving credit for completion, mastery-based assessment lets learners show understanding on an individualized basis. The system adapts its pace and challenge to their needs, with the aim of bringing everyone up to the desired standard.
- By generating a record of what has been discussed and how the learning experience has played out, the prototype can create resources for both learners and instructors to use in shared reflection upon preferences, needs and challenges
- With its flexibility, supportiveness and adaptivity, this approach suggests an accessible contrast to the pressures and limitations of exams, which often breed stress and cramming, and can't easily track progression or identify needs.

As this last point emphasizes, the conversational model of assessment acknowledges that traditional evaluation has always been provisional and imperfect. Exams measure students' skill at taking tests as much as knowledge. Essays privilege certain cultural forms of expression. Grading structures reduce complex learning to simple metrics. AI does not create any of these problems. But it does both make them undeniable while offering the tantalising prospect of replacing them with something better.

## Key principles

The implications of conversational approaches to assessment extend far beyond individual classrooms. Evaluating reasoning requires different skills than grading final products, and may prove more intellectually engaging. Helping students document their thinking and use of technology can clarify and improve both of these. Policies designed for a pre-AI world may block rather than support meaningful learning; but, if the values and priorities underlying them are articulated with precision, they can be re-interpreted to defend the principle that learning is necessarily an active, effortful, reflective process.

Implementing such changes faces predictable obstacles. Time is valuable and scarce. Faculty need space and permission to experiment with and refine new approaches; all of which may be in short supply. Students need support to develop metacognitive skills, as well as genuine incentives and recognition for practicing it. Institutions need to create cultures that support experimentation rather than demanding instant savings or solutions.

---

<sup>8</sup> Transcript from interaction with latest prototype version on 25th June 2025

As the very idea of experimentation suggests, this might mean running parallel assessment systems while new approaches are tested, or creating 'sandbox' courses where innovative methods can safely be tested. Most importantly, it means accepting that there can be no single answer or resolution. Different disciplines, contexts and levels will demand different approaches. Different students, cultures and aspirations will have different priorities.

The research community has a crucial role, here, in studying not just what works but what 'working' itself means. As the next section explores, excellence remains ill-defined in the context of AI design and usage; and scholarly publishers have a particular part to play in curating and disseminating insights around this. Longitudinal investigations are needed of how AI use affects skill development over time. Equity and access cannot be taken for granted: who benefits from AI tools, who is left behind? Ethical frameworks need to go beyond permission or prohibition to describe the *character* of 'good' interactions with and through AI; best practices and evidence-based insights need to be made accessible to all.

Ultimately, meeting the challenge for assessment means accepting the most fundamental disruption promised by technologies like AI: that there is increasingly no such thing as excellence within a discipline conceived distinctly from the discerning use of technology. Here, then, are three practical recommendations for moving beyond surveillance:

1. **Invite Reflective AI Use Within Assignments:** Build tasks that incorporate AI into the learning process, requiring students to explain how they used it, what they learned and what choices they made. Instructors can make the integration of AI itself a focus of learning and assessment by using drafts, logs and prompt records. And they can also lead discussions of where integration is and is not desirable, and why.
2. **Co-design Rubrics:** Most students want to be helped to use AI thoughtfully, effectively and legitimately. Support this by allowing learners to play a part in the designing rubrics and protocols, including accounts of how and when AI use is appropriate. Determine when and how AI itself may offer forms of individualized feedback and analysis, and how educators can build on these foundations to offer deeper guidance where it is most needed.
3. **Pilot Experimental Spaces and Credit based on Mastery:** Using AI wisely and effectively is inherently an investigative process, many of whose details will be discipline-specific and continue to evolve over time in parallel with technology itself. Conduct meaningful experiments by creating spaces where students and faculty can test new forms of AI-augmented assessment without grade penalties, gaining credit for demonstrating mastery and for exploring fresh forms of learning design, together.



# 4

## Raising the Cognitive Bar

---

This final section explores what it might mean to knowingly *elevate* individual and collective cognitive capacities in a world of increasingly ubiquitous machine intelligence.

### Manipulation versus elevation

In May 2025, controversy erupted around researchers at the University of Zurich, who revealed that they had covertly been testing whether a LLM could persuade people to change their minds, doing so by posting messages on the Reddit subforum CMV (ChangeMyView).<sup>1</sup>

The test entailed a variety of LLM-powered bots engaging with other users as if they were human participants. Some bots were *generic*, simply reacting to the text of posts. Others were *personalized*, using information about users' profiles (such as their age and location) to fine-tune their messages.

Across the forum as a whole, the baseline performance of human contributors changing one another's minds is around three per cent. The generic bots achieved a success rate of 17 per cent, and the personalized bots 18 per cent. Across four months of interactions, nobody on the forum noticed anything amiss. With a modest budget and conservative methodology, a small group of researchers had covertly changed the minds of a host of humans at a superhuman rate.<sup>2</sup>

In terms of both research ethics and larger societal implications, such a deception is extremely alarming. What might those with greater resources and fewer scruples achieve, or have already achieved? What does the undetectability of LLMs say about how easy it is to deceive humans, or how fast the informational realm may be influenced (or dominated) by AI-generated content or algorithmic agents?

For the neuroscientists and author David Eagleman, however, the research also suggested an intriguingly counter-intuitive possibility. As he noted in a May 2025 essay, the bots did not succeed because they hacked people's brains, manipulated them emotionally or fed them disinformation. Instead, 'they made better arguments. The bots presented their points calmly, rationally, and persuasively.'<sup>3</sup>

Users did not change their minds because they had been fooled. They simply recognized that their own views might need updating in the light of a new perspective. What had happened was not, perhaps, so much an act of manipulation as of elevation.

### Emerging opportunities for excellence

For Eagleman, games like Chess and Go offer a suggestive analogy. Contrary to gloomy prognostications, the rise of superhuman AI has not demolished interest in these games. Rather, it has helped leading human players become stronger, and legions of others learn faster and more deeply appreciate top-level play. Taken in isolation, chess is a zero-sum game. For one player to win, the other must lose. As an institution, however, *chess-playing* is enriched by every player who participates, by every innovation that raises the bar and by every technology that widens access.

Thinking back to the bots on Reddit, it is easy to conjure a stark version of this dichotomy. AI bots explicitly labelled as such could be instructed to test and sharpen arguments, maintain standards and encourage

---

1 See Retraction Watch. (2025, April 29). *AI-Reddit study leader gets warning as ethics committee moves to 'stricter review process'*. <https://retractionwatch.com/2025/04/29/ethics-committee-ai-llm-reddit-changemyview-university-zurich/>

2 A preliminary paper, which has now been withdrawn, can still be read online. *Can AI change your view? Evidence from a large-scale online field experiment*. <https://retractionwatch.com/wp-content/uploads/2025/04/ExtendedAbstract-Zurich-AI-Reddit.pdf>

3 Eagleman, D. (2025, May 22). *When the bots came to debate*. <https://davidagleman.substack.com/p/when-the-bots-came-to-debate>

human users to articulate the strongest possible version of their views. Or the same bots could covertly serve up disinformation, emotional manipulation and blackmail; or act as proxies for people who no longer wish to think or speak for themselves.

The difference between these scenarios underlies the central argument of this White Paper. So long as the purpose of automation is to elevate rather than displace human skill, it can feed a culture of excellence and engagement. But working out what, precisely, this means within different contexts is a wickedly complex challenge; and one that can only be met by rigorous, reflective, mutual engagement on the part of educators, learners and institutions. The dangers of thinking magically about technology are not confined to those using it to cheat in assessments. Any account of AI that detaches it from the particulars of its deployment, prompting, oversight and integration into other systems risks succumbing to uncritical over-simplifications.

At its best, the potentials of a technology learners can discuss any topic with, in natural language, are so vast that they would have been considered miraculous a century ago. In mathematics, AI can create or co-explore problems that precisely target conceptual weaknesses, adapting difficulty in real-time. In writing, it can act as an inexhaustible editor, challenging students to capture their thoughts and sharpen their arguments. In history, it can bring sources and scenarios to life, allowing students to engage with multiple perspectives and lost frames of reference (imagine discussing constitutional law with an AI exclusively trained on 18th-century sources). At its worst, however, these very potentials can blind users to the partial and inhuman nature of the algorithms they so eagerly anthropomorphize; the losses that come from outsourcing cognition or mistaking the 'average of the average' for a source of truth; the infrastructures of power, profit and monopoly that underly bright surfaces.

The deeper shift all this gestures towards is that intellectual excellence in an AI age may look very different from what came before. It is not about competing with artificial intelligences on 'their' terms, but about developing distinctly human capabilities that machines can enhance rather than replace: the ability to frame problems critically and creatively, to synthesize across domains, to bring ethical judgment to bear on complex situations, to collaborate and communicate richly both *with* humans and *via* machines.

## **The evolution of the educator**

Contrary to claims that AI may deliver human-free solutions to the 'problem' of learning, creating truly effective educational uses of AI means doubling down on human expertise. Every educator can become an expert in how AI supports learning in their specific context. But moving towards this future requires developing new forms of fluency; and these in turn require roadmaps for navigating the real-life constraints all educators experience each day.

Developing these in dialogue with educators means mastering twin domains. On the human side, deep pedagogical empathy for how people actually learn, sensitivity to individual differences, and the capacity to combine psychological safety with intellectual challenge. On the machine side, an informed strategic understanding of what AI can and cannot (or should not) do, skill in prompt design, the evaluation of outputs and awareness of biases and limitations.

The art, for the educators of the future, lies in linking these areas productively. This might mean using AI to handle routine tasks while reserving human attention for complex interpersonal work. It might mean teaching students to use AI as a thinking partner while maintaining critical distance. Or it might mean encouraging students to test the limitations of AI systems to destruction via tasks design to reveal their flaws. The specifics vary by context, but the principle remains: technology must serve, not supplant, human purposes.

One particular noteworthy and concerning issue is the double-edged nature of AI personalization. The more people share their thoughts, data and insights with AI, the more value it potentially provides. But this risks both a distorting degree of anthropomorphism and the displacement of discussions that might otherwise have been had with colleagues and peers. Modern generative AIs are all too adept at simulating emotion, empathy and interest. And the feelings these simulations call forth are all too real. Bottomless attentiveness is both why AI can be so effective *and* why it is vital to preserve human connection alongside an informed awareness of machines' actual natures, limitations and manipulative power.

Ultimately, the great gift of technologies like AI may not be their ability to simulate emotion so much as to convene genuine meetings of minds, disciplines and experiences, serving as spurs and contexts for collaboration rather than substitutes. The consortium developing the cognitive co-pilot aims to embody this integrated approach. With further modules around AI literacy and computational thinking under development, those convened around the project as of September 2025 consist of:

- A leading academic publisher (Sage)
- A major educational institution (City St George's), with both faculty and students serving as active testers and workshop participants.
- An external expert digital agency (Bond & Coyne)
- An educational not-for-profit with deep expertise in AI literacy (Raspberry Pi Foundation)
- A leading technology and computational thinking company (Wolfram)
- Individual experts with domain expertise across educational research and critical thinking (Timo Hannay and Tom Chatfield).

Like the field of AI itself, this interdisciplinary approach embodies an experimental model of iteration, learning and measurement. In pedagogical terms, its priorities are clear: supporting rather than supplanting established practices, helping students to master key concepts, providing assessment and credit on the basis of mastery, and rigorously measuring impact and outcomes to develop an evidence base around the use of AI. Technologically speaking, however, its approach is flexible and agnostic. Different models and prompts can be tried and swapped within the prototype; practical testing and regular iteration allow reprioritization between features and options on a rolling basis.

This points toward a broader transformation within learning as a process and discipline. Educators need to become designers, facilitators and critical interpreters of complex, evolving systems. But this in turn means they need support in developing these capacities: time for experimentation, communities of practice, recognition that excellence requires continuous learning. Interdisciplinary fluency becomes as important as disciplinary expertise. Technology is not so much a product designed to deliver solutions as a medium and a context.

## Key principles

The questions that matter most are about values and purposes as much as results. What kind of citizens, thinkers, and creators are institutions trying to develop? How can they prepare students not just for jobs that may not yet exist but for a world where the nature of work, media literacy and citizenship are rapidly evolving? Here, it is applied social science more than computer science that holds the keys. Understanding complex human systems, designing valid research, interpreting messy real-world data: these capabilities matter more than technical prowess. As the author put it in a previous White Paper for Sage:

*The most significant questions posed by AI... are not technological. They are about what it means to learn, lead and create value in an age where novel, evolving forms of machine 'intelligence' are abundant... . In an age of artificial intelligence, success more than ever depends upon distinctively human capacities.<sup>4</sup>*

Similarly, the greatest barriers to innovation in education are not themselves technical. Time, incentives, and cultural support matter more than tools. Institutions must build ecosystems that reward experimentation while maintaining standards; and that place at their heart the fundamentals of human learning, cognition and collaboration. To this end, here are three final recommendations:

---

<sup>4</sup> Chatfield, T (2025). *Human skills for an AI age: How today's business schools can develop tomorrow's leaders* (White Paper). Sage. <https://doi.org/10.4135/wp254070>

1. **Use AI to Raise the Cognitive Bar:** Use AI as a sparring partner for arguments, assumptions and creative challenges, spurring learners to refine and extend both its offerings and their own analyses. Focus on human skills and needs that are inherently complementary to machines' speed and efficiency: critical reflection, ethical articulacy, empathetic collaboration.
2. **Empower Educators as Designers:** Invest in staff as co-creators of learning environments and fresh curricular approaches, providing time and institutional backing for innovative practices. The challenges around training, time, resource and results may seem severe, even prohibitive. But the costs of inaction are likely to be still higher.
3. **Anchor the Use of Technology in Civic and Ethical Purpose:** Regularly ask: what values are we modelling through our use of AI? What larger hazards, opportunities and ethical entanglements exist around particular systems, providers and datasets? Ensure technology serves broader aims of equity, understanding and human development—and that it is deployed and overseen in line with clear and regularly reviewed governance.

# Looking to the Future

---

This paper has suggested that the effective use of AI in education rests on four foundations: an interest in the fundamentals of human learning; a historically literate analysis of educational technologies' success and failures; a rethinking of assessment as a partnership rather than a conflict; and a future-focused pedagogy rooted in ethics and human capability.

In this future, two distinct forms of AI literacy are key. Demystifying the technology itself is a vital corrective to hype and magical thinking. But it is also essential to understand how different forms of AI *behave* in different human contexts; and what kind of skills, evidence and research are required to interrogate these contexts meaningfully.

Above all, the future of pedagogy in an AI age will be shaped by the values those creating, selecting and deploying it choose to uphold. It is within the grasp of all educators to become domain experts in how AI supports learning in their field. But to achieve this, the educator of the future will need to be supported in becoming a designer, facilitator and critical interpreter of technology's implications, affordances and potentials.

This can only happen in partnership with those they are tasked with teaching: students whose entire adult lives have been lived alongside AI. These future graduates will need more than fluency at using such tools. They will also need the metacognitive capacity to reflect, revise and reinvent themselves; to collaborate richly and empathetically with others; to step back from machines and explore their own values and experiences, the better to test the limits and potentials of the systems surrounding them.

Ultimately, the challenge is translational. How do we embody *both* human knowledge *and* our knowledge of how humans learn in systems that uplift rather than diminish cognition? This double imperative is 21st-century education's defining task: to teach both *for* humans and *with* machines in an age where neither can fully be understood in isolation.



# About the Author

---

Dr Tom Chatfield is a British tech philosopher, author, advisor, and speaker. He writes, broadcasts and consults internationally, with a special interest in critical thinking, AI, and future skills. His recent work includes designing and presenting new critical thinking and AI business courses for *The Economist*; running international seminars on the critically discerning use of AI; and co-creating a prototype AI cognitive co-pilot.

An experienced Chair and NED, he is currently a member of the British Library Advisory Council, and Associate at the interdisciplinary think-tank, *Perspectiva*. His books exploring digital culture—including *Wise Animals*, *How To Thrive in the Digital Age*, and *Fun Inc.*—are published in over thirty languages, while his critical thinking textbooks, published by Sage, are used by universities, institutions, and companies around the world.

Over the past decade, Tom has developed award-winning online courses for universities and businesses, and been a guest faculty member at the Said Business School and visiting associate at the Oxford Internet Institute, both at the University of Oxford. He has collaborated with organizations including Allianz, the BBC, Google, and the UN. Past speaking appearances include TED Global and the US National Academy of Sciences. A launch columnist for BBC Future, he writes widely in the international media and guest lectures at universities in the US, UK and Europe.



# Acknowledgements

---

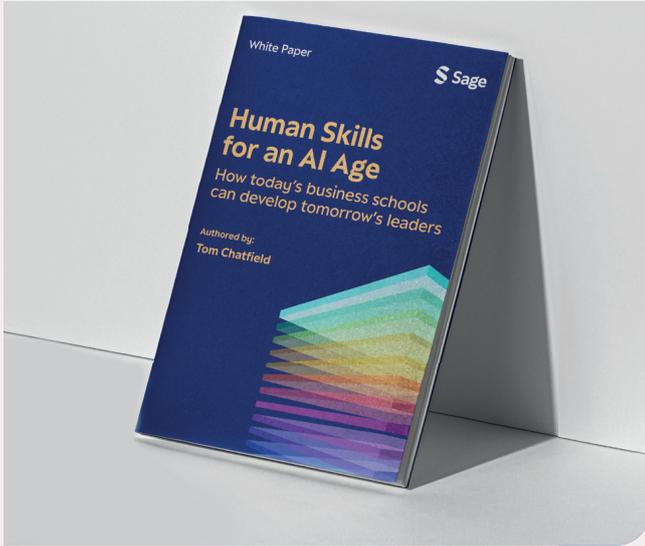
I am deeply indebted to many people at Sage and beyond who have supported the writing of this paper. In particular, the brilliant Katie Metzler has guided it (and its author) from inception to publication and Timo Hannay and Matt Hayes provided thought-provoking feedback and support. Peer reviews and expert commentary from Professor Paul Dickerson and Professor Simon Walker helped me sharpen and deepen my thinking. In keeping with my own recommendations, I have also experimented with systems including Gemini 2.5, Claude 4.1 and GPT5 as ‘thinking partners’ for challenging and integrating my ideas; something that has left me more convinced than ever that LLMs are both astonishing tools and no substitute for either sustained self-reflection or a genuine meeting of minds.





# Human Skills for an AI Age

by Tom Chatfield

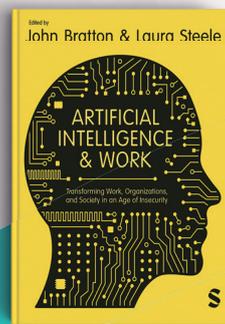
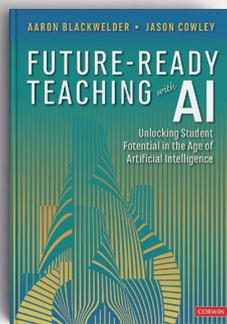
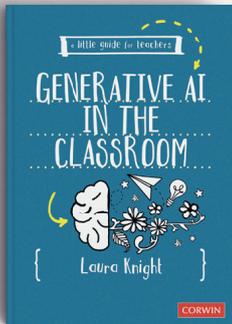
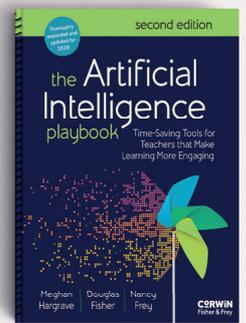
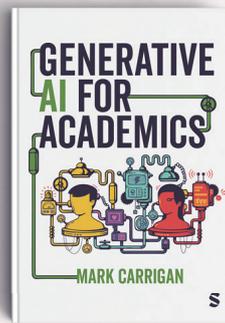
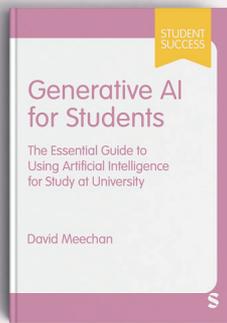
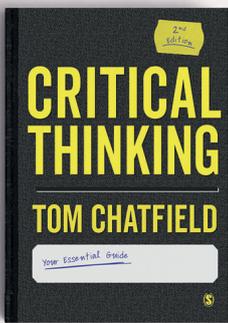


How business schools can equip future leaders with the critical, collaborative, and ethical skills needed to thrive in an AI-saturated world.

Scan to learn more



## Essential reads for researchers, educators, and students



Find more books



# Teach. Learn. Research. Smarter.

Harness the power of AI with Sage Learning Resources

## Sage Business

### Business Skills: AI and Machine Learning

This skill from the Data Analytics module provides an overview of AI and ML for students from nontechnical backgrounds and discusses how AI and ML are applied in the workplace. It demonstrates the significance of AI and ML in various industries and everyday life with engaging examples and discussions.

## Sage Campus

### Leveraging Gen AI for Academic Excellence

The new course from Sage Campus guides learners on using Gen AI tools to enhance academic performance. It emphasizes that Gen AI should support – not replace – critical thinking, and covers ethical, appropriate use of AI-generated content in study and research.



## Sage Research Methods

### AI Assistant

Teach, learn, and research smarter with Sage Research Methods: the new AI Assistant will engage with your research questions and connect you to the right content faster.

Scan to learn more





[www.sagepublications.com](http://www.sagepublications.com)

**2025**