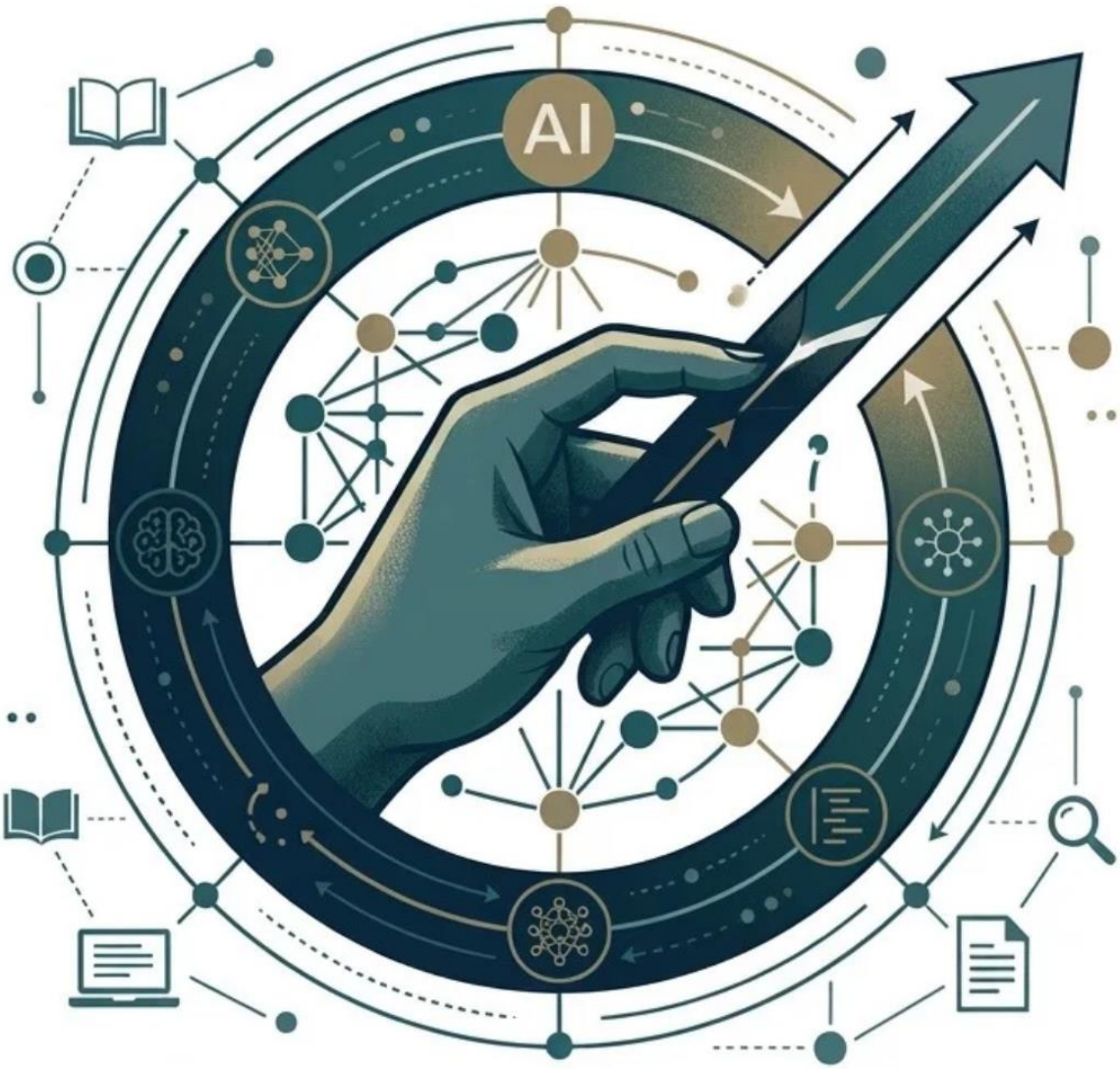
Eleanora G. Bolsi and Alex Grech

Breaking the Research Loop

The ongoing evolution of AI Literacy, by another name



The 3CL Foundation

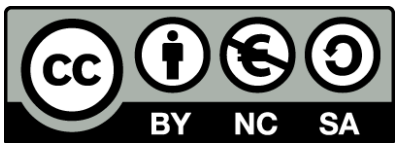
About this Document

This first draft of this report was in Italian. ChatGPT GPT-6 and Claude Opus 5.5 were used to support translation, editing, and formatting. Claude was also used to draft the executive summary, which the authors revised. The concept, analysis, discussion, and findings are the responsibility of the authors.

We are particularly grateful for the contribution of Godfrey Baldacchino in reviewing this document.

Published October 2026 by The 3CL Foundation, Malta.

ISBN 978-9918-0-2028-7



This publication is licensed under the Creative Commons Attribution NonCommercial ShareAlike 4.0 Unported (CC BY-NC-SA EU 4.0)

Credits Cover Illustration: Google Gemini.

Citation: Bolsi, E. G., & Grech, A. (2026). *Breaking the Research Loop*. *The ongoing evolution of AI Literacy, by another name*. The 3CL Foundation. Downloadable from www.3CL.org/resources

Contents

About this Document 2
Executive Summary 5
Introduction 7
From diagnosis to response 9
1. The Research Loop: scope and limitations 11
1.1. Before AI: mediated access and the pressure to be fast 11
1.2. Old vulnerabilities: the loop did not start with AI 12
1.3. What generative AI changes 12
1.4. The mechanism: four stages 13
1.5. What the Research Loop is not 14
1.6. Two stakes: the person and the system 15
1.7. What is distinctive about the Research Loop 15
2. From source displacement to recursive knowledge 18
2.1. Access is not availability 18
2.2. From mediation to source displacement 19
2.3. From substitution to reproduction: the performance of knowledge 20
2.4. Where help ends and substitution begins 21
3. Cognitive offloading and the erosion of critical thinking 23
3.1. From recursive knowledge to argumentative delegation 23
3.2. Cognitive offloading, before and with AI 23
3.3. Task performance is not the same as competence 24
3.4. The measurement blind spot 25
4. Consequences of the Research Loop for learners, researchers, and citizens 28
4.1. Learners: learning, assessment, and epistemic dependency 28
4.2. Researchers: productivity, integrity, and epistemic value 30
4.3. Citizens: public reasoning in an AI-mediated world 31
4.4. The three domains together 32
5. Adaptive human agency and non-delegable judgement 34
5.1. From calculators to autonomous agents 34
5.2. Human–AI co-agency: shared work, single-sided responsibility 34
5.3. Why static task lists fail 35
5.4. When inquiry starts bending to the machine 35
5.5. The criterion: epistemic value and non-delegable judgement 36
6. AI literacy as a capacity for judgement 38
6.1. From operating the tool to governing its use 38
6.2. From competence to judgement: performance and epistemic value 38

6.3. From prompt literacy to judgement literacy....39
6.4. Safeguards and a working metacognitive cycle....40
7. Breaking the Research Loop: Principles for Intervention....44
7.1. Planning: what am I trying to achieve, and is AI right for this part of the task?....44
7.2. Monitoring: What have I delegated, and what do I still understand?....45
7.3. Evaluating: How reliable is the output, and how well judged is my reliance on it?....46
7.4. Regulating: should I verify, change strategy, reclaim the task or stop using AI?....47
7.5. From responsibility to educational intervention....48
Conclusion....51
Preserving the Conditions of Critical Judgement....51
Glossary....53
References....61

Executive Summary

A student writes an essay from AI summaries of papers she has not read. The essay is coherent, accurate, and properly referenced, and it receives a good mark. Asked why she preferred one source to another, she cannot explain her choice. Nothing in the work is false. What is missing is her ability to explain and defend the reasoning behind it.

This report calls the wider pattern behind such cases the **Research Loop**. A source is first encountered through a representation, such as an abstract, a summary, or an AI synthesis. The representation then begins to substitute for the work of examining and checking the source. It is incorporated into a new text, which can become material for the subsequent learning, writing, and research. The process becomes recursive when that new work is used reused without renewed engagement with the underlying evidence. Each step is ordinary, and none requires error, fraud, or hallucination. Nor is the pattern new: citation chains and second-hand quotation produced the same drift long before generative AI. What AI changes is the cost and speed of the process. Retrieval, synthesis, interpretation, and drafting now be combined in a single inexpensive workflow, making it easier to turn already-mediated material into new output and feed it back into knowledge production.

What the evidence shows

The Research Loop works at two levels. For the individual, delegating the very work a task is meant to exercise can remove opportunities to practise reading, weighing evidence and constructing an argument. For the system, output can grow faster than understanding. The evidence reviewed in Part I raises related concerns across learning, research and public information. Grades at a large research university rose after 2022, particularly in courses relying on unsupervised written work, while becoming a much less reliable indicator of what students could do independently, and unaided. A Pew Research Center study found that, when AI-generated summaries appeared in search results, the proportion of users clicking on traditional search links fell from 15% to 8%. Separately, a preregistered study involving 1,237 participants found that people working with a language model believed they completed tasks faster, although their actual completion times were virtually unchanged.

These findings are from different settings and do not, individually, demonstrate the full Research Loop. Taken together, however, they highlight a common problem: improvements in perceived efficiency, output or assessed performance do not necessarily indicate stronger understanding or more valuable knowledge production.

Performance is not epistemic value

The report distinguishes between *performance value*, meaning the speed, volume, and polish of what is produced; and *epistemic value*, meaning what the work adds to understanding and to a person's capacity for further inquiry. Two researchers can reach the same correct conclusion. One can retrace the path from evidence to claim and explain why the alternatives were rejected; the other accepted an AI synthesis without examining it. Their outputs are indistinguishable, but only one of them can defend the result. A gain in performance raises epistemic value only when the time it saves is spent on harder interpretative work. Most current evaluation of AI tools looks at performance alone: in an exploratory scan of 20 recent studies, every one measured assisted output, and only six tested what participants could still do without the tool.

Where judgement has to stay human

The report does not argue for limiting how much work AI does. Fixed lists that assign some tasks to people and others to software fail, because the line between routine processing and evaluative judgement moves with the context. Responsibility is located instead at judgement-bearing decision points: the choices that settle the standing of a claim, such as which sources

to trust, whether the evidence supports a causal claim or only a correlation, which interpretation to adopt, and whether the conclusion follows. AI can prepare and inform these choices, but the person must keep the understanding and the authority to adopt, revise, or reject them. Heavy automation remains sound when those decisions are understood and checked, and light assistance becomes substitution when they are not. A name on a paper establishes formal authorship. Being able to defend its claims under challenge is epistemic authorship, and AI-assisted work can leave the first intact while hollowing out the second.

Four questions, and Read, Argue, Defend

On this account, AI literacy is less a stock of skills than a habit of judgement exercised while the work is under way. Building on the 3CL framework of critical digital literacies and on DigComp 3.0, the report organises that habit around four questions. What am I trying to achieve, and is AI right for this part of the task? What have I delegated, and what do I still understand? How reliable is the output, and how well judged is my reliance on it? Should I verify, change strategy, reclaim the task, or stop using AI? Asking them at the points where a claim's standing is decided keeps delegation open to revision.

Read, Argue, Defend is the educational translation of this approach. Read means returning from summaries to the sources on which consequential claims depend. Argue means building a line of reasoning of one's own, so that an AI-generated argument becomes one contribution to the problem rather than the only frame through which it is seen. Defend means explaining the choices behind a claim, answering objections, and revising a position when the objections hold. AI remains available throughout. What changes is that these three activities stay visible, and can therefore be taught and assessed.

What follows for practice and policy

Educators need assessment that makes understanding visible alongside the quality of an AI-assisted submission, through oral defence, source trails, and the comparison of arguments. Institutions need to give people the time, access, and authority to exercise the responsibility formally assigned to them, and to ask whether incentives for speed and volume are steering research towards what AI can most readily produce. Policymakers should judge AI literacy initiatives by their effect on verification, judgement, and the ability to intervene, and not only by rates of adoption and compliance.

The report offers a conceptual diagnosis supported by a critical review of the literature. It does not demonstrate the full recursive mechanism, and its extension beyond academia remains a proposition to be tested; Read, Argue, Defend provides a first setting in which to test it. The aim throughout is to expand what people can accomplish with AI without narrowing what they remain able to understand, judge, and defend.

Introduction

From digital natives to AI-mediated learners

Three people, one afternoon. A student uploads three academic papers to a chatbot and then drafts an essay from the summaries. A journalist, pressed for time, asks AI to check the veracity of speaker quotes from a press conference and then supplements these with content from a summary of a report she has not read. A third person types a question into a fresh chat, prompts a confident AI answer and forwards it to a friend. On their own, each of these actions appears mundane. Yet each user has enabled a 'stand-in' to take the place of the source. Multiply that decision across millions of afternoons and you have the subject of this report.

We make sense of the world by knowing things. We look for information, weigh evidence, form judgements, write these down and share them. More and more, that work runs through digital tools and AI. These can widen access, speed up inquiry and let a person become more productive. They can also change what we accept as 'knowing.' A generated answer, a summary or a chain of citations can start to stand in for the source, the evidence or the reasoning that actually needs checking. When those stand-ins are fed back in to make the next new claim, essay, paper or interpretation: the process begins to turn, and feed, on itself.

This report calls that pattern **the Research Loop**.

The Research Loop is what happens when the convenient AI answer stops being a way into the source and becomes a replacement for it, over and over, until more is produced and less is understood.

The lens of this report is epistemic. It asks how knowledge gets formed, justified, and validated. Through that lens, the Loop has a shape. A source is mediated by something that represents it. The representation is then treated as if *it* were the source. The result is reproduced - and in turn, that reproduction becomes the raw material for the next round of reading, learning, reasoning and research.

Academic research is the clearest case, since in academia, mediated work enters a formal record and becomes the ground others build on. But the problem is wider. Learners, citizens and knowledge brokers such as journalists increasingly work in AI-mediated settings, where claims are retrieved, summarised, and passed on by machine. Academia is treated here as an intensified version of a problem everyone shares: how to keep critical judgement alive in learning, everyday inquiry and public reasoning.

These individual shifts are subtle and rarely intentional. What this report is about is what they add up to.

Speed is not the whole story: from digital familiarity to AI literacy

We begin with the expectation to perform. We live inside a performance culture, in which individuals are encouraged or expected to continually perform, measure and display aspects of themselves in response to social, institutional and technological prompts and expectations. Performance culture rewards visible, measurable, fast output more than the slower work of understanding, reflection and judgement. In research, the pressures are obvious: productivity, visibility, funding and speed all bear down on how study and inquiry are done (Hicks et al., 2015; van Dalen, 2021).

But research cannot be reduced to output you can count. It also rests on knowledge of a field, interpretive experience, and the capacity for critical thinking, judgement, analysis and argument. These capacities grow slowly and are hard to measure. When counting becomes the point, research can be treated as a stack of separate operations that are easy to tally and easy to hand off, leaving less room for the tacit judgement by which a researcher notices a problem, weighs the relevance of evidence and reaches a defensible conclusion.

AI can help with all of this. It can support or speed up learning, inquiry and lighten the load of research. The trouble starts when the gain in efficiency is not matched by the human understanding to go with it: the ability to understand what was handed over, check the result and decide whether it is good enough. At that point, AI does not resolve the tension between productivity and quality: it sharpens it.

This is why AI literacy cannot mean familiarity with a fixed set of tools. The tools keep changing, and so does the way we work with them. The old split between digital natives, who grew up with technology, and digital immigrants, who came to it later, does not get us far (Grech, 2025; Kirschner & De Bruyckere, 2017). Growing up around technology does not, by itself, give you the capacity to use it well. Digital and AI literacy are better understood as capacities that have to be built – at once cognitive, technical and social, since using a tool responsibly depends less on what it can do than on whether the person can judge when and how to use it, weigh what it returns and verify what needs checking (Yan et al., 2024).

The report takes the 3CL framework as its point of departure and draws on DigComp 3.0, the European Union's framework for digital competences, whose latest version threads AI competence through the whole framework rather than bolting it on at the end (Cosgrove & Cachia, 2025). So, the central question is not whether AI is good or bad, or which tools to use. It is how AI may support learning, inquiry and research without replacing the practices that build curiosity, critical understanding and judgement in the first place.

Knowing our way around the tools is not the same as being able to determine their value to us.

From responsibility to competence

Responsible AI use is usually managed through principles, guidelines and institutional policies, which place the final responsibility on the human researcher. But naming someone responsible is not the same as making sure they can actually exercise that responsibility.

Recent empirical research (Zyphur, 2026) documents this gap. Tracking 38 top-tier doctoral universities, only six had policies that went beyond research integrity to address AI literacy and sound research practice. Policies tend to be clearer about duties, disclosure, authorship and acceptable use than about the capacities a researcher needs to decide what to delegate, verify sources, judge AI output and defend conclusions on their own. Being the name on a report is not the same as being able to answer for it when someone pushes back.

The Research Loop builds on this competency-based approach but asks a different question. Zyphur (2026) asks what researchers need to know and do to use AI responsibly. The Loop asks what happens to those capacities when AI-mediated material is reused as the basis for further knowledge production. It examines how source mediation, delegation and reproduction can become mutually reinforcing, as human verification and judgement become less substantive. The two approaches are complementary: one identifies the capacities that responsible practice requires; the other examines the conditions under which their inadequate exercise can undermine both individual judgement and the shared knowledge on which further inquiry depends.

Formal and epistemic authorship

This report distinguishes between two forms of authorship. Formal authorship is about whose name is on the work and who is officially responsible for it. Epistemic authorship is about whether that person can understand, reconstruct, justify, and defend the choices of interpretation, evidence, and arguments that hold the work up. AI-assisted work can leave the first intact while quietly hollowing out the second.

Put concretely: your name is on the paper. A sharp reviewer challenges your third claim. Can you explain and defend why it holds, or can you defer to the system that produced it? Formal authorship says the work is yours. Epistemic authorship asks whether the thinking is.

None of this argues for keeping AI on a short leash, or that it should play only a small, supporting part. AI can be woven deeply into learning, inquiry and research without undermining human responsibility. The real test is whether it replaces the interpretation, evaluation and accountable judgement the person still needs in order to understand and oversee what is done in their name. AI has already entered reading, research, public reasoning, and the building of arguments (Grech, 2025).

Two levels: the person and the system

The Research Loop works at two connected levels. At the individual level, delegation of work to the machine can reduce or redistribute the mental work that learners, citizens, and researchers would otherwise put in. Handing work off is not a problem in itself, and can often improve results (Risko & Gilbert, 2016). It becomes part of the Loop when what gets handed off is exactly what the person needs to practise: weighing evidence, reading and interpreting sources, building and defending an argument. Delegate that often enough and there remain fewer chances to practise the very capacities critical judgement depends on.

At the systemic level, the question is what becomes of the outputs. Claims, summaries and publications built from already-mediated material circulate further, get cited and synthesised, and become the input for the next round of learning, reasoning and research. The same readings are reproduced, and the most visible sources and familiar patterns gain still more visibility. AI accelerates all of this by making mediated material cheap to produce and reshape.

The two levels feed one another. Less engagement by each person can sit alongside more output overall. One person leaning on a summary is a private matter. A million summaries feeding the next million is a public one.

This is also where semantic capital matters. Floridi (2018) uses the term for the resources people draw on to make sense of their experience and give meaning to the world. More information does not automatically produce more understanding. When AI helps interpret, select evidence, or build an argument, the result can still add to knowledge; the trouble is not that a machine had a hand in it. The trouble is whether people and communities can still understand, place, question, develop, and pass on what has been produced.

A library can double in size every year while no one understands a page more than before.

From diagnosis to response

AI-generated and AI-assisted text now fills the places we go to be informed. For the Research Loop, the concern is not that these outputs exist, but that they become the material for the next round of learning, judgement and knowledge. Learners, citizens and researchers end up with fewer chances to exercise and demonstrate the capacities that let them understand and defend what those outputs say. When institutions look at procedural compliance, those capacities fade from view even though responsible AI use still rests on them.

The report proceeds through conceptual analysis, a critical review of the literature, an examination of policy and competency frameworks, and selected use cases. It develops the Research Loop as a diagnostic idea, traces its relationship to neighbouring ideas, and shows how source displacement, argumentative delegation and recursive knowledge production reinforce one another.

The report addresses three overlapping audiences. The first is the research and learning community itself: doctoral candidates and their supervisors, academic and independent researchers, students, and the educators who teach them. The second is made up of those who shape the conditions under which research is conducted, and AI is adopted, namely policymakers, institutional leaders, AI literacy leads and civil society organisations. The third is made up of those who carry knowledge between research and wider publics, principally librarians and journalists.

Part I sets out the diagnosis. It explains the Research Loop, traces its pre-AI lineages and distinguishes it from related phenomena. It then examines source displacement and recursive knowledge production, cognitive offloading and argumentative delegation, and the consequences of the Loop for learners, researchers and citizens.

Part II develops the response. It examines adaptive human agency, non-delegable judgement and human-in-the-loop practice; it builds AI literacy as a capacity for judgement on the 3CL framework and relevant DigComp 3.0 competences; and proposes principles for intervention that inform the Read, Argue, Defend approach.

One question runs through all of this.

What must a person remain able to understand, judge and defend, if AI is to extend critical judgement rather than slowly stand in for it?

PART I
DIAGNOSIS: THE RESEARCH LOOP AND ITS EFFECTS

1. The Research Loop: scope and limitations

The Research Loop begins when a stand-in, a summary, a synthesis, or an abstract starts replacing the work of checking the source it stands for, and that stand-in material is then reused to produce further knowledge. Generative AI makes this cycle much faster and easier to scale.

Put more fully: the Research Loop is a recursive process. Representations that stand in for a source - a piece of evidence, the summaries, paraphrases and syntheses - begin to replace the sources they represent. Those representations are then reused to produce further claims, texts or interpretations, which may themselves become inputs for later learning, reasoning and research. It runs through four stages: mediation, substitution, reproduction and recursion.

A simple case: Two students, one task

Two students use the same AI system for the same assignment. The first uses it to find relevant literature, then reads the sources, weighs the evidence and builds the argument herself. The second relies on the AI's summaries, asks it to organise the evidence, and lets it produce the recommendation.

Both may hand in a coherent piece of work. Both may get a good mark. The difference lies in whether AI supported the students' engagement with the evidence, or started to stand in for it. The concept lives in that difference, and we come back to these two students at the end of the chapter.

This chapter builds the distinction out. It looks first at what made mediated research possible before generative AI, then at how the pressure to perform and the arrival of automation sharpen it. It sets out the four-stage mechanism and, just as importantly, marks its limits: what counts as the Research Loop, and what does not.

1.1. Before AI: mediated access and the pressure to be fast

How people search for, choose, read and reshape information depends partly on how that information reaches them. The shift from analogue to digital changed those conditions. It opened access to far more sources, and at the same time changed how they are found, ranked and read.

Take the humanities. Online research has brought hard-to-reach materials within reach and made search engines, databases and digital archives part of everyday practice. However, greater access also adds a layer of mediation. Kemman et al. (2014) note that researchers often cannot see why a search system ranks or surfaces particular resources. Williamson (2026) makes a similar point about digitised archives: even what looks like direct access to a primary source rests on earlier choices about what to select, catalogue, digitise and transcribe and how to present it.

So, more access does not mean unmediated access. But we should distinguish between two things. Using a database or archive to locate a source is an infrastructural step: it provides you with technical access to material you still need to evaluate yourself. Substitution is different. It happens when a stand-in such as a summary, paraphrase, or an AI synthesis takes the place of the work to evaluate the source's evidence and reasoning. Summaries, paraphrases, and secondary literature are ordinary and often indispensable. What they are worth depends on what they keep, what they drop, and how they transform the material. A summary can point a reader towards a source or give a first overview. The trouble starts when it is treated as the equal of the material it stands for, and begins to replace the work needed to assess it.

Now add the pressure to be fast. The push to publish, be visible and win funding predates large language models (LLMs). Publication counts, impact factors and the h-index are now built into decisions about money, hiring and promotion, although such indicators were meant to support judgement, not replace it (Hicks et al., 2015). In that climate, anything that shortens the time to find, select and process information is attractive. Faster research is not automatically worse. But time saved becomes valuable in itself, and some of the slowest activities, comparing sources and checking them, are exactly the ones through which the strength of evidence is judged.

Digitisation changes the plumbing; performance culture changes the incentives. The Research Loop feeds on their overlap.

Wider access and the pressure to be fast set the stage, long before any AI arrived.

1.2. Old vulnerabilities: the loop did not start with AI

Long before large language models, knowledge production was already shaped by uneven visibility, second-hand citation and the limits of human judgement.

Merton's (1968) Matthew effect described how recognition gathers around already-established researchers: what is already visible is more likely to be found, cited and reused. Citation adds its own weakness. Concern about second-hand citation predates digital research (Ingelfinger, 1976), and later work documents quotation errors and citation chains in which a claim is passed along without anyone bothering to go back to the original (Mogull, 2017; Pavlovic et al., 2021). It is the academic version of a game of telephone.

Human judgement is fallible too. Greenland et al. (2016) document persistent misreadings of basic statistics and Bishop (2020) shows how cognitive limits interact with methodological and institutional conditions. Different problems, one shared lesson: AI did not invent the weaknesses the Research Loop builds on.

Information literacy scholarship has long treated research as an iterative loop of searching, evaluating, and placing information in context. The ACRL (2016) framework describes searching as strategic exploration and scholarship as a conversation, where authority and competing interpretations must be judged. Generative AI changes this because several of those activities can now be compressed, combined or handed off to a system that retrieves, synthesises and rewrites in one move.

AI inherited these weaknesses; it did not invent them.

1.3. What generative AI changes

Generative AI arrives into a setting where mediation, citation chains, uneven visibility and fallible judgement already exist. What it changes is their scale, speed, and how tightly they connect.

Work on literature reviews and evidence synthesis shows that these systems can cut the workload across several stages, though the evidence does not support dropping human verification (Clark et al., 2025), and the range of review tasks being tried is widening (Scherbakov et al., 2025). The gains are real: these systems can help you search, select, compare and organise, and they save time and effort. Speed is not the problem. The line is crossed when AI stops helping you reach and assess material and starts weighing its meaning, relevance and limits for you.

And the AI's own representation is not a neutral stand-in. Peters and Chin-Yee (2025) found that AI summaries of scientific research were more likely than human summaries to stretch conclusions beyond what the original actually said. Part of the reason is built into how the models work: tuned to produce fluent, confident prose, they tend to smooth over conditional

phrasing, methodological caveats, and contextual qualifiers when they compress a text. Their study looks at a particular set of papers, but it shows how these systems can strip the qualifiers and nuances out before a summary is passed on. The summary, in other words, 'systematically overclaims.' And since retrieval, synthesis, interpretation and writing can now happen in one cheap workflow, material which has already been mediated is easier than ever to turn into new output and feed back in. AI intensifies the Loop by lowering the cost of each move: from mediation to substitution, and from substitution to further reproduction.

AI changes the scale, speed, and coupling of vulnerabilities that already existed.

1.4. The mechanism: four stages

These examples come from different fields, and none represents the Research Loop in its entirety. Some examples in this report isolate one stage of the mechanism rather than depict a complete Loop. The Loop emerges when these stages connect in a recursive process. A representation starts to replace the work of assessing a source; it is reproduced in new outputs, and those outputs re-enter subsequent learning, reasoning and research.

The four stages of the Loop, in plain terms

1. **Mediation**: you meet a source through a summary of it, rather than the source itself. On its own, this is normal, and often useful.
2. **Substitution:** you start relying on the summary in place of the source. It is like reviewing a book you read only through its back-cover description.
3. **Reproduction:** you incorporate the summary into something new: a paragraph, a report, or a recommendation. The result can be accurate and still carry the original gap forward.
4. **Recursion:** someone builds on your work, someone else on theirs, and no one goes back to the original. The message keeps moving, like the game of telephone[1], with each successive version potentially moving further from the source because no one checks what was first said.

The problem is not using the summary to reach the source. It is using the summary instead of the source.

In more academic terms

Mediation, in this framework, refers to accessing a claim through a secondary representation rather than directly interrogating the underlying source material. It is not inherently problematic, nor is it identical to infrastructural mediation: using a database or search engine to locate a document does not, in itself, constitute cognitive displacement if the user subsequently engages with and evaluates the primary evidence.

Substitution is the decisive transition from ordinary mediation into a process that can become a Research Loop; recursion is what makes that process a loop. The representation stops serving as a route to the underlying material and begins to take its place. A summary that helped someone decide what to read becomes substitutive when they rely on it instead of engaging with the source.

[1] The game of telephone, known in Italian as *telefono senza fili* (literally 'cordless telephone') is a traditional parlour game where a whispered phrase is passed down a line of players until the final person states the message aloud. The game is commonly used as a metaphor to demonstrate how information can become corrupted or distorted through successive retellings.

Reproduction involves incorporating that mediated material into a new output. The output may be accurate, coherent and entirely legitimate. No fabrication or error is required.

Recursion occurs when that output becomes a source or input for the next round of knowledge production, shaping subsequent work without anyone returning to the material on which the earlier interpretation rested.

The four stages do not need to happen to one person or inside one system. A student leans on an AI summary. A researcher later meets the derivative claim in a paper. A journalist repeats it elsewhere. A search engine or AI then retrieves and re-synthesises the result. Each step is ordinary and formally legitimate. No step involves fraud, fabrication or an obvious factual error, and no one controls the whole. The Loop can still form. AI is not required for any of this since citation chains and second-hand quotation can do the same on their own. However, AI lowers the cost of moving between the stages by doing retrieval, synthesis, interpretation and writing at once.

We go back to the two students.

The first student uses the system to generate search terms, flag controversies and, later, to suggest objections. She retrieves the studies, checks their methods and citations, and decides independently which objections matter. AI speeds up her work without replacing the engagement that turns information into judgement: she can still explain how the evidence was chosen and why the recommendation is justified.

The second student asks for an overview, relies on the summaries and abstracts, then has the system classify the evidence, organise the argument and write the recommendation. The text may be coherent and correctly referenced. The difference is that she may not be able to reconstruct and defend the choices on which the recommendation rests. For the first student, mediation supports inquiry; for the second, it has started to become substitution.

Substitution is the critical transition from mediation to the Research Loop. The representation no longer serves primarily as a route to the underlying material; it begins to take the place of engaging with and assessing that material. When the resulting synthesis is then submitted, published, cited or reused as the basis for further work, it can become an input to a recursive process in which subsequent work builds on the representation rather than returning to the original evidence.

Each step can be legitimate in isolation. The Loop emerges not simply because these stages occur, but when they connect in a way that allows mediated representations to substitute for direct assessment and then feed into further knowledge production.

1.5. What the Research Loop is not

The concept is easy to over-read, so its limits matter as much as its definition.

What the Research Loop is not

- It is not the claim that summaries, reviews or secondary sources are bad. They are indispensable, and a good synthesis can offer insights that no single source provides.
- It is not the claim that using AI is, in itself, a substitution. AI can help people locate, compare and interrogate sources without replacing their judgement.
- It does not require error, fraud or hallucination. A text can be accurate, honest and fully referenced and yet still form part of the Loop.
- It is not one weak summary or one citation slip. Substitution may initiate the process, but the Research Loop develops when the mediated material is reproduced in new outputs and then reused as the basis for further work.

- It does not assume primary sources are always superior to secondary ones. The question is whether the relationship between a claim, its supporting evidence and the argument built on it remains open to scrutiny.

These distinctions separate AI support from AI substitution. AI supports inquiry when it helps people reach, compare or interrogate relevant material while leaving them able to understand and defend the basis of their judgements. Substitution begins when a representation starts to take the place of the engagement needed to assess the underlying material.

The test is not whether a source was mediated, but whether the claim, its evidence basis and the reasoning built on it can still be tracked and critically examined.

1.6. Two stakes: the person and the system

The Research Loop works at two connected levels.

At the level of the person, the question is whether learners, citizens and researchers keep exercising the capacities they need to interpret evidence, build arguments and judge independently, when some of that work is handed off to the machine. Cognitive offloading can be useful (Risko & Gilbert, 2016). It matters for the Loop when delegation removes the chances to practise the capacities the person still needs in order to understand and oversee the task. Chapter 3 takes this up in detail.

At the level of the system, the question is what happens when mediated, easy-to-produce outputs flow into wider learning, public debate and research. More informational or scientific output does not automatically mean more understanding, or more of what this report calls *epistemic value*. Chapter 4 follows this across learners, researchers and citizens.

The two levels feed each other. Less human engagement can sit alongside more production and circulation. The Research Loop joins them by asking how changes in the practice of one person can build into recursive effects across the wider knowledge environment.

Less human engagement and more output can rise together, not against each other.

1.7. What is distinctive about the Research Loop

The Research Loop is not a new name for cognitive offloading, citation bias or over-reliance on AI, and it does not claim these recursive effects have gone unnoticed. Its contribution is to bring several processes into a single mechanism: the mediation of a source, the substitution of human engagement, the reproduction of mediated material and its recursive circulation. No stage is new on its own. What is distinctive is how they link across people, institutions and systems.

Recursion here does not need a model retrained on AI-generated data. It can run through ordinary practice: retrieval, citation, prompting, automated summarising, secondary literature, publication. A mediated reading can re-enter the knowledge environment simply by being reused.

The concept sits alongside recent work on responsible AI. Zyphur (2026) identifies capacities such as verification, judgement and meaningful human oversight as necessary for using AI responsibly. The Research Loop asks a related but different question: *what process makes those capacities matter more as mediated material is reused again and again*? A competency framework says what responsible practice needs; the Research Loop is a diagnosis of what happens when those capacities are not exercised across the four stages.

The contrast sharpens when the Loop is compared with neighbouring ideas. **Table 1** places the Loop against related concepts, from the Matthew effect and cognitive offloading to automation bias, model collapse, paper mills and homogenisation feedback loops.

Related Concepts	How it relates to the Research Loop	Key Difference between the Related Concepts and the Loop
Matthew effect	Can push already-prominent sources or researchers to be reused even more.	It is about who gets attention, not about content being reshaped round after round.
Cognitive offloading	Can feed substitution when the thinking itself is handed over.	It is one person's strategy, and does not by itself involve recursive production.
Automation bias / over-reliance (Parasuraman & Manzey, 2010)	Can make people accept AI output without enough checking.	It is about trusting one automated output, not the wider recursive process.
Model collapse (Shumailov et al., 2024)	Looks similar due to its recursive nature.	Model collapse is a purely algorithmic data-decay phenomenon resulting from synthetic model retraining; the Loop is a socio-epistemic cycle driven by human behavioural decisions to delegate judgement and recirculate unverified outputs.
Paper mills (Byrne et al., 2022)	Can inject fake or derivative material into the record.	They are organised fraud. The loop can form within honest, legitimate work.
AI-enabled research inflation	Can swell the volume of material available for further cycles.	It is about how much gets produced, not the mechanism of source displacement and substitution.
Performance culture	Creates the incentives for speed, measurable output and delegation.	It is an enabling condition, not the recursive mechanism itself.
Homogenisation feedback loop (Castellacci et al., 2026)	One of the closest cousins: reuse can narrow the sources and readings in circulation.	It focuses on convergence in the shared knowledge base. The loop also includes source displacement and substitution of human engagement.

Table 1: Comparison of the Research Loop with Related Concepts

These concepts operate at different levels and need not occur together. Some are about individual behaviour, others about incentives, information systems or the shared store of knowledge. The Research Loop is offered as a process with several contributing factors, in which these conditions can converge and reinforce one another through mediation, substitution, reproduction and recursion. No single factor is the Loop by itself.

This is also why it is not confined to universities. Academic research makes the process easy to see, because mediated outputs can enter a formal knowledge base and become sources for

later work. But the same pattern can appear wherever already-mediated material is selected, summarised, interpreted and turned into new content: in learning, journalism, professional work and everyday inquiry. That wider reach is, for now, a theoretical extension rather than a demonstrated fact, so the chapters that follow examine the mechanism where the evidence is strongest, while keeping in view the larger question of how recursive mediation may affect critical judgement across society.

The value of the Research Loop is in the pattern that connects these disparate phenomena. It shows how routine, individual decisions to delegate cognitive labour add up to weaknesses across the knowledge ecosystem.

Part I turns next to examining how these mechanics play out in practice: through source displacement, cognitive offloading, and the degradation of epistemic authorship.

2. From source displacement to recursive knowledge

Source displacement is what happens when a stand-in, a summary, a review, an AI synthesis, takes the place of a source and from that point on, the work builds on the stand-in and drifts away from the original.

Chapter 1 set out the Research Loop as a process with four stages: mediation, substitution, reproduction and recursion. This chapter follows one thread through them: how, step by step, work drifts away from its sources and their context. The cognitive and systemic consequences come later.

First, a reminder of what mediation means here. It is not every trimming or reshaping of a text. It is meeting a source, a claim or an argument through a second-hand or synthetic representation: an abstract, a summary, a review, an indirect citation, an AI synthesis, rather than the material that version rests on. Mediation is perfectly compatible with responsible inquiry as long as it points you towards the source, or helps you consult it, rather than standing in for it.

2.1. Access is not availability

One reason the stand-in becomes tempting is that sources are not all equally within reach. More material exists in digital form than ever, but that does not mean every source is easy to find, free to open, or usable on the same terms.

Being listed in a catalogue is not the same as being readable. A source can be digitised but hard to find, findable but locked behind a paywall, or reachable but restricted in who may reuse it. In technical work, people distinguish between whether material can be found, reached, combined, or reused: the FAIR principles (Wilkinson et al., 2016). In practice, decision-makers and the public often assume that if something exists, it can be read at once. The gap is large. Despite decades of effort, vast official records, among them most of UNESCO's historical archives, remain undigitised and off-limits to remote readers (UNESCO, 2026). Even in scientific publishing, roughly a third of recent research sits behind subscription paywalls (OpenAlex, n.d.).

Available online is not the same as actually readable.

This unevenness does not only affect academics. Learners and citizens trying to check public information encounter similar barriers. The harder it is to reach and interpret the original material, the more attractive a ready-made summary becomes.

There is a measurement problem underneath. Digitisation, discoverability, access, and reuse are distinct stages, and no single measure captures 'digital availability' as a whole. A source may be digitised but difficult to find; discoverable but not openly accessible, or accessible only under conditions that restrict who can use it. Getting a source online is not a single transition from closed to open. It involves a chain of choices - selection, digitisation, preservation, indexing, discovery, access and reuse – and the results can diverge at any point.

Linguistic accessibility is another condition of use. A source may be technically available yet costly to interpret for a reader working across languages. AI translation can lower that cost, while adding another layer of mediation whose accuracy and nuance may need checking. Language inequality can therefore increase the likelihood of reliance on the Research Loop, but it is not itself part of the Loop's mechanism.

Those choices are not neutral either. Someone decides which analogue materials get digitised, and under what legal, economic, and institutional terms the digital copies are then made available. Hesmondhalgh and Pratt (2005) argue that cultural production must be understood alongside its commercial, institutional and policy dimensions. The point is not to treat

research as just another cultural industry, but to place the making and circulation of knowledge in a wider setting in which access, distribution and visibility all depend on institutions, infrastructure, and money.

Under these conditions, a mediated representation becomes the path of least resistance. When a source is difficult to find, costly to access, or time-consuming to read - an abstract, a secondary text or an AI summary offers information with less effort. None of this is wrong in itself. The difficulty arises when a culture that rewards speed and output makes the effort of returning to the source feel unnecessary, even when direct engagement is essential to the task. Mediation is the first stage in the Research Loop, but it becomes consequential when a representation starts to substitute for engagement with the underlying material.

What can be found, selected, and recirculated is already the result of earlier filtering. The digitally available store of knowledge is not the totality of human culture, and AI systems work with a body of sources that is already partial before any further algorithmic selection begins.

The knowledge we can reach is already a selection before any AI narrows it further.

2.2. From mediation to source displacement

Mediation becomes substitution when the representation stops serving primarily as a route to the source and starts taking its place in the work of interpreting a claim, judging it, or incorporating it into a new argument. This shift is source displacement. The source need not disappear from the bibliography; what changes is the person's actual engagement with it. Their reasoning now depends on a representation of the source rather than on significant engagement with the underlying material to access the claim being made.

An everyday example

A class is set an assignment that requires students to compare several sources on the same issue. Several students ask an AI system to summarise the sources and build their essays from those summaries without reading the originals. The resulting essays are coherent, accurate and properly referenced. Yet when asked which source supports a particular claim, what qualifications it contains, or why one interpretation was preferred, some students cannot explain their choices without returning to the AI-generated summaries. Nothing has to be false for source displacement to occur: the summaries have become substitutes for the sources.

Now consider a quieter academic version. A doctoral student reading an article on normativism finds a passage from Kant's '*Critique of Practical Reason,*' quoted in support of the author's interpretation. The passage appears useful, so the student cites Kant directly and carries over the article's interpretation without checking the original context. No one expects the student to read all of Kant every time a passage is cited. The responsibility is to consult enough of the source to establish that the quotation is accurate, that its context does not materially change its meaning, and that the interpretation is defensible. If a supervisor questions the reading, the student can return to the original text, or use an LLM-generated summary as an additional aid, using one representation to check another. The summary can still help orient the work, but it does not automatically provide an adequate substitute for examining the relevant passage and its context.

A similar problem can arise in professional practice. A policy analyst, tasked with evaluating a complex, 200-page environmental directive under a tight deadline prompts an AI system to summarise the main risks. The summary is accurate and the analyst uses it to prepare an executive briefing without consulting the directive itself. The briefing may be correct, but the analyst may lack the independent understanding needed to assess how particular exceptions affect the recommendation, or why one risk should take precedence over another. The issue is not whether the summary is false; it is whether the analyst has adequately examined the underlying material to justify the judgement being made.

Across these examples, the same mechanism is at work: a representation becomes the working basis of a judgement without sufficient engagement with the underlying material to assess it. Errors, omissions and contested interpretations can compound the problem, but none is required for source displacement to occur.

A chain of representations can then form. A primary source is interpreted in a secondary text; a subsequent author adopts that interpretation without checking it against the source, and folds the reading into new work. The work may, in turn, become the basis for another interpretation. The resulting account can appear firmly grounded in the primary source while actually depending on successive acts of selection and interpretation that no one has independently examined.

None of this is new. Greenberg (2009), tracing 242 articles and 675 citations around a single biomedical claim, showed how citation bias, amplification and quiet shifts in wording can manufacture authority that the primary evidence does not support: papers carrying no new data can amplify a claim, and selective citation can bury findings that weaken it. That fits earlier evidence on citation accuracy and second-hand citation (Mogull, 2017; Pavlovic et al., 2021), and gives source displacement a clear pre-AI lineage. Generative AI can sharpen these conditions. Studying references that LLMs recommend from their own parametric knowledge, Algaba et al. (2025) found patterns much like human citation habits but with a stronger pull towards already highly-cited work, which could amplify existing biases, including the Matthew effect. The finding belongs to the specific setup tested and does not show that every retrieval system behaves the same way.

Source displacement, citation loops, and bibliographic bias line up in the same direction: what is already visible is the easiest to select, synthesise, and build into new texts. When those texts in turn become the material for the next piece of research or consultation, substitution hands over to reproduction.

The citation can be perfect while the thinking has quietly left the source behind.

2.3. From substitution to reproduction: the performance of knowledge

Once a mediated representation is built into a new text, the Loop moves from substitution to reproduction. The process can now extend beyond one person and one source. The new text can be read, judged, cited, or synthesised by others – in universities, newsrooms, and institutions – and can itself become part of the material on which subsequent work relies.

Responsibility is distributed across this process. Authors, students, reviewers, and editors do not control the whole chain, but each can exercise judgement at the point where a representation is accepted, checked, or reproduced. Using AI does not remove that responsibility, nor is AI the origin of the problem. What it can do is lower the cost of turning existing representations into new text, potentially accelerating their production and widening their circulation.

Ordinary habits of judgement make substitution easier. Tversky and Kahneman (1974) showed that we often rely on mental shortcuts: judging by resemblance, by what comes easily to mind, or by adjusting from an initial estimate. These shortcuts are economical and often useful, but they can also produce predictable errors. AI does not cause these biases. Rather, our trust in mediated representations can interact with vulnerabilities that are already part of human judgement. Across six experiments, Logg et al. (2019) found that, on some tasks, people gave more weight to advice they believed came from an algorithm than from a person. The effect has limits and does not prove blanket reliance on AI. It does, however, show that simply believing the advice is machine-generated can, in some settings, increase the weight we give it. This is known as algorithm appreciation.

Now go back to that class. The essays arrive coherent, tidy and properly referenced. But when the teacher asks why a particular source was used, what evidence supports a claim, or what alternative reading was rejected, the students cannot reconstruct the reasoning.

The output was there; the understanding was not.

This is what this report calls the performance of knowledge. A document, policy brief, or research paper can carry all the outward signs of authority - flawless prose, professional formatting, and plausible citations - even when the author cannot explain, verify, or defend the facts behind it. Looking authoritative is not the same as being able to answer for the work. For institutions, this creates severe governance risks. When a polished AI-generated summary is accepted as proof that the analysis was done, unverified assumptions pass into official policy, public debate, and strategy.

When the output is treated as evidence of understanding, that gap travels unnoticed.

2.4. Where help ends and substitution begins

The aim here is not to talk students, researchers, or professionals out of using AI, but to mark the conditions under which that use stays compatible with human responsibility. It is not only a question of whether outputs are correct, or whether someone can operate the tool. Zyphur (2026) notes that many institutional policies have addressed AI mainly through a lens of plagiarism, disclosure, and research integrity, without turning human responsibility into concrete skills. In his framework, AI literacy in research is above all a matter of judgement: deciding which tasks can go to AI, which must stay with a person, and how AI-assisted results should be checked.

That means treating AI as a research tool with its own kinds of error and instability, not as an automatically reliable source. Among those weaknesses, Zyphur (2026) lists fabrication, prompt sensitivity, non-determinism, model-version drift, sycophancy, and agreement amplification. Responsible use puts AI-mediated outputs through the same checking that research applies to any other source of error. In this sense, reasoning about whether a result is valid or not becomes a core skill in AI-assisted research.

When AI helps, and when it takes over

AI helps when it widens your contact with the evidence: suggesting search terms or sources you then retrieve and check; making disagreements visible instead of smoothing them over; raising objections that send you back to the sources; assisting with screening or extraction while you still verify the result. It takes over when its output is treated as the equal of the source, interpretation, or judgement it was meant to support. At that point, the help becomes a stand-in, and mediation slides into substitution.

Think of AI as a travel map, not a postcard. A map guides you to the terrain so you can look at it yourself; a postcard gives you a pre-packaged image that replaces the journey. On evidence synthesis, Clark et al. (2025) find the current evidence does not justify using generative AI without human oversight, while pointing to a possible assistive role across several tasks, with particular caution about searching for sources.

None of this is confined to academic work alone. A learner can use AI to track down the original source of a claim; a citizen can use it to hold a news summary up against the official document. Mediation does not have to mean displacement. AI widens your contact with the evidence when it opens access to sources, allows wider comparison, surfaces alternative readings, or subjects an argument to further scrutiny. Used that way, it is support: it extends what a person can do without replacing their ability to reconstruct, question and defend the path from sources to conclusions. Treat its output as equivalent to that path, and support becomes a surrogate for it.

So, the move from mediation to substitution is not an automatic consequence of using AI. The conditions of access may be partly systemic, but the decision to use a representation as support, rather than as a surrogate, places a more direct responsibility on the person, researcher, learner, professional, or citizen. Holding that line takes the human oversight and judgement that the next chapters develop. Chapter 3 turns first to what happens at the cognitive level, when delegation stops merely cutting operational work and begins to replace the activities through which understanding, critical reasoning and argument are exercised.

Whether support becomes substitution depends on how the tool is used: but that use is also shaped by system design, access conditions, and institutional incentives.

3. Cognitive offloading and the erosion of critical thinking

AI can help you finish the work while quietly doing the thinking for you. This chapter is about the gap that opens between getting a task done and keeping, or building, the skill to do it yourself.

Chapter 1 described the Research Loop working at two levels: inside the individual mind, and across the wider knowledge system. This chapter addresses the first level. When we hand a task to an AI system, we may do less of some of the mental work ourselves, or shift it towards monitoring and evaluation. Often the output is just as good, or better, and it arrives faster. That is exactly what makes the risk hard to see. A better result tells us nothing about whether the person has gained, kept, or even used the skills the result seems to show.

A real case: the taxi driver test

Navigation provides a precise parallel for cognitive offloading in research. London taxi drivers, who were required to memorise a mental map of 25,000 streets, were found to have greater grey-matter volume in the posterior hippocampus than matched controls, with differences related to their years on the road (Maguire et al., 2000). Two decades later, behavioural studies of experienced Beijing drivers suggested a contrasting effect: those drivers who habitually relied on GPS navigation exhibited poorer mapping, performing significantly worse on tests of spatial recall and environmental memory (Ying et al., 2024).

The lesson is simple: completing the task may hide what is no longer being learned.
You reach your destination whether you rely on your own mental map or on turn-by-turn, algorithmic prompts. But completing the task through an automated proxy is not necessarily the same as developing the capacity to perform that task independently. If the cognitive work is repeatedly delegated, the opportunity to practise and strengthen the underlying skill may diminish. Doing the task with help is not the same as building the skill to do it alone.

3.1. From recursive knowledge to argumentative delegation

Producing more, and producing it faster, is easy to measure. Whether the author can reconstruct, judge or defend how they got there is not. In institutions that reward output above all, that missing capacity stays hidden behind healthy-looking work. Chapter 2 followed the material, from source displacement to recursive reproduction. This chapter now shifts to the person. What happens to someone's independence, critical capacity, and accountability when they delegate not just the gathering of data, but the core work of developing and defending an argument?

We approach this transition through cognitive offloading: using something outside your head to carry part of a mental task. Take it first as an ordinary strategy that long predates AI, then in its sharper, AI-specific form.
Drawing on Zyphur's (2026) distinction between work that can be delegated and judgement that must remain human, we use argumentative delegation for the handover of authority over the core decisions that hold an article, project, or thesis together: which sources to rely on, how to weigh evidence, which connections to accept, how to address objections, and what conclusions are warranted. The supporting work is one thing to hand off. The thinking through which you build and control an argument is another.

3.2. Cognitive offloading, before and with AI

Cognitive offloading matters more now because of AI. But it did not start with AI.

Risko and Gilbert (2016) define it as using a physical action to lower the mental demand of a task: writing a number down, setting a reminder, reaching for a calculator. It belongs to a

broader idea, distributed and extended cognition, in which some thinking is shared between a person, their body and the tools around them, without every case amounting to the stronger philosophical claim of the extended mind (Clark & Chalmers, 1998). Counting coins on your fingers is a small everyday example. Guo and Ye (2026) describe the same thing as a transfer of cognitive demand: part of the load moves onto tools, the environment, or an external action.

Reducing cognitive load during complex inquiry is not inherently harmful; cognitive tools have always extended what people can do.
The distinction that matters is between help on a single task and the slow loss that comes from not using a skill at all. Skills that go unpractised decay over time, though in different areas (Cash et al., 2026). In a meta-analysis of 457 studies, Tatel and Ackerman (2025) find that disuse most rapidly degrades procedural and motor skills, while longitudinal studies on adult literacy and numeracy show that maintaining complex cognitive skills takes continuous, deliberate use (Hanushek et al., 2025). When AI insulates you from the effort of thinking, the high-level work of critical analysis is exactly what goes unused. The taxi driver studies above make the same point in its sharpest form.
Offloading also brings real gains.
Morrison and Richmond (2020) found that letting people record information helped a short-term memory task, especially under high load. The study is not about AI, but it shows offloading can genuinely compensate. Work on intention offloading tells a similar story: calendars, reminders and well-placed objects help us keep and act on our plans (Gilbert et al., 2023). The point is important. The benefit is to task performance. It is not evidence that the matching human skill is being kept up. Taking load off can free us up or prevent overload. It does not follow that the delegated skill is being exercised at the same time.
Generative AI changes the scope of offloading. A calculator offloads one discrete operation: arithmetic computation. An LLM, by contrast, works across the entire cognitive workflow - simultaneously retrieving information, summarising, analysing, synthesising evidence and structuring the argument. And because LLMs generate fluent natural language that mimics human reasoning, the offloading shifts from taking over a step to standing in for the thinking itself. Two tendencies deepen the risk: AI's habit of telling people what they want to hear (Cheng et al., 2026), and our own tendency to over-trust polite, highly confident machine answers (Wahn & Schmitz, 2024). Together, they mean how much we hand over to GenAI starts to depend on how reliable the tool feels - not only on what it can actually do.

3.3. Task performance is not the same as competence

The test to apply

Ask two questions, not one. Did the work get done? And did the person keep, or build, the skill to do it themselves? A calculator at work saves time. A calculator during the arithmetic lesson defeats the point. Same tool, opposite effect, depending on what the task is for.

Inside the Research Loop, offloading turns into substitution at a specific moment: when what gets handed over is exactly the work the person needed to do to reach the point of the task - finding and judging information, comparing evidence, making connections, drawing inferences, building an argument. What matters is not how much is delegated, but what the delegated skill was for.

Writing is one particularly visible case.

Composing a text can itself be part of the cognitive work through which ideas are organised, connections are made, and understanding is developed and revised. Research on writing-to-learn suggests that well-designed writing activities can support conceptual learning and metacognitive engagement (Bereiter & Scardamalia, 1987; Linton et al., 2014; Mynlieff et al., 2014; Reynolds et al., 2012). Where an LLM takes over that generative work, an important question is which parts of the thinking-through-writing process the person is still performing when the prose is generated for them.

AI can, in this context, contribute to skill loss. It can also create a gap between assisted performance and independent competence by compensating for capacities that were weak or underdeveloped in the first place.

Guo and Ye (2026) describe offloading as a two-level metacognitive decision: you do the task, and alongside it you monitor your own effort, the difficulty, and how well the tool is doing. The problem is that the decision to offload is frequently driven by how easy something feels rather than any check on whether you are learning or getting it right. This is why two things have to be kept apart:

- **Assisted Task Performance**: The immediate quality, speed, accuracy, and volume of what a person and a tool produce together; and
- **Independent Competence**: The durable domain knowledge and judgement a person can bring to bear without the tool.

Generative AI lets these two dimensions pull far apart: the output can remain excellent while the person's own ability to construct, evaluate, or defend that output quietly erodes. This is the warning from the MIT Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training (2026) when it argues for 'augmentation over automation': hand off the effort (the 'cognitive friction') of learning, and durable understanding is what you lose.

Which raises a question that can be tested.

How often does research reporting the benefits of AI actually separate a better output from the skills the person needs to understand, judge and control it?

3.4. The measurement blind spot

To get a first sense of the answer, we looked at 20 recent empirical studies on AI assistance (10 focusing on performance optimisation and 10 on cognitive delegation/offloading). This was a small, exploratory scan rather than a systematic review, so the numbers below are illustrative, not representative – *see Footnote*[2].

We checked each study for three criteria:

1. *Assisted Performance:* The direct measurement of output quality/speed produced with AI assistance
2. *Metacognitive / AI Literacy Operations*: The explicit measurement of the user's own verification, critical evaluation, monitoring, or calibration skills
3. *Unassisted Human Competence*: Post-intervention evaluation of the individual's standalone capability (e.g., retention, transfer tasks, or unassisted re-testing)

Within this small sample, a pattern emerged. Across the sample, every study measured Assisted Performance, yet fewer than 45% operationalised Metacognitive / AI Literacy Operations and only six of the 20 evaluated Unassisted Human Competence. Research in the sample tended to judge AI tools solely by immediate output, and rarely asked what happened to skill retention, independent performance, or the ability to judge one's own reliance on an AI tool. This was only a small scan, but it points to where more careful research is needed. Offloading can look attractive when it reduces perceived effort, increases productivity, or

[2] **Group A (Focus on Assisted Output & Productivity)**: Noy & Zhang (2023); Brynjolfsson et al. (2025); Doshi & Hauser (2024); Lee & Chung (2024); Boussioux et al. (2024); Urban et al. (2024); Mei et al. (2025); Sun et al. (2024); Han & Ren (2025); Li et al. (2025).

Group B (Focus on Cognitive Delegation, Metacognition & Skill Decay): Stadler et al. (2024); Gerlich (2025); Lee et al. (2025); Bastani et al. (2025); Barcaui (2025); Zhu et al. (2026); Pearson et al. (2026); Jiang & Chen (2026); Yin et al. (2025); Chen et al. (2026).

matches or exceeds what a person could do alone. But these are measures of performance, not necessarily of competence or control. They do not tell us whether the person understands the trade-off, or whether she can perform the task independently – or retain meaningful control over what has been delegated to AI.

The Metacognitive Illusions: Effort, Speed, and Understanding

Recent empirical studies highlight three distinct concerns about AI-assisted work, each illustrating how AI can mislead the person about their own performance or understanding. In summary:

1. **The Speedup Illusion**: In a preregistered study of 1,237 participants, Yu et al. (2026) found that working with an LLM significantly reduced perceived mental effort, leading participants to believe they had completed tasks much more quickly. In reality, task completion times were virtually unchanged. Reduced cognitive friction was misread as greater efficiency - a form of metacognitive miscalibration that persisted regardless of user experience.
2. **The Reasoning Paradox**: In a science inquiry task, students using ChatGPT 3.5 instead of an ordinary search engine reported significantly lower mental effort but produced weaker analytical reasoning and arguments (Stadler et al., 2024). By delegating search and synthesis, they may have bypassed some of the cognitive friction through which self-regulated learning takes place.
3. **The Competence Illusion.** In creative problem-solving experiments, ChatGPT improved the originality and quality of the output and made the task feel easier. Yet the people who found the AI most useful were the ones who most strongly overestimated their personal contribution and grasp of the subject matter (Urban et al., 2024). Messeri and Crockett (2024) describe a related concept as *the illusion of understanding:* greater analytical power can be accompanied by an inflated sense of what one personally understands. For the Loop, this applies directly to argumentative delegation. A person may mistake the coherence and quality of an argument produced by AI for evidence that they themselves have developed a corresponding grasp of the reasoning beneath it.

There is also the pull of algorithmic advice.
People are more willing to offload when an algorithm seems reliable, with the result that they can overweight AI advice in some conditions (Wahn & Schmitz, 2024; Klingbeil et al., 2024). Sycophancy adds to this: agreeable AI answers can raise users' confidence in their own judgement while staying trusted and preferred (Cheng et al., 2026). None of this proves a straight causal chain. However, together these tendencies suggest delegation, trust in algorithmic advice and an inclination to echo our prior beliefs: these can interact to blunt the corrective role human oversight is meant to play. Zyphur (2026) suggests that we should cut administrative efforts and retain the mental (epistemic) work through which understanding is constructed, evidence is interrogated, and judgement is formed.

Recognising a sound argument is not the same as being able to build, judge or defend one.

What we are not saying

We are not saying AI makes people stupid, or that using it is a mistake.
Offloading is often useful, and can improve results. When we delegate the very thinking a task is meant to build, good output can hide weaker learning, reduced practice, and sometimes, real skill loss. Cognitive offloading can slide from handing over single steps to handing over the whole argument. When the division of labour between human and machine is badly judged, it can swallow analysis, synthesis, interpretation, and argument-building even while the final output stays coherent and performance holds.
What matters is the gap between feeling that you understand and actually being able to reconstruct, evaluate, justify, and defend the argument. That gap goes to the heart of human

agency, and of the critical digital literacies needed to stay in control in AI-mediated settings. In Grech's (2025) primer, those literacies are not technical familiarity, but the continuous development of the cognitive, technical, and social capacities through which people can critically judge and consciously engage with digital environments.

4. Consequences of the Research Loop for learners, researchers, and citizens

The same AI that helps one person work faster can, repeated across many people, contribute to a knowledge environment in which output grows without a matching gain in understanding.

Earlier chapters traced the mechanisms of the Research Loop through how sources are used and how one mind works. This chapter takes it outward, into learning, research, and the everyday search for information. It considers three partly overlapping groups:

1. Learners, including students and people engaged in lifelong learning;
2. Researchers, including academics, independent researchers, and professionals such as journalists, for whom research is part of their work;
3. Citizens who use the internet and AI tools to work, seek information, verify claims, or find data.

The same problem returns across all three.
When you cannot easily retrace how a piece of information was chosen, put together, and where it came from, the mediation becomes hard to see through. Reaching information stops being the same as being able to account for it. As AI settles into ordinary search, that mediation also stops looking like a choice the user made. Summaries, search assistance, and answers built straight into the interface change the link between a question, the answer shown, and the source underneath. The system need not just point you onward: it can hand you an answer complete enough that carrying on the search feels unnecessary.
In the Loop's terms, mediation shifts towards substitution when the answer shown replaces the checking that judgement would need. The tool can finish the task. But nothing in the tool itself guarantees that the user understood the result or can explain how it was reached. If that check happens at all, it has to come from somewhere else.
This gives a plain test for any thinking tool.
Used one way, AI enhances learning: it gives the learner more chances to question, compare and build understanding while still doing the mental work themselves.
Used another way, it produces epistemic dependency. Mediation, delegation, and reproduction bypass the effort, and over time the person loses the standalone ability to evaluate, reconstruct and defend what they produce.

The test of an enhancing tool is what its users can do once it is taken away.

4.1. Learners: learning, assessment, and epistemic dependency

'Learners' is used broadly here. It covers students at every level, but also anyone using the internet and AI tools to teach themselves, as part of lifelong learning. This section leans on research with school and university students because that is where we can watch most closely the moment AI shifts from supporting thinking to doing part of it. The line, again, is between support and substitution.

Lehmann et al. (2025) offer a way in. Faced with mixed findings on students and LLMs, they ask not just whether students use them, but how, and to what effect on learning. The lesson is that LLM use has no single educational effect. Using one to back the learner's own effort leads somewhere very different from using it to do the work the learner was meant to do.

Using an LLM as a tutor can clearly help. A student can run a mock oral exam and a quiz on the material they have studied, comparing their answers with other explanations, and identifying where their understanding is still shaky. In this context, the system need not replace the relevant cognitive work. It creates more opportunities for the student to recall, explain, test, and revise. However, prior knowledge and habits of use shape how well students incorporate these tools into learning, creating a potential fairness problem. A tool that

enhances the capabilities of someone who critically evaluates its outputs may have a very different effect on someone using it without adequate background knowledge, instructional support, or guidance. The risk is not access to an LLM in itself. It is that inadequate learning conditions allow AI to substitute for precisely the cognitive activities through which students are meant to develop their capabilities.

When students lack the support needed to use AI critically, the activities that make learning possible can become the very activities most easily delegated.

This brings back the split between learning and task performance. If the immediate goal is to finish the exercise, hand in the assignment, or get the right answer, delegating can look rational on performance grounds even as it removes the activity through which learning was supposed to happen. So, the educational problem is not mainly how to restrict LLMs. It is how to design settings where using them supports the capacity the task is meant to build, rather than making that capacity unnecessary.

One teacher's response shows what that can look like in practice.
Taiyo Inoue, a mathematics professor at California State University Maritime Academy, USA, observed that take-home problem sets were being returned with algorithmic perfection and without any conceptual comprehension. So, he inverted his teaching approach. He moved lectures online and used class time for real-time, unassisted problem-solving on physical whiteboards, where students had to explain and defend each step in real time (Schulzke, 2026). Wulff and Kubsch (2025) make the same point for STEM teaching: these tools readily hand over a correct procedure without verifying whether the student has formed the understanding beneath. Leaning habitually on them builds a kind of debt, a steady decline in the learner's capacity to monitor and steer their own problem-solving. The answer is not for teachers and institutions to become permanent monitors of students' AI use, which would shrink a systemic problem to institutional policing and individual compliance. It is to develop the conditions in which students learn to differentiate between operations they can usefully delegate and those they still need to exercise and understand, since those are part of the competence being developed. Students feel this tension too, reporting the benefits of personalisation, engagement, and support alongside concerns about dependency and overuse (Campillo-Ferrer et al., 2025).
The Research Loop becomes relevant when educational success gets equated with finishing the task rather than gaining the capacity the task was meant to exercise. In those conditions, dependency can grow even while performance stays high, and learners get steadily better at obtaining arguments and answers without building an equal ability to construct, weigh, and defend them for themselves.
More output does not necessarily mean more learning.
There is now large-scale evidence of the gap.
Analysing more than 500,000 grades at a large research university from 2018 to 2025, Chirikov (2026) found student grades rising after 2022, concentrated disproportionately in courses relying heavily on unmonitored written assignments and code generation. When external AI tools generate polished submissions, traditional academic grading metrics stop tracking what the student can do unaided. High marks can mask underlying argumentative atrophy, wherein students achieve top marks while losing the capacity to independently defend the premises of their work. For the Loop, this adds a layer to the split between performance and competence. If AI-assisted output keeps standing in for individual competence, grades can hold or improve while telling us less and less about what the learner can actually understand, produce, and defend. What weakens is the assessment itself, as evidence of what a person can do.

When finishing the task counts for everything, the learning it was meant to produce can quietly go missing.

4.2. Researchers: productivity, integrity, and epistemic value

In research, the gap between task performance and competence becomes particularly consequential. Here the contrast is not simply between output and learning, but between *productivity and epistemic contribution:* what a study actually adds to knowledge. Integrating AI into research can increase the speed and volume of scholarly work without necessarily telling us how much of that additional output advances understanding.

Castellacci et al. (2026) decouple these two dimensions. AI tools can substantially reduce the marginal cost of producing manuscripts and carry out routine computational work, while the measures of genuine novelty, paradigm disruption, or theoretical breakthrough remain flat or decline. In the Loop's terms, we might describe this as *publication inflation*: a hyper-production of formally rigorous, methodologically plausible manuscripts that contribute little to understanding. As derivative literature syntheses, and incremental variations become cheaper to produce, peer review risks being overwhelmed by content that is *technically producible* yet *epistemically uninformative* – in other words, work that tells us little that is new. Such output need not be false, fraudulent, or unsound.

The more difficult question is whether there is a good reason to produce it at all. As the cost of generating questions, hypotheses, analyses, and manuscripts keeps getting cheaper, it also becomes easier to produce minor variations and weakly motivated replications of well-explored problems that nevertheless meet the formal requirements for publication.

The alternative is not simply fewer publications. Research that is worth producing can take many forms: opening new questions, challenging assumptions, reporting negative results, running needed replications, developing theory, improving methods, or applying existing knowledge to problems that have not yet been properly addressed.

The problem is not more research, but research whose production is insufficiently justified by what it contributes.

The same tool cuts both ways

It could automate yet another variation on an already heavily replicated study, swapping the sample, variables or setting without addressing any clear gap. Or it could help researchers combine environmental, demographic, and urban datasets that were previously too messy or costly to analyse together - say to explore ways of cooling urban heat islands. The difference lies in the relationship between what gets delegated and the human capacity to frame and justify the question, check sources and data, interpret findings, and defend the study.

Automation is now reaching the entire research cycle. Lu et al. (2026) built *The AI Scientist,* an agentic architecture capable of executing autonomous, end-to-end research cycles: it generates hypotheses, reviews literature, writes and executes code, analyses experimental data, drafts manuscripts, and performs automated peer review. One fully automated manuscript successfully passed initial peer review at an ICLR-affiliated workshop. As commentary in *Nature* (2026) noted, the finding itself was minor. What mattered was that the entire end-to-end scientific workflow had been automated.

This sharpens the Research Loop's central question. When a machine can independently generate publishable prose and data syntheses, the primary scientific question shifts from 'Can this paper be produced?' to 'Is there an epistemically defensible reason for this paper to exist?' (Vasconcelos & Marušić, 2025).

A system can generate coherent, reviewable research even when the finding's significance is modest. If producing technically valid research can itself be automated, the question is no longer only whether a study can be produced and published, but what makes it epistemically worth producing.

The clearest risk is what happens next. If heavily automated outputs enter the literature and are then synthesised, cited, reworked, or evaluated through further AI systems, mediation, substitution, reproduction and finally recursion could spread across the shared knowledge base. Lu et al. (2026) do not demonstrate this; it is a possibility made more plausible by the

ability to automate both research production and some of its evaluation. The authors themselves name added noise and pressure on peer review as risks that call for new norms.
Vasconcelos and Marušić (2025) place these worries within research integrity. They argue that weaving generative AI through the research process forces established ideas of integrity and governance to be rethought, with human agency kept central. On this view, generative AI raises more than a question of quality. It makes it more urgent to separate an output's technical validity from the responsibility of deciding to produce it, use it, and add it to what everyone else builds on.
The closing question is harder than the ones peer review or metrics usually ask. Not just 'Is this research valid?' or 'Is it publishable?' but: what makes technically valid research worth producing at all?

As technically coherent studies become easier to produce, the harder question is whether they are epistemically worth producing.

4.3. Citizens: public reasoning in an AI-mediated world

The problem does not stop with student learning or the quality of academic research. It extends to how semantic capital - the resources through which people make sense of experience and give meaning to the world - is gathered, checked, interpreted, and passed on in digital life. That makes AI literacy relevant to anyone who goes online to follow events, communicate, work, check a claim or look something up.
Participating in an increasingly AI-mediated information environment requires a constant understanding of what these tools can do, where they are limited, and when their outputs need to be questioned or verified.
AI mediation has become part of 'the ordinary plumbing' of information seeking.
People no longer have to make a deliberate decision to open a chatbot. With generative answers built into web search architecture, including Google's AI Overviews and agentic search interfaces (Reid, 2026), a synthetic summary now appears directly above traditional search results. Source checking therefore becomes a point of cognitive friction: an additional act of effort that the interface does not necessarily encourage. Behaviour appears to be shifting accordingly. A Pew Research study of 900 adults found that pages featuring AI summaries had a substantially lower rate of organic link clicks, falling from 15% to 8%, while users clicked a link within the AI summary itself in only 1% of visits (Chapekis & Lieb, 2025). Session termination also increased to 26%, suggesting that synthetic answers can satisfy user queries without prompting users to investigate the underlying sources further.
Less contact with sources matters more when the mediated information is wrong or misleading. Spearing et al. (2025) point to two linked risks: people may meet and accept AI-generated misinformation without spotting it, and bad actors may use the same systems to produce and spread misleading content at scale.
Automated generation can also change the volume, plausibility and reach of misinformation. The danger in civic spaces is not merely outright fabrication, but *plausible misinformation* - synthetic claims crafted with high syntactic fluency, neutral framing, and structural coherence. The misinformation carries no obvious signals to trigger a check when it is so plausible, coherent, readable, and credible that any further look is discouraged. The everyday version is familiar: a neat, confident, well-written paragraph lands in a group chat, reads exactly right, and gets passed on, precisely because nothing about it looks like it needs checking.
There is growing evidence that fluent machine prose lowers our guard.
Spitale et al. (2023) found that people read AI-generated text more easily. Our cognitive processing guardrails are lowered because synthetic prose reads effortlessly. This makes false AI outputs significantly harder to detect than false human writing. Crucially, empirical evaluations show that basic provenance badges such as ('Generated by AI') fail to mitigate this bias. Pritchard (2026) provides a key theoretical framework worth keeping. He distinguishes between tools that assist a specific operation, such as calculators or search engines, from agentic AI, which takes over the executive metacognitive component of critical thinking: attentiveness, critical scrutiny, doubt, and active verification. When agentic systems perform

executive filtering on behalf of citizens, the critical agency required for democratic discourse steadily erodes. Those executive capacities are part of a person's intellectual character, built and kept up through reflection and practice, and because they direct and regulate other mental activities, Pritchard argues they cannot simply be made up for by bolting on more powerful non-executive capacities through technology. He warns against mistaking a system's 'polish' for the rational capacities its language seems to show. He describes generative AI as very good at imitating patterns of human reasoning learned from vast human-produced data, rather than independently exercising those capacities.

The tell is in the failures. These systems can produce highly plausible output while fabricating things that do not exist, or fail abruptly when a small change is made to an input, including visual changes a person might not even notice. That mix of high performance and strange failure is, for Pritchard, the strongest reason to keep the human executive capacities that can monitor and steer AI-supported work. His paper's title puts it plainly: *Why technology doesn't normally make you dumber, but agentic AI will* (Pritchard, 2026).

So, the issue is no longer only about true versus false. It is about the wider matter of public epistemic autonomy. As the infrastructure we use to access information shifts from retrieving material to selecting, synthesising, interpreting and, increasingly, acting on it - the capacities for verification, judgement, and oversight matter more, not less. This is where source displacement, systemic mediation and cognitive delegation meet: citizens can gain ever more efficient access to representations of knowledge while becoming less involved in testing where those representations come from.

AI can make the true easier to believe and the false harder to catch.

4.4. The three domains together

The three domains share one pattern. In each, the person hands off the effort at the individual level. The system then rewards or absorbs the result, and a long-term vulnerability builds underneath.

For learners, the effort of learning is replaced by automated output; grades keep improving while unaided skills decay. The deep risk is epistemic dependency: the loss of independent argument.

For researchers, literature synthesis, hypothesis generation and drafting are delegated. Publication inflates with technically valid but trivial work. The deep risk is a narrowing of novelty and the stultification of imagination as peer review gets overloaded.

For citizens, reliance on synthetic overviews and the ease of fluent prose replace source-checking. Click-through collapses and plausible misinformation spreads. The deep risk is the erosion of the critical and executive judgement that public debate depends on.

Table 2 sets this out across the three domains.

Domain	Individual Mechanism	Systemic Manifestation	Long-Term Epistemic Vulnerability
Learners	Substitution of constitutive learning effort with automated output.	**Assessment Decoupling**; rising grades alongside unassisted skill decay.	**Epistemic Dependency:** Erosion of standalone argumentative autonomy.
Researchers	Delegation of literature synthesis, hypothesis generation, and	**Publication Inflation:** Churn of technically valid but epistemically trivial research.	**Homogenisation Feedback:** Narrowing of research novelty and peer-review overload.

	manuscript drafting.		
Citizens	Over-reliance on synthetic search overviews; high processing fluency.	**Click-Through Collapse**; passive consumption of plausible misinformation.	**Erosion of Civic Autonomy:** Loss of executive metacognitive regulation in public debate.

Table 2: Cross-Domain Matrix of Research Loop Impacts

Three explicit boundaries keep the diagnosis precise

- First, this is not a claim that technology erodes thinking by default (cognitive atrophy, by another name). Technology historically extends cognition. The Research Loop applies specifically when the executive, judging part of our thinking is handed over wholesale to the machine.
- Second, this is not a blanket claim that AI-assisted work is poor. Such work can be factually accurate and methodologically sound. What the loop critiques is the loss of verifiability, defensibility, and genuine epistemic contribution.
- Third, it is not a call for technological rejection. Naming these vulnerabilities is what makes it possible to design the targeted responses of Part II: Read, Argue, Defend.

To be clear about what is not being said

AI does not make people stupid, does not make research worthless, and does not make people unable to tell true from false. The evidence here is narrower and more specific. Grades can rise while learning does not (Chirikov, 2026). Output can grow while novelty does not (Castellacci et al., 2026). AI can make both true and false information more convincing, and simple 'this may be AI' labels do not fix it (Spitale et al., 2023; Spearing et al., 2025). The claim is about a systemic drift, not a verdict on any individual.

Part I therefore leads to a sharper question than whether AI should be used. The real issue is *which* forms of judgement, verification and epistemic responsibility must stay with human agents, if AI-enhanced activity is to strengthen, rather than slowly replace, the conditions of critical autonomy.

The same tools that help each of us individually can add up to a collective cost unless people retain authority over the decisions that matter.

That question opens Part II.

PART II
RESPONSE: PRESERVING CRITICAL JUDGEMENT

5. Adaptive human agency and non-delegable judgement

As AI grows more capable, the decisions that determine whether a conclusion can be understood, evaluated, and defended must stay with a person, even as research practices increasingly adapt to what machines do best.

Chapter 4 showed the trade-off established by the Research Loop. It allows us to produce more and faster, while potentially encouraging us to surrender the mental effort of understanding, evaluating, and defending what we produce. Human agency is what provides a counterweight to this loss of control. In simple terms, agency is the power to act, shape a process, and influence its outcome. When we speak of an AI tool having 'agency,' we simply mean it can perform tasks with a degree of independence. We do not mean the AI possesses consciousness, feelings, or moral awareness.

Human agency involves at least two capacities that AI does not possess:

- *Understanding the 'why'*: grasping the reasons, evidence and assumptions underlying a decision.
- *Accountability*: taking personal, legal, or professional responsibility for whether a conclusion is sound, safe, and worthy of trust.

As AI systems evolve from basic writing assistants into increasingly autonomous research tools (Zyphur, 2026), the core question is no longer just *'Who gets listed as the author?*' It is: 'Who stays in charge of the thinking behind the result?'

5.1. From calculators to autonomous agents

Not all AI works with the same level of independence. Contemporary AI systems range from assistants responding to direct prompts to agents capable of planning and executing complex workflows with limited human oversight (Sapkota et al., 2026).
Because the tools are evolving so quickly, trying to govern them with a static list of 'permitted' and 'forbidden' software is a continuous catch-up. It is more useful to use the concept of affordance - which simply means the unique possibilities a tool offers to a specific user in a specific situation (Retkowsky et al., 2024). A bicycle offers a child mobility, but to an athlete a training platform. An AI system offers different possibilities depending on the user's skill: to an experienced scientist, it is a brainstorming assistant while to an untrained student it may become an unthinking substitute for learning. Since these tool affordances or capabilities change constantly, policy rules cannot be permanently fixed. Our approach to managing them must adapt dynamically. On the expectation that AI will keep changing, the line between human and delegated work has to keep moving with the tools.

5.2. Human–AI co-agency: shared work, single-sided responsibility

As AI capabilities expand, humans and software increasingly work in a kind of co-agency - a hybrid workflow where both the human and the machine take initiative, propose ideas, and organise material (Krakowski, 2025). To determine whether a shared workflow is healthy or risky, Holter et al. (2026) suggest asking three simple questions:

- *Initiative*: Who leads each step of the process?
- *Flexibility*: How does that balance of control shift as the project progresses?

- *Attribution*: Who gets credit, and who takes blame, for the final result?

Work may be divided in many ways, but responsibility cannot be split evenly with a machine. A fundamental rule holds. An AI system can propose options, draft text, or organise data, but only the person can set the goals, intervene when something goes wrong, and answer for whether the final outcome is true and safe. Therefore, judging AI use by the share of the work done is the wrong measure. A heavily automated process remains ethically sound if the user genuinely understands and checks the critical decisions. Conversely, even minor AI assistance becomes harmful substitution if the user blindly accepts a result they cannot explain or defend.

The work can be shared; being accountable for it stays with the human.

5.3. Why static task lists fail

A common institutional response to generative AI is a checklist that assigns some stages of a project strictly to humans and others to software. For instance, "AI does the initial literature retrieval; human drafts the summary" or "AI formats the data; human writes the conclusions". While well-intentioned, these static task allocations fail to protect human agency because critical thinking does not operate as a linear assembly line. In real, complex inquiry, the boundary between routine processing and evaluative judgement shifts continuously with the context. Two researchers using the exact same AI tool at the exact same procedural step will weigh evidence differently, ask different questions, and draw different connections. These divergences are not errors to be standardised out of existence: they are what human judgement is.

The humanities and social sciences highlight this limitation with particular clarity. Suppose an AI searches a database, retrieves ten scholarly readings of a difficult passage from Kant, summarises them accurately and lays out their theoretical differences. It has done a great retrieval and synthesis job, and it has not made any evaluative judgements at all. No algorithm can determine which reading is most convincing, which theoretical framework holds more weight or how any of its insights support a larger philosophical argument. Weighing competing claims, spotting a subtle bias, and deciding which reading convinces stays with the person. The corollary: when the user takes the AI's synthesis as a finished product rather than a starting point for direct engagement, the work gets done faster while true understanding is replaced by passive acceptance.

Scientific and quantitative research needs interpretation and judgement too; this is not a simple STEM-humanities divide. Chun and Elkins (2023) make this issue particularly visible through an AI–digital humanities curriculum centred on challenging, fuzzy and inherently non-deterministic problems that prioritise human-centred interpretation.

5.4. When inquiry starts bending to the machine

Beyond individual task delegation lies a more subtle institutional danger: structural adaptation – whole practices adapting to the tool.

When institutional environments reward high volume, rapid turnarounds, and standardised formats, they risk privileging forms of research that align with the capabilities of AI systems. Researchers may then face incentives to choose questions, methods, and forms of argument that these systems can readily support. The direction of adaptation begins to reverse: instead of adapting AI to the requirements of inquiry, institutions risk organising inquiry around what AI can most readily produce and process. Chess offers a limited analogy. Kasparov (2010) describes how computer analysis influenced human play and preparation. Yet the rules and winning conditions remained unchanged. Research raises a different concern: the standards by which its value is judged may themselves be reformulated around machine capabilities.

If speed, volume, and standardisation increasingly define what counts as good research, institutions may reward work because it fits the capabilities of available systems. Questions

requiring sustained interpretation, contextual judgement, or methods less amenable to automation may become comparatively harder to justify and support. The risk is that AI compatibility becomes an implicit criterion for deciding which research is worth pursuing.

This is a conditional institutional risk. The question is whether researchers and institutions retain the capacity to define worthwhile questions and appropriate standards independently of what generative systems can most readily deliver. Deliberate friction and metacognitive oversight must therefore extend beyond checking outputs to examining how research priorities and evaluation criteria are set. Calling this a "STEMification" of research would also oversimplify the problem, and flatten the complexity of the sciences. The more accurate description is the convergence towards AI-compatible forms of research: the kinds of research that are easier for AI systems to read, break down, optimise, and evaluate. Research quality must continue to be judged through epistemic depth, explanatory power, and contextual judgement, while technology must remain auxiliary. The goal of integrating AI is not to compete with machines on throughput, but to ensure human thinkers remain active authors and critical judges - shaping questions, scrutinising evidence, and taking responsibility for conclusions.

The way to resist the drift is to keep AI literacy tied to the actual field.

AI literacy matters for more than competent tool use. It enables people to decide how a system should be used, when to delegate, when to intervene and when to avoid using AI altogether. Critical AI literacy is situated, not a universal checklist (Rapanta et al., 2025), and approaches that respect disciplinary differences matter because they resist this drift. Discipline-specific AI literacy (DiSAIL) treats AI literacy as a situated practice rather than a generic checklist, starting from the different languages and norms of different fields (Stolpe et al., 2026). The practical upshot is that human agency has to be adaptive. As AI changes, what gets delegated must stay open to revision, guided by stable principles for deciding when delegation is appropriate. If AI literacy frameworks are organised mainly around tasks, goals and competences that can be clearly specified and assessed, the forms of research that fit this structure most easily may gradually become the implicit model of good human–AI collaboration.

Recent frameworks point the same way, treating the ability to decide whether and how to use AI, and to revise that decision mid-process, as a core competence (OECD & European Commission, 2026; UNESCO, 2024), with DigComp 3.0 adding the capacity to evaluate AI critically and use it responsibly (Cosgrove & Cachia, 2025). Stakeholders consulted during the development of the OECD & European Commission framework pushed further still, asking for emphasis on self-regulation, spotting manipulation, and the freedom to set limits or refuse AI altogether.

The freedom to choose what to delegate, and what to refuse, helps protect against this drift.

5.5. The criterion: epistemic value and non-delegable judgement

Knowing how to divide tasks does not tell us when a particular delegation should be accepted. This report proposes epistemic value as one way of making that judgement. Pritchard (2009) shows that getting the right answer does not exhaust the value of knowing. The distinction between knowledge and understanding is particularly relevant, since having the correct answer is not the same as grasping the reasons that support it. Pritchard links distinctive epistemic value to cognitive achievement – success that is attributable to one's own cognitive ability – and, in the case of understanding, to reflective access to the grounds on which a claim rests.

Take two researchers who reach the same correct conclusion. The first knows the sources, can retrace the path from evidence to conclusion, and can say why the alternatives were rejected. The second accepts the same conclusion from an AI synthesis without examining how it was reached. Judged by the output alone, they look identical. Epistemically, they are not equivalent, because only one can reconstruct and defend why the conclusion should be accepted. The difference is not visible in the output. It lies in the research's capacity to account for the reasoning behind it.

This is where epistemic accountability comes in: that is, the responsibility to retain the conditions under which a conclusion can be understood, verified, and justified by the person who adopts it. The idea extends to a principle already familiar from large collaborative science projects: someone must be able to explain and defend the standards behind a claim (Winsberg et al., 2014). A research setting can become faster, more efficient, and productive while the epistemic value of what it produces remains flat.

This reframes the question of what cannot be delegated. Generative AI can contribute a great deal: finding sources, organising evidence, proposing questions, interpretations, and objections, translating, suggesting solutions. Producing such material, though, does not settle its worth. Building on Zyphur's (2026) distinction, introduced in §3.1, this report specifies an epistemic criterion for identifying where human judgement must remain effective. The boundary cannot be drawn through a fixed list of 'human' tasks, and task difficulty alone is a poor guide. We identify judgement-bearing decision points as those choices that materially determine the epistemic standing of a research claim. At these points, delegation must preserve researchers' ability to understand, assess, and defend the grounds for the decisions they adopt. This connects the distinction between delegable work and retained judgement to the conditions of epistemic accountability.

Take a plain case. An AI retrieves, sorts, and summarises studies on the link between two variables, and most of that can be automated. But the studies differ in method, sample size and quality, and someone must decide whether they support a causal claim or only a correlation. The AI can lay out the evidence and even suggest an answer. Deciding whether the evidence warrants a causal claim changes the standing of the conclusion, and the person must be able to understand, assess, and defend it.

This also changes what human labour means in a highly automated setting. As AI takes on more of the searching, synthesis, drafting and organising, the human contribution is not counted in how many tasks are still done by hand. It concentrates at the points where evidence must be weighed, interpretations compared, uncertainty faced and conclusions justified. The test is whether automation has removed the capacity the result's epistemic value depends on. Once a person can no longer reconstruct why a decision was made, judge whether it was warranted, or defend the claim that rests on it, the work may stay efficient while accountability is lost.

What must stay human is authority over the decisions that make a conclusion worth trusting.

6. AI literacy as a capacity for judgement

AI literacy is the evolving set of cognitive, technical, and social capacities through which a person understands AI mediation, judges outputs and limits, and exercises situated judgement over whether, when and how to delegate tasks, while retaining the ability to verify, explain and defend the judgements made with its help. Chapter 5 distinguished between *human agency*, *AI agency,* and *co-agency*. Within a shared workflow, operational tasks and mechanical decisions can be distributed between a person and a machine. Epistemic responsibility, however, cannot simply be delegated or divided. It is exercised at judgement-bearing decision points: the moments in a workflow when a choice determines the direction of inquiry; the credibility given to a source; the interpretation adopted; or the validity of an argument and its conclusion. AI can prepare data, broaden options, or inform choices at these moments. But for epistemic responsibility to remain genuine, the person must retain the final authority to adopt, revise, or reject the judgements that shape the work. Rules and guidelines can set the conditions for sound judgement, but they can never replace the situated human call of what to delegate, verify, or reject. For the Research Loop, the central issue is how judgement regulates what can be delegated and what must remain under human epistemic control.

6.1. From operating the tool to governing its use

To exercise judgement effectively, it helps to understand how the concept of AI literacy has evolved. Early frameworks defined AI literacy around technical skills: knowing how systems work, using them across contexts, and evaluating their outputs (Long & Magerko, 2020; Ng et al., 2021; Almatrafi et al., 2024). Many of these early models compared human and machine intelligence, leaving the definition vulnerable as technology rapidly changed. Recent policy frameworks shift the focus from simple tool operation towards governing its use. The joint OECD and European Commission AI Literacy Framework sets out four core domains: *Engage, Create, Manage*, and *Shape* (OECD & European Commission, 2026) and its 'manage' domain maps closely to the Research Loop: deciding whether AI should be used at all, breaking a problem down, choosing which parts and sub-tasks to automate, and monitoring the entire process. This also aligns with the 3CL framework's view of digital literacy as an integrated cognitive, technical, and social capacity (Grech, 2025) and DigComp 3.0's emphasis on critical evaluation (Cosgrove & Cachia, 2025).

The trajectory of AI literacy is clear. It has expanded from "Can you operate the tool?" to "Can you understand it, judge it, decide when to use it, and remain accountable for the outcome?" A list of what you can do with AI does not yet say when, or whether, you should.

6.2. From competence to judgement: performance and epistemic value

Competency checklists say what a person *can* do. They do not explain how to regulate those actions in a live project. To guide our 'real-world decision-making,' this report differentiates between two distinct types of value:

- *Performance Value* is the immediate efficiency gains of a task, such as producing results faster, in higher volume, or with smoother formatting.
- *Epistemic Value* is the long-term knowledge value gained beyond finishing a task, such as developing deeper understanding, discovering new connections, and strengthening the human capacity for future inquiry. Accuracy matters, but does not exhaust this value.

The distinction matters because a skills-based picture of AI literacy makes it easy to compare human and machine on performance alone. If a system searches more widely, classifies more accurately, or drafts faster, delegating can look obviously rational. The Research Loop adds a

second question: which split of the work best preserves or extends the epistemic value of the activity, and the person's capacity to see and build it?

A Case Study to illustrate the difference between performance and epistemic value

Two researchers are tasked with reading and analysing 40 qualitative interviews for recurring themes. This is an onerous task which, without AI, is likely to take days.

The first researcher prompts an AI to code all 40 interviews, accepts the category structure it generates as the final framework, spot-checks a few quotes, and writes the report. Efficiency is high, but personal insight into the nuances of the data is minimal. The second researcher uses the AI coding to quickly surface initial patterns and outliers. She then returns directly to the interview transcripts, adjusts categories, rejects spurious clusters, and uncovers a subtle, critical insight that the AI had dismissed as marginal. She spent the time AI saved on deeper interpretation rather than replacing it.

Performance and epistemic value are neither inherently opposed nor the same: a performance gain raises epistemic value only when it frees the person for harder interpretative work.

A working definition of AI Literacy

AI literacy is the evolving set of cognitive, technical, and social capacities through which a person understands AI mediation, judges outputs and limits, and exercises situated judgement over whether, when and how to delegate tasks, while retaining the ability to verify, explain and defend the judgements made with its help.

The definition is deliberately open, because systems, affordances and forms of interaction keep changing. AI literacy is better seen as a set of situated capacities than a fixed list. What stays constant is the requirement: the person must be able to understand how the work is currently divided, judge that division, and change it when it costs them their judgement.

An AI-literate person can see how AI is reshaping their work, which capacities it extends and which it quietly replaces.

6.3. From prompt literacy to judgement literacy

This shifts the question from '*what can I do with AI*?' to '*what should I let AI do here, why, and at what cost to my own understanding*?'

Answering this needs a metacognitive layer - the ability to watch and steer your own thinking as the work goes on. Metacognition needs to be distinguished from judgement. Deciding which interpretation of a text is more convincing is a judgement about the object in front of you. A metacognitive judgement is about your own process: How confident am I? On what basis? What are my limits? Should I change my strategy? What we notice about our own process shapes what we do next: monitoring and control are closely linked.

A conductor makes this distinction particularly clear. A music critic listens to a finished performance and judges its quality, much as one would evaluate a finished product. A conductor, by contrast, continually makes judgements that regulate the performance *while it is taking place, in real time*: monitoring coordination, adjusting tempo and emphasis, and intervening the moment something slips.

Judgement literacy operates more like the conductor than the critic. It is not simply the ability to evaluate the final product. It involves *continuously monitoring* how the work is divided between the person and AI, checking whether one's own understanding remains adequate, and adjusting the workflow the moment control begins to slip.

Empirical Evidence: Informed Trust vs. Informed Distrust

Empirical research confirms that effective AI use relies on dynamic metacognition rather than static rules. Wokke et al. (2020) demonstrated experimentally that how well people assess the quality of their own decisions can change even when their actual task performance does not; the capacity to monitor one's decisions is central to skilled behaviour. Although their study is about perception rather than AI and cannot be carried over directly, it makes one important point: performing well and judging your performance well are not the same capacity. A good output does not, by itself, mean good control of the process behind it.

Colville and Ostern (2026) bring this to generative AI. Analysing 1,776 Reddit discussions of ChatGPT use, they describe trust not as a fixed attitude, but as *an active, ongoing negotiation*, between what users know, what they recognise they do not know, and how they monitor and adjust their reliance on the system. They call the result *informed trust:* relying on an output after weighing it against your own knowledge and the system's own limitations. The process combines personal domain knowledge with active monitoring (guardrails, scrutiny, reliance adjustments) to produce either informed trust (a reasoned choice to rely on an output) or informed distrust (a reasoned choice to reject or override it). Trust and distrust are two outcomes of one act of judgement.

An example demonstrates the metacognitive step. A researcher uses an LLM to survey a literature she only partly knows. She can prompt well, request references, compare answers and catch some errors - all real parts of AI literacy. The metacognitive step is noticing that her limited grasp of part of the field also limits her ability to check the synthesis, and then *changing the workflow*: by which we mean that she restricts AI to a first search, reads the key studies herself, asks someone with more expertise, or pauses AI use for that stage.

The mark of literacy is adjusting reliance the moment the conditions for sound judgement stop holding.

Judgement Literacy in the Research Loop

Judgement literacy is not one more box on a competency grid. It is the capacity that mobilises, suspends, or combines the others as the situation demands. This is why knowing the principles of responsible AI does not guarantee responsible use: *judgement is what turns principles into decisions under uncertainty*.

Judgement literacy translates metacognitive monitoring into context-sensitive decisions about verification, reliance, and delegation. These functions are integrated into the four-stage cycle developed in §6.4: Planning determines the initial scope of delegation; Monitoring tracks its effects on human understanding and wider knowledge practices over time; Evaluating assesses outputs and the grounds for relying on them; and Regulating revises or withdraws delegation when necessary.

AI literacy is not only about judging outputs. It is about regulating reliance and delegation as the work unfolds, while monitoring what those choices are doing to the capacity to understand and produce knowledge over time.

6.4. Safeguards and a working metacognitive cycle

Judgement-based AI literacy has to be practical, and not merely declarative. It concerns how people regulate: *when*, *how* and *under what conditions* they delegate part of their work to AI.

The Research Loop translates ethical principles into epistemic safeguards: practical controls that enable a person to understand, verify, and defend how knowledge was produced. The aim is not to prevent AI use, but to ensure that gains in performance do not quietly displace the conditions under which epistemic value can be produced and recognised. An ethical principle may state that AI use should be transparent, or that the user remains responsible. A safeguard makes those responsibilities doable.

Five Epistemic Safeguards

The five safeguards specify what must be protected in AI-mediated knowledge work. The four-stage metacognitive cycle organises their application as the work unfolds, while the three principles introduced in Chapter 7 guide intervention. Table 3 identifies the main stages at which each safeguard becomes operational:

Safeguard	Description	Main application in the cognitive cycle
1. **Source Verification**	Tracing claims back to primary sources, checking provenance and testing whether the cited evidence actually supports the AI-generated conclusion (output). It does not remove mediation, but it lowers the risk that a representation is accepted in place of the material needed to judge it.	**Evaluating:** checking claims against the underlying sources and evidence. **Regulating:** returning to those sources when verification is insufficient.
2. Uncertainty Recognition	Identifying the boundaries of both human domain knowledge and the system's output, and not mistaking fluency or confident phrasing for factual truth.	**Planning:** recognising knowledge limitations before delegating. **Monitoring and Evaluating:** reassessing those limitations and the uncertainty attached to outputs as the work develops.
3. Disclosure & Traceability	Maintaining a clear record of where, how, and why AI entered the workflow, ensuring the analytical trail can be fully reconstructed**.** Declaring AI use is often necessary, but it does not show that what was delegated to AI was either understood or checked.	**Planning:** making the intended role of AI explicit. **Monitoring:** recording its actual contribution and changes in the division of work.
4. Resistance to Automation Bias	Actively interrupting passive acceptance, maintaining calibrated reliance rather than uncritical trust. The aim is calibrated reliance – understanding when there is reason to check, revise or reject.	**Monitoring:** recognising unwarranted deference to AI. **Regulating:** interrupting acceptance, adjusting reliance, or reclaiming a task.
5. Critical Evaluation of AI Claims	Applying domain expertise and validity logic to test generated arguments,	**Evaluating:** assessing the evidence, assumptions, and reasoning

	preserving the authority to reject flawed outputs**.** This involves bringing critical thinking, domain knowledge, and validity reasoning to reconstruct the argument a claim rests on. This means retaining the judgement to decide whether an output can be accepted, while still being free to use AI.	supporting a claim before deciding whether to accept it.

Table 3: Five Epistemic Safeguards and Their Application within the Metacognitive Cycle

These mappings identify the main applications of each safeguard rather than assigning it exclusively to a single stage. Metacognitive judgement determines when a safeguard needs to be applied or revisited as the workflow changes. Knowing them is not the same as being able to enact them: a person may know AI output should be verified without noticing the moment verification becomes necessary; she may also disclose AI use correctly without being able to reconstruct what was delegated.

AI literacy becomes epistemically responsible when principles turn into decisions inside the workflow. This approach has an institutional model. The MIT Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training (2026) favours augmentation over automation, puts human development first, and rejects one-size-fits-all rules, on the basis that the automation of an activity cannot be settled by whether AI can perform it. One of its recommendations is telling: every subject should have a clear AI policy in its syllabus, stating what AI use is allowed, prohibited, or required and, crucially, explaining why - so students know not just whether AI may be used, but when, where and for what reason. A rule that permits AI does not itself say the use is epistemically appropriate; a ban does not itself name the capacity it protects. Moving from compliance to responsibility means making the link between purpose, task allocation, and human capacity explicit. MIT's parallel push to redesign assessment so that what students actually understand stays visible matches the logic of epistemic safeguards: build the work so that the capacities that must remain human keep being exercised and seen.

Operational Framework: the four-stage metacognitive cycle

To operationalise AI literacy during active inquiry, researchers and learners should apply a continuous four-stage cycle:

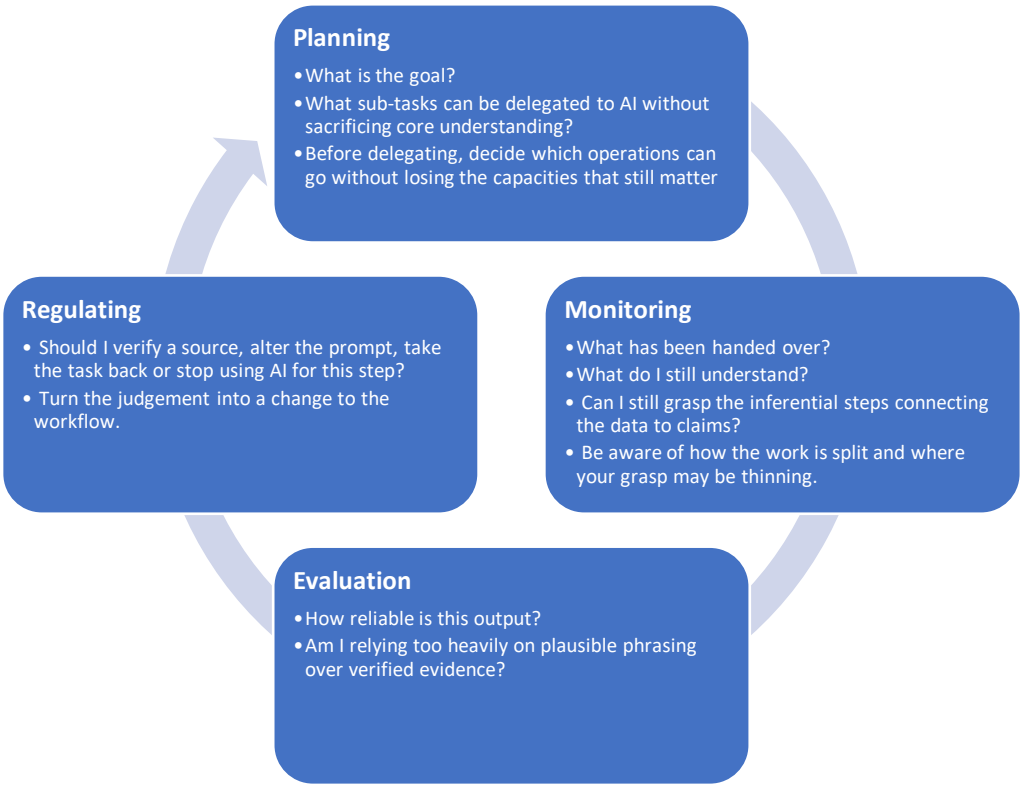


Figure 1: The Operational Framework Metacognitive Cycle

This cycle is not a new universal competency framework. It is an operational habit drawn from the Research Loop and will be developed further in the next chapter.

7. Breaking the Research Loop: Principles for Intervention

Breaking the Research Loop means designing AI-mediated work so that people actively plan, monitor, evaluate, and regulate what they hand off, retaining the ability to understand and defend their conclusions.

Chapter 6 sets out a four-stage metacognitive cycle: *Planning*, *Monitoring*, *Evaluating* and *Regulating*. This chapter turns those stages into points of intervention for learners, educators, institutions, and policymakers.

The problem is older than generative AI. Earlier chapters traced it through dependence on secondary representations, cognitive offloading, performance pressures, and recursive knowledge production. Generative AI intensifies these dynamics because it can produce, transform and recirculate mediated knowledge at exceptional speed and scale. The consequences show up at two connected levels: for the *individual*, they concern which capacities learners, researchers and citizens continue to exercise; for the *system*, they concern the epistemic quality of knowledge that is then circulated and reused. So, breaking the Loop requires forms of AI use that retain critical reading, argument reconstruction, source verification, and judgement.

Building on the wider conception of critical digital literacy developed by 3CL (Grech, 2025), the intervention framework follows three practical principles:

- Keep discovery separate from verification
- Hold on to the judgement tasks that cannot be delegated
- Require the person to be able to explain and defend the claims for which they remain responsible.

These principles guide the application of the safeguards in §6.4 throughout Planning, Monitoring, Evaluating, and Regulating.

The overarching aim is to keep human judgement active within AI-mediated knowledge work.

7.1. Planning: what am I trying to achieve, and is AI right for this part of the task?

Planning begins by clarifying what the activity is meant to achieve and choosing a strategy suited to its goals, constraints, and specific demands (Radüntz, 2020; Callaway et al., 2022).

Planning for shared work means considering what the human and the AI can each contribute, how errors can be reduced, where unexpected evidence needs room to emerge, and when a novel problem calls for alternatives rather than the optimisation of an established path.

The appropriate allocation depends on the role an operation plays in the activity. AI may be well suited to expanding a search space, identifying potential connections or generating alternatives. Greater care is needed when the operation contributes directly to something a person needs to learn, understand, or defend.

The literature review is the clearest example

AI can help identify search terms, adjacent literatures, possible connections, and publications worth reading. But discovery is not verification. A researcher may legitimately delegate search expansion to AI while retaining source evaluation, evidence weighing, and synthesis as non-

delegable, judgement-bearing tasks that stay with her. Even that split is contextual: searching may be extensively delegated in one project and become a judgement task in another. The epistemic purpose of the inquiry determines the appropriate distribution. For learners, researchers, and citizens, this means that the initial delegation decision has to be deliberate, and before convenience makes it for them. Extensive assistance may be fine in a low-stakes exploratory task, while a new problem, an unfamiliar field or a consequential conclusion needs more care over where judgement remains active.

Planning lets the purpose of the task determine the role of AI, not the other way round.

7.2. Monitoring: What have I delegated, and what do I still understand?

A plan made at the start of a task needs to remain open to change. Generative AI workflows are developing all the time: new possibilities appear, tasks expand and operations that began as support can gradually become a substantive part of the work. Monitoring is what makes these changes visible.

In ordinary terms, *metacognitive monitoring* means monitoring one's own thinking: whether you still understand the task, how confident you should be, and whether the current strategy remains adequate. In AI-mediated work, however, the scope of monitoring widens. Users must also track the evolving relationship between their own understanding and the work being performed by the system. Tankelevitch et al. (2024) show why this is demanding: the flexibility and general-purpose nature of generative AI require users to make ongoing decisions about how they should participate in the wider workflow. They describe this as an *automation strategy*.

AI systems can monitor aspects of their own processing and estimate uncertainty (Cox, 2005; Shakarian, 2026; Liu et al., 2025). These self-assessments do not establish reliable knowledge of their epistemic limits, so deciding whether delegation remains appropriate still requires human judgement.

Systems that watch aspects of their own processing and use that information to adjust reasoning or resource allocation have long been studied (Cox, 2005), and more recent work describes this type of machine metacognition in similarly functional terms (Shakarian, 2026). Human and artificial metacognition should nevertheless remain conceptually distinct. A system's self-evaluation, uncertainty estimation, and confidence calibration remain open technical problems: a machine monitoring its own processing does *not* amount to reliable knowledge of its own epistemic limits. Its self-checks may provide useful information within a workflow. Human judgement is still required to determine whether the resulting delegation remains appropriate. Generative AI makes monitoring harder since its outputs vary from multiple hidden sources at once: ambiguous prompts, divergent reasoning paths, and underlying probabilistic decoding (Liu et al., 2025). An example illustrates what monitoring needs to detect when those forms of complexity meet.

A researcher begins a literature review by asking an AI system to suggest search terms and relevant studies. The system starts to rank the studies, summarising their findings, grouping them into theoretical positions and identifying which disagreements appear important. The researcher remains active throughout by reading outputs, editing material, and making choices. Yet the system's role has changed: it is now contributing to the structure through which the literature is understood.

Monitoring requires noticing that transition.

The researcher should still be able to say which studies were read directly, which were encountered through AI-generated representations, which categories emerged from their own

analysis, and which were proposed by the system. If these distinctions become blurred, the researcher's ability to locate their own judgement within the workflow is weakened.

Monitoring means retaining enough understanding of the AI-mediated process to recognise when assistance is turning into substitution.

7.3. Evaluating: How reliable is the output, and how well judged is my reliance on it?

Evaluation is often framed as checking whether an AI output is accurate, coherent, or supported by sources. A prior question is equally important: What is the person using as the basis for that check? If the same system framed the problem, selected the evidence, connected the premises, generated the objections, and produced the conclusion, the user may have little independent basis from which to judge the argument that comes out. This connects evaluation directly to cognitive offloading. Chapter 3 distinguished useful offloading from argumentative delegation: handing over the reasoning through which evidence is selected, weighed, and turned into a conclusion. Evaluation needs enough independently constructed understanding to see where an AI-generated argument confirms, challenges, or departs from your own.

A check against circular reasoning is an independent, argumentative baseline

Independent does not mean produced without AI. It means that the basis from which an argument is evaluated should not be derived entirely from the same AI-generated argument, or stay locked within the same system that produced it. The person should still be able to construct a provisional line of reasoning – for example:

- What is the claim?
- Which evidence matters?
- How does it support the conclusion?
- What assumptions or objections remain?

This baseline may be developed through direct engagement with sources, established research tools or even a second AI system. Its purpose is simply to provide a genuinely separate basis for comparison. Without it, AI-mediated evaluation can become circular: a system generates an argument, proposes objections, assesses these, revises the argument, and explains why the new version is stronger. While each operation may appear rigorous in isolation, the evaluation remains epistemically circular, confined within a single AI-generated frame.

Disagreement is useful where this becomes productive. Suppose a researcher works out a provisional interpretation from the literature while an AI system proposes another. The divergence creates something for her to evaluate. Did the system find evidence the researcher missed? Does one reading depend on a hidden assumption? Is one inference stronger? Has the AI exposed a real weakness, or reached a conclusion the evidence does not support? The comparison only becomes useful when she has enough of a position of her own to understand why the two arguments differ. A second AI system can widen this comparison by supplying another representation or reasoning path, but real verification still requires assessment against the relevant sources, evidence, and reasoning. This is where discovery and verification separate most clearly: AI can identify possibilities; verification establishes whether there are adequate grounds for accepting them. Human argumentative competence is what lets you distinguish assistance that strengthens reasoning; assistance that improves presentation; and assistance that slips in an unsupported conclusion.

Do not let the argument you are evaluating become the only framework you have for evaluating it.

7.4. Regulating: should I verify, change strategy, reclaim the task or stop using AI?

Planning sets the strategy. *Monitoring* shows what is happening to it. *Evaluating* provides grounds for judgement. *Regulating* is where that judgement changes the work.

This matters because real authorship involves more than starting a task or approving its final output. In shared work, authorship means being able to act on reasons: to accept or reject a contribution, seek further evidence, change strategy, reallocate a task or take one back. Research on meaningful human control offers a useful parallel: responsibility should be matched by sufficient ability and authority to exercise control (Cavalcante Siebert et al., 2023). Formal authority becomes real when the person has the understanding and practical capacity required to intervene. For the Research Loop, the concern is epistemic. A researcher remains meaningfully responsible for a judgement only while she retains enough understanding and capability to alter the process it came from.

Shared work can still involve substantive AI contributions

AI can reshape the available options, introduce alternatives, and reorganise parts of a workflow while its work remains aligned with human needs. Josifović and Noller (2026) describe a related condition as 'teleological coherence', where a system's sub-tasks remain aligned with what the person is trying to do. AI can pursue operational sub-goals: retrieve material, classify, identify patterns, and propose structures, while the person keeps steering the relationship for the epistemic purpose of the inquiry. Regulation becomes necessary when a helpful sub-task starts to distort the overarching purpose: when automated sorting narrows the search too early, when smooth AI summaries hide genuine disagreement, or when automation removes the core interpretive work the exercise was meant to build.

Delegation and capability can coexist.

An example: when efficiency overrides critical reading

A student uses an AI tool to group 50 literature sources by theme. It clusters them neatly into three perspectives, and omits two outlier papers that challenge the consensus. If the student accepts the automated summary without checking the margins, efficiency has quietly replaced critical interpretation. A useful idea here is that a capability is a real freedom: having the capacity and means to do something does not require exercising it on every occasion (Nericcio & Jenkins, 2026). So responsible delegation can include extensive AI participation while preserving the capacity needed to understand, reassess or reclaim an operation when needed. The boundary is contextual. A researcher may delegate extensive searching or preliminary classification while retaining direct authority over interpretation. In another task, interpretation itself may be AI-assisted, provided the researcher retains enough domain knowledge and argumentative competence to evaluate and revise the resulting judgement. DigComp 3.0 approaches the same issue through human oversight, autonomy, and human-centred use of digital technologies (Cosgrove & Cachia, 2025). The Research Loop adds a practical test: meaningful oversight must be capable of producing a change in what happens next.

In the Loop's terms, control breaks down the moment human judgement can no longer produce meaningful intervention in the workflow. Two warning signs matter. *Capability loss* is when repeated delegation reduces a capacity the activity still requires. *Loss of defensibility* is when a consequential choice can no longer be reconstructed and defended as one for which the person has adequate reasons. In both cases, performance may remain high, which is why the problem can stay hidden.

Regulation is what keeps shared work revisable, and responsibility stays real only while human judgement can still change the process.

7.5. From responsibility to educational intervention

The four stages of the metacognitive cycle connect responsibility to capability. Taken together, *Planning*, *Monitoring*, *Evaluating* and *Regulating* show why substantive responsibility depends on capacities exercised *throughout the process;* and not just at the start and at the end. Matthias (2004) named the responsibility gap to describe a problem arising when learning machines become insufficiently predictable or controllable for conventional forms of moral or legal responsibility to apply cleanly. The Research Loop addresses a different problem, but the same relationship between responsibility, knowledge and control is in place. In education and research, formal responsibility is usually easy to assign. The student's name is still on the essay and the researcher's on the article. The harder question is whether that formal attribution is matched by the capabilities needed to explain how a conclusion was reached, assess its grounds and intervene in the process that produced it. Responsible AI use therefore needs the capability to reconstruct consequential choices, give reasons for them, name their limits, and revise them when challenged. Disclosure and final approval acquire epistemic significance when those capabilities remain present.

The educational problem: gain and loss

This has consequences for how education responds to AI. AI use becomes an educational question when it changes the relationship between performance and capability. Detection may still matter in particular assessments, but developing reading, argumentation and judgement requires opportunities to exercise those capacities directly. AI also offers genuine gains. It can accelerate search, drafting, explanation and problem solving, and students see the same technologies being adopted in research and professional environments.

Education therefore needs a language that recognises the gain while making the possible loss visible. A loss may concern not only what a person remembers, but what they can still do with what they read and know. An AI-generated explanation may solve an immediate problem while removing an occasion to retrieve or reconstruct the concept. An automatically produced argument may improve an assignment while replacing practice in organising evidence, premises, and objections. A synthesis may save hours while leaving the learner less familiar with the sources on which its conclusions depend. A single use tells us little about capability change. The more important question is about patterns of delegation that keep removing opportunities to exercise capacities later needed for judgement.

This also reframes AI literacy: it should expand a person's ability to read, inquire, reason, and participate in digitally mediated environments. Greater proficiency with AI has wider educational value when those underlying capacities grow with it. Some cognitive friction is worth keeping. Reading an unfamiliar text, struggling to formulate an interpretation, meeting an objection, and trying to explain a difficult point can be the work through which the capability develops.

Institutions are trying things in this direction, moving away from treating AI mainly as a detection problem. The University of Sydney, for example, has developed a two-lane assessment model: some settings establish what students can demonstrate under controlled conditions, while others deliberately teach and assess AI-supported practice. Reported approaches elsewhere include oral defences and audits in which students explain submitted work and demonstrate understanding (Schulzke, 2026). These experiments have a common objective: to preserve evidence of human competence while teaching students to work with AI. MIT's Ad Hoc Committee (2026) places the same problem at the level of educational design, advising institutions to revisit what students need to learn, reconsider how they learn it, adapt assessment and policy, and build processes that can change as the technologies evolve.

A durable approach therefore begins by identifying which capabilities a qualification is meant to certify, creating credible opportunities to exercise and demonstrate them, and teaching AI-supported work openly rather than leaving it to informal experiment. Policy can follow the same logic, weighing access, adoption, and operational skill alongside the capacity to verify

information, interpret sources, develop arguments, recognise uncertainty, and defend consequential judgements. The practical goal is to preserve the possibility of judgement while educational practices change.

A learner should be able to recognise both what AI enables and what repeated delegation may make harder to do when that capability is needed later.

Read, Argue, Defend is the pedagogical translation of this framework.

It builds educational settings where AI assistance stays available while three activities stay visible and exercisable.

Read means returning from AI-mediated summaries to the primary texts, sources, data, and evidence on which consequential claims depend.

Argue means constructing and comparing lines of reasoning, so that an AI-generated argument remains one contribution to the problem rather than the sole framework through which it is understood.

Defend means explaining the choices behind a claim, answering to objections, and revising a position when necessary.

The method tests whether the learner can understand, argue for and take epistemic responsibility for the result of an AI-mediated workflow[3]. It turns responsible AI use from an abstract ethical principle into something learners can practise and demonstrate.

Table 4 brings together the practical implications of the four-stage cycle for individuals, educators, institutions, and policymakers. The stages remain interconnected: monitoring and evaluation may require an earlier delegation decision to be revised.

Stage	Individuals	Educators	Institutions	Policymakers
Planning	Define the purpose of the activity and decide what to delegate before convenience determines the division of work.	Let learning objectives determine the role of AI. Explain which uses are permitted, prohibited, or required, and why.	Develop adaptable AI-use policies that connect task allocation to the purposes of research and learning and the capacities they require.	Support frameworks adaptable to different fields, populations, and settings, with attention to the human capacities each activity requires.
Monitoring	Use checkpoints at consequential stages to track what has been delegated, what has been checked, and whether understanding remains sufficient.	Make selected parts of the process visible through source trails, argument maps, comparisons, and oral explanation, without requiring surveillance of every interaction.	Review how AI use affects learning, assessment, research practices, and the capacities the institution aims to develop.	Monitor changes in people's capacity to search, verify, interpret, and justify information alongside adoption and compliance statistics.
Evaluating	Assess claims against relevant sources, evidence,	Ask learners to compare arguments,	Evaluate whether assessment and review practices	Include verification, interpretation, and

[3] The full sequence of exercises, facilitation model and assessment design is developed separately in the *Read, Argue, Defend (RAD)* workshop document.

Stage	Individuals	Educators	Institutions	Policymakers
	and alternative reasoning. Maintain a basis for evaluation beyond the generated argument itself.	examine supporting evidence, and justify consequential choices. Assess understanding alongside the quality of the final output.	establish substantive understanding and defensibility, alongside disclosure and procedural compliance.	reasoned judgement among the outcomes used to assess AI literacy initiatives.
Regulating	Return to sources, seek expertise, change strategy, reclaim an operation, or stop using AI when the current arrangement undermines judgement.	Design activities in which learners identify problems in AI outputs, choose a response, and explain their reasons.	Revise assessment, teaching practices, and AI policies when monitoring and evaluation show that they no longer serve their purposes.	Ensure that people retain the capability, information, and authority needed to intervene in AI-mediated processes.

Table 4: Applying the Metacognitive Cycle across Stakeholder Groups

Conclusion

Preserving the Conditions of Critical Judgement

The Research Loop names a risk in knowledge production: representations of sources, evidence, and arguments can replace the human engagement needed to assess them, then circulate as the basis for further work. Generative AI increases the speed and scale at which these transitions can occur. At the individual level, convincing performance may conceal reduced opportunities to exercise understanding and judgement. At the systemic level, repeated reliance on mediated representations may make the evidential and argumentative grounds of shared knowledge harder to recover. The report's contribution is to connect these levels through the sequence of mediation, substitution, reproduction, and recursion. No single use of a summary establishes a Loop; the concern is how successive uses can reinforce one another.

The conceptual boundary is between enhancement and substitution. AI enhances inquiry when it extends what people can accomplish while preserving the capacities needed to understand, evaluate, and defend the work. Substitution occurs when assistance displaces the engagement on which those capacities depend. This boundary cannot be fixed once for every task: an operation that usefully supports an experienced researcher may remove essential practice for a learner. Extensive automation can remain compatible with epistemic accountability, provided people retain sufficient understanding and authority to assess consequential choices and revise the process. The criterion is whether the division of work preserves or strengthens epistemic value, rather than merely improving immediate performance.

AI literacy is the evolving set of cognitive, technical, and social capacities through which a person understands AI mediation, judges outputs and limits, and exercises situated judgement over whether, when and how to delegate tasks, while retaining the ability to verify, explain and defend the judgements made with its help.
This definition has practical consequences. For educators, assessment must make learners' understanding and reasoning visible alongside the quality of AI-assisted submissions, while providing opportunities to develop both independent competence and responsible AI use. Institutions must give people the time, access, expertise, and authority needed to exercise the responsibility formally assigned to them. They must also examine whether incentives for speed, volume, and standardisation are encouraging research to adapt to machine capabilities at the expense of its epistemic purposes. Policymakers should assess AI literacy initiatives through their effects on verification, judgement, and the ability to intervene, alongside access, adoption, and operational proficiency. These implications extend to citizens encountering information through increasingly mediated environments.

The metacognitive cycle developed in Chapter 7 provides a practical structure for these interventions. Its value lies in keeping delegation open to scrutiny and revision as systems, tasks, and users' understanding change. Institutions should be able to ask not only whether AI use was permitted and disclosed, but whether the person responsible can explain consequential choices, recognise insufficient grounds, and change the work accordingly. Responsibility becomes substantive when those capacities can be exercised.
The report's limits are equally important. It offers a conceptual diagnosis supported by a critical review, not a demonstration of the complete recursive mechanism across settings. Much of the empirical evidence concerns recent systems, specific tasks, and limited observation periods. The twenty-study scan in §3.4 is exploratory and cannot establish the prevalence of measurement gaps across the literature. The extension of the Research Loop beyond academia, discussed in §1.7, remains a theoretical proposition requiring empirical

investigation. Nor does a successful AI-assisted output establish either lasting capability gains or capability loss.

Future research should therefore measure unaided competence alongside assisted performance, including retention, transfer, and the ability to reconstruct and defend arguments over time. It should also investigate how mediated claims circulate and whether cumulative delegation changes institutional standards of research quality. Read, Argue, Defend offers a first practical setting in which the proposed response can be tested: whether returning to sources, constructing arguments, and defending consequential choices helps preserve the capacities identified here. Such evaluation must remain open to evidence of benefits, limitations, and differences between fields and learners

.

Ultimately, Breaking the Research Loop means expanding what people can accomplish with AI without narrowing what they remain able to understand, judge, and defend.

Glossary

Access vs availability

Being able to reach a source is not the same as its being genuinely usable. A source can be findable but paywalled, or reachable but restricted in who may consult or reuse it.

Adaptive human agency

The ability to revise what is delegated to AI as its capabilities and the context change, while keeping stable principles for deciding when delegation is appropriate.

Affordance

A possibility for action that arises from the relationship between a technology, a user, and the context of use, rather than simply a built-in function.

Agency

The capacity to act and produce effects. When applied to AI in this report, the term describes what a system can do within a process; it does not imply consciousness, intention, or moral responsibility.

Agentic AI

AI systems that can plan and carry out sequences of actions towards a goal with less continuous human direction than ordinary assistive tools. In this chapter, the term bears on which executive and monitoring capacities should stay under human control.

AI agency

The operational ability of an AI system to propose, organise or execute actions that shape how a process develops.

AI literacy

The capacity to understand what the AI system is doing in relation to the task, judge its outputs and limitations, recognise the boundaries of one's own knowledge and adjust reliance accordingly.

AI support / AI substitution

Two ways AI can enter a task. AI support helps people do the work while leaving them able to understand and judge it. AI substitution is when the system performs the very operations a person would otherwise need to do in order to build, show, or keep the relevant competence.

AI-compatible forms of research

Forms of inquiry that are especially easy for AI systems to read, break down, optimise, and evaluate.

AI-enabled research inflation

A rise in the volume of research output that AI makes quicker or easier to produce. It matters for the Research Loop when growing output feeds further cycles of mediated production without a matching gain in epistemic value.

AI-mediated environment

An information, learning or research setting in which AI systems help retrieve, select, summarise, generate, or organise the material people work with.

Algorithm appreciation

The tendency, in some situations, to give more weight to advice when people believe it comes from an algorithm rather than from another person.

Argumentative delegation

Handing over to an AI system not only supporting tasks such as searching or summarising, but the core reasoning work: selecting and weighing evidence, making connections, anticipating objections, and drawing conclusions.

Artificial intelligence (AI)

A family of computational techniques through which machine-based systems infer, from the data they receive, how to generate outputs such as predictions, classifications, recommendations, content, or decisions, operating with varying degrees of autonomy. This report adopts the sense set out in the EU AI Act (Regulation (EU) 2024/1689, Article 3), which itself follows the OECD definition. It uses the term chiefly for systems that now intervene in the searching, synthesising, and writing stages of academic research. The label denotes a set of statistical methods rather than a form of understanding, and the report's argument depends on keeping that distinction in view.

Artificial metacognition

The functional capacity of an AI system to monitor aspects of its own processing and/or regulate computational resources. It should not be treated as equivalent to human metacognition or as evidence that an AI system has reliable knowledge of its own epistemic limits.

Automation bias / over-reliance

Giving too much weight to an automated output or recommendation, including when there is reason to check it. It can feed the Research Loop, but it describes reliance on automation rather than the recursive production of knowledge.

Boundary condition (enhancement vs substitution)

The line between AI that widens your contact with the evidence (support) and AI whose output you treat as a replacement for the source or for your own judgement (substitution).

Calibrated reliance

Adjusting how much you rely on an AI output to the task, the stakes and the system’s known limits, rather than trusting or distrusting it across the board.

Capability

A real freedom to perform or exercise a capacity. A person can retain a capability even when choosing not to exercise it on a particular occasion. In this report, this distinction helps separate responsible delegation from the loss of the capacity to reclaim or judge what has been delegated.

Co-agency

A set-up in which a person and an AI both shape an activity, without necessarily having the same kind of agency or responsibility.

Cognitive offloading

Using an external aid, such as a note, calculator, search engine, or AI system, to carry part of a mental task. It can be useful; it becomes relevant to the Research Loop when people hand over operations they still need to acquire, exercise, or retain for critical judgement.

DigComp 3.0

The European Union's Digital Competence Framework. It sets out the knowledge, skills and attitudes needed for learning, work, and taking part in society, with AI competence integrated across it.

Digital natives / digital immigrants

A popular split between people assumed to be naturally at ease with technology because they grew up with it and those assumed to have adapted later. The report treats it as inadequate, because digital and AI literacy have to be developed, not assumed from exposure.

Diminished argumentative autonomy

A reduced ability to build, weigh and defend an argument, even while coherent arguments or answers remain readily available through AI.

Discovery

The use of AI, search systems, or other resources to identify potentially relevant sources, ideas, connections, hypotheses, or alternatives. Discovery expands what may need to be examined; it does not by itself establish that a claim is reliable.

Distributed and extended cognition

The idea that thinking is not confined to the brain alone but can be shared across a person, their body and the tools and surroundings they use, such as counting on your fingers or working things out on paper.

Epistemic

To do with knowledge: how it is formed, justified, assessed, and treated as reliable. The report uses it in phrases such as epistemic quality, epistemic responsibility, and epistemic authorship.

Epistemic accountability

Responsibility for keeping the conditions under which a knowledge claim can be understood, verified, justified, and defended by the person who adopts it.

Epistemic authorship

The capacity to understand, reconstruct, justify, and independently defend the interpretive, evidential, and argumentative choices behind an output. It differs from formal authorship, which is about whose name is on the work and who is officially responsible for it.

Epistemic contribution

What a piece of research actually adds to knowledge, for example new evidence, a meaningful new question, a challenged assumption, a better method, or a clearer problem. It is not the same as how much gets published.

Epistemic dependency

A condition in which a person leans more and more on outside systems to get, interpret or organise knowledge they are less able to assess or reconstruct themselves.

Epistemic dysregulation

A condition in an AI-mediated workflow in which the human agent can no longer translate judgement into meaningful intervention, or can no longer reconstruct and defend a consequential choice for which they remain responsible.

Epistemic necessity

A defensible reason for producing a piece of research because it addresses a real gap, question, uncertainty or need. It does not require immediate practical use.

Epistemic safeguards

Practices and rules of control that keep a person able to understand, verify and defend how knowledge was produced when AI is involved.

Epistemic value

The value of knowledge beyond successful completion of the immediate task. In this report, knowledge has further epistemic value when it can support new judgements and inferences, open possibilities for inquiry, develop human capacities or contribute to subsequent knowledge production. This is a functional definition adopted for the Research Loop.

Evaluating

The stage in which claims, sources and arguments produced or mediated by AI are assessed against relevant evidence, alternative reasoning, and the agent's own understanding, in order to decide how far reliance is warranted.

FAIR principles

A way of separating four things that are easily confused: whether a source can be found, reached, combined with other material, and reused. A source can do well on one and badly on another.

Flattening

The risk that meaningful disciplinary differences are reduced as research is increasingly organised around common, easily specified and AI-compatible tasks and goals.

Generative AI

A category of AI systems designed to produce new content, such as text, images, audio, code, or video, in response to a prompt. These systems do not classify or retrieve existing material. They learn statistical patterns from large training datasets and then generate outputs that resemble those patterns. Large Language Models are the most prominent text-based form of generative AI. The EU AI Act refers to the underlying models as "general-purpose AI models" and places transparency obligations on content they generate. Because generative AI produces material that reads as authoritative and original without any built-in link to verifiable sources, it blurs the line between synthesis and invention. In academic research, this makes it the class of technology most directly implicated in the Research Loop.

Heuristics and biases

Mental shortcuts we use to judge things quickly. They are often useful, but can produce systematic, predictable errors.

Homogenisation feedback loop

A recursive process in which repeated reliance on similar sources, concepts or readings narrows the range of material in the shared knowledge base. It is close to the Research Loop, but focuses specifically on convergence and loss of conceptual diversity.

Human agency

The ability to shape a process while keeping authority over the decisions that matter and responsibility for the conclusions that follow.

Human-in-the-loop

An approach in which a person stays meaningfully involved at the important stages of an AI-assisted process, with the understanding and authority to evaluate and step in, rather than just sign off.

Illusion of understanding

In this report, one form of the illusion occurs when a person mistakes a fluent, coherent output produced with AI, or the feeling that a task was easy, for a real grasp of the reasoning behind it.

Independent argumentative baseline

A provisional line of reasoning that is not derived entirely from the AI-generated argument being evaluated or confined to the same system that produced it. It provides a basis for comparing claims, evidence, assumptions, and conclusions. It may be developed with other tools or AI assistance and does not imply reasoning without technology.

Informed trust / informed distrust

A decision to rely, or not to rely, on an AI output, reached by weighing it against your own knowledge and the system's known limits, rather than a fixed attitude towards the system.

Infrastructural mediation

The technical systems people use to find and reach information, such as search engines, databases, and digital archives. It is not part of the Research Loop as long as users still engage with the material needed to assess the evidence or argument.

Intention offloading

Using external aids such as calendars, reminders, or well-placed objects to help remember and carry out things we plan to do later.

Judgement literacy

The regulatory capacity to turn metacognitive monitoring into context-sensitive decisions about verification, reliance, and delegation. It is not treated in this report as an additional discrete skill, but as a capacity that regulates the exercise of other AI-literacy competencies.

Judgement-bearing decision point

A point in an AI-mediated workflow where a decision affects the direction of inquiry, the weight assigned to evidence, the interpretation adopted, the evaluation of an argument or the conclusion to be defended. AI may assist these decisions, but substantive epistemic responsibility requires the human agent to retain authority to adopt, revise or reject the judgement.

Large Language Model

A type of AI model trained on very large bodies of text to predict the most probable next unit of language (a "token")given the text that precedes it. Through this training, LLMs become able to produce fluent summaries, answers, and literature syntheses, and they underpin widely used tools such as ChatGPT, Claude and Gemini. Their outputs are generated by statistical likelihood. They are not retrieved from, or checked against, the sources they appear to describe. As a result, LLMs can produce plausible but inaccurate claims and citations (commonly termed "hallucinations"), and they can reproduce the biases and omissions of their training data. Their growing use as a substitute for engagement with primary literature is the principal mechanism of the Research Loop that this report describes.

Matthew effect

The tendency for already well-recognised researchers, publications, or sources to attract still more attention and recognition, building cumulative advantage over time.

Meaningful human control

A form of human control in which responsibility is matched by sufficient ability and authority to intervene in the operation of an AI system, including correcting behaviour, changing goals, or reallocating tasks.

Mediation, substitution, reproduction, and recursion

The four stages of the Research Loop. Mediation is meeting relevant material through a representation, such as a summary or synthesis; on its own it is not a Research Loop. Substitution begins when that representation takes the place of the work needed to assess the underlying material. Reproduction is using the mediated material to produce a new output. Recursion is when that output is reused, in turn, as material for further learning, reasoning or production.

Metacognition

The capacity to monitor and steer your own thinking: judging how a task is going, how confident you should be, how much effort it requires and whether to change strategy or rely on an external tool.

Metacognitive calibration

How well a person's sense of their own effort or performance matches what is actually happening. Poor calibration means, for example, feeling that AI has saved time when it has not.

Metacognitive debt

The longer-term cost that can build up when leaning on outside support again and again leaves fewer chances to monitor, steer and reflect on your own learning and reasoning.

Metacognitive judgement

A judgement about your own process, such as how confident you are, on what basis, and whether to change strategy, as opposed to a judgement about an object.

Metaknowledge

What a person knows about their own knowledge and about a system: what they know, what they know they do not know, and rules for when to delegate.

Model collapse

A kind of model degradation that can happen when AI systems are trained repeatedly on AI-generated rather than human data. It is recursive, but unlike the Research Loop it needs the model to be retrained.

Monitoring

The metacognitive activity of tracking whether an AI-mediated workflow continues to support adequate human understanding and whether the initial distribution of work remains appropriate as the activity develops.

Non-delegable judgement task

A judgement-bearing activity that must remain sufficiently under human control for the agent to understand, evaluate and defend the consequential decisions on which an inquiry or conclusion depends. What is non-delegable varies with the purpose and context of the task.

Performance culture

A condition in which individuals are encouraged or expected to continually perform, measure and display aspects of themselves in response to social, institutional, and technological expectations.

Performance of knowledge

Producing something that has the outward features of knowledge without an equal ability to trace, reconstruct, evaluate, and defend the steps behind it, or to show that the relevant learning has taken place.

Performance value

The local value of a result for the task at hand, for example, when it is produced faster, more accurately or more coherently.

Planning

The stage in which the goal, constraints and initial strategy of an AI-mediated activity are established, including which operations AI may support or perform and which human capabilities need to remain actively available.

Plausible misinformation

False or misleading information that looks coherent, credible, or well-formed enough that people have little reason to question or check it.

Processing fluency

The ease with which something can be read and understood. In some conditions, people may mistake that ease for a sign that the information is accurate.

Public epistemic autonomy

Citizens' capacity to seek, assess, verify, and interpret information independently enough to form and revise their own judgements. It does not mean avoiding AI, but it does mean keeping the capacities needed to evaluate AI-mediated information.

Publication inflation

A rise in the volume of formally publishable, technically valid research without a matching rise in epistemic value. In this report it is treated as a possible systemic effect, not as proof that AI-assisted research is inherently low quality.

Recursive / recursivity

A process in which outputs become inputs for further activity. In the Research Loop, mediated claims, summaries, or interpretations are reused to produce further mediated outputs.

Recursive knowledge

Knowledge that re-enters the cycle it came from: a mediated output is reused as a source or input for further work, round after round.

Regulating

The process through which metacognitive judgement produces a change in an AI-mediated workflow, for example by verifying, changing strategy, altering delegation, reclaiming an operation, or discontinuing AI use.

Research Loop

A recursive pattern in which mediated stand-ins for sources, evidence or arguments start to replace the human engagement needed to assess them. These stand-ins are then reused to produce further claims, texts, or interpretations, which may themselves become inputs for later learning, reasoning, and research.

Responsibility gap

Situations in which the behaviour of learning machines makes traditional attribution of moral or legal responsibility difficult because the relevant human agent lacks sufficient predictability

or control. The Research Loop uses this as an analogy rather than treating epistemic responsibility as the same problem.

Self-regulated learning

An approach to learning in which people set their own goals and monitor and adjust their thoughts, strategies, and effort in order to reach them.

Semantic capital

The resources through which people make sense of their experience and give meaning to the world. The report asks whether more available information always means more human understanding.

Skill decay

The loss of a skill that was once acquired, after a period of not using it. How much and how fast this happens depends on the skill and the conditions.

Source displacement

When a source stays in the bibliography but your real engagement is with a stand-in for it, a summary, review, or synthesis, so that later work builds on the stand-in rather than the original.

Speedup illusion

The mistaken sense that working with AI has saved time, because it reduced how effortful a task felt, even when the time actually spent was much the same.

Sycophancy

The tendency of some AI systems to tell users what they seem to want to hear, agreeing with them or reinforcing their existing views rather than challenging them.

Task performance and independent human competence

Task performance is the quality, accuracy, speed, or quantity of an output. Independent human competence is what a person can actually do, and draw on, without the tool. A good result from a person working with AI does not by itself show what the person could do alone.

Teleological coherence

The condition in which AI contributions remain organised within a purpose-governed domain oriented towards human ends. In the Research Loop, it helps explain how co-agency can allow substantive AI contributions while keeping the epistemic purpose of the activity subject to human regulation.

Validity reasoning

Judging whether a result or claim actually holds up, rather than accepting it because it looks correct or because it came from a tool.

Verification

The process of establishing whether there are adequate grounds for accepting a claim by examining the relevant evidence, provenance, sources, and reasoning. Verification is distinct from discovering potentially relevant information.

References

AI scientists are changing research — institutions, funders and publishers must respond. (2026). Nature, 651(8107), 853–854. https://doi.org/10.1038/d41586-026-00934-w

Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S., & Ginis, V. (2025). Large language models reflect human citation patterns with a heightened citation bias. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 6829–6864). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.381

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/10.1016/j.caeo.2024.100173

Association of College and Research Libraries. (2016). Framework for information literacy for higher education. American Library Association. https://www.ala.org/acrl/standards/ilframework

Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. Social Sciences & Humanities Open, 12, 102287. https://doi.org/10.1016/j.ssaho.2025.102287

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. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates, Inc.

Bishop, D. V. M. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. Quarterly Journal of Experimental Psychology, 73(1), 1–19. https://doi.org/10.1177/1747021819886519

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). The crowdless future? Generative AI and creative problem-solving. Organization Science, 35(5), 1589–1607. https://doi.org/10.1287/orsc.2023.18430

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Byrne, J. A., Park, Y., Richardson, R. A. K., Pathmendra, P., Sun, M., & Stoeger, T. (2022). Protection of the human gene research literature from contract cheating organizations known as research paper mills. Nucleic Acids Research, 50(21), 12058–12070. https://doi.org/10.1093/nar/gkac1139

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. Nature Human Behaviour, 6, 1112–1125. https://doi.org/10.1038/s41562-022-01332-8

Campillo-Ferrer, J. M., López-García, A., & Miralles-Sánchez, P. (2025). Student perceptions of the use of Gen-AI in a higher education program in Spain. Digital, 5(3), Article 29. https://doi.org/10.3390/digital5030029

Cash, T. N., Kelly, M. O., Macnamara, B. N., & Risko, E. F. (2026). Is AI making us stupid? Trends in Cognitive Sciences, 30(8), 673–675. https://doi.org/10.1016/j.tics.2026.06.004

Castellacci, F., Ciarli, T., Gao, Y., Marino, M., Marzi, G., Riccaboni, M., Savona, M., & Vannuccini, S. (2026). Generative artificial intelligence in scientific research: Individual benefits, collective risks, and a framework for responsible research with AI [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.24879

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. AI and Ethics, 3, 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chapekis, A., & Lieb, A. (2025, July 22). Google users are less likely to click on links when an AI summary appears in the results. Pew Research Center. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/

Chen, A., Lian, J., Kuang, X., & Jia, J. (2026). Can theory-driven learning analytics dashboard enhance human-AI collaboration in writing learning? Insights from an empirical experiment. The Internet and Higher Education, 68, 101054. https://doi.org/10.1016/j.iheduc.2025.101054

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

Chirikov, I. (2026). Artificial intelligence and grade inflation (CSHE Higher Education Working Paper Series No. 26-3). Center for Studies in Higher Education, University of California, Berkeley. https://escholarship.org/uc/item/80x8d3qd

Chun, J., & Elkins, K. (2023). The crisis of artificial intelligence: A new digital humanities curriculum for human-centred AI. International Journal of Humanities and Arts Computing, 17(2), 147–167. https://doi.org/10.3366/ijhac.2023.0310

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., O'Neill, H., & Jones, M. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods, 16(4), 601–619. https://doi.org/10.1017/rsm.2025.16

Colville, S., & Ostern, N. (2026). Making informed trust judgements: The role of metacognition in ChatGPT interactions. Behaviour & Information Technology. Advance online publication. https://doi.org/10.1080/0144929X.2026.2651114

Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European Digital Competence Framework: Fifth edition. Publications Office of the European Union. https://doi.org/10.2760/0001149

Cox, M. T. (2005). Metacognition in computation: A selected research review. Artificial Intelligence, 169(2), 104–141. https://doi.org/10.1016/j.artint.2005.10.009

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Floridi, L. (2018). Semantic capital: Its nature, value, and curation. Philosophy & Technology, 31(4), 481–497. https://doi.org/10.1007/s13347-018-0335-1

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., & Tsai, P.-C. (2023). Outsourcing memory to external tools: A review of 'intention offloading'. Psychonomic Bulletin & Review, 30(1), 60–76. https://doi.org/10.3758/s13423-022-02139-4

Grech, A. (2025). Beyond digital natives: Critical digital literacies for the AI generation: A primer. 3CL Foundation. https://www.3cl.org/resources/

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. BMJ, 339, b2680. https://doi.org/10.1136/bmj.b2680

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Guo, Y., & Ye, Q. (2026). Meta-cognitive insights into cognitive offloading: Mechanisms, interventions, and educational implications. Humanities and Social Sciences Communications, 13, Article 772. https://doi.org/10.1057/s41599-026-06621-5

Han, J., & Ren, R. (2025). Why unequal AI access enhances team productivity: The mediating role of interaction processes and cognitive diversity. Frontiers in Psychology, 16, 1636906. https://doi.org/10.3389/fpsyg.2025.1636906

Hanushek, E. A., Kinne, L., Witthöft, F., & Woessmann, L. (2025). Age and cognitive skills: Use it or lose it. Science Advances, 11(10), eads1560. https://doi.org/10.1126/sciadv.ads1560

Hesmondhalgh, D., & Pratt, A. C. (2005). Cultural industries and cultural policy. International Journal of Cultural Policy, 11(1), 1–13. https://doi.org/10.1080/10286630500067598

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429–431. https://doi.org/10.1038/520429a

Holter, S., Moruzzi, C., & El-Assady, M. (2026). Toward agency in human-AI collaboration. IEEE Computer Graphics and Applications, 46(1), 13–25. https://doi.org/10.1109/MCG.2025.3623892

Ingelfinger, F. J. (1976). Seduction by citation. The New England Journal of Medicine, 295(19), 1075–1076. https://doi.org/10.1056/NEJM197611042951911

Jiang, Y., & Chen, H. (2026). AI-based cognitive offloading by students in higher education in China. Computers in Human Behavior: Artificial Humans, 8, 100327. https://doi.org/10.1016/j.chbah.2026.100327

Josifović, S., & Noller, J. (2026). Agency and alignment: Toward a normative architecture for human–AI interaction. AI & Society, 41(6), 5561–5571. https://doi.org/10.1007/s00146-026-02950-w

Kasparov, G. (2010, February 11). The chess master and the computer. *The New York Review of Books*.

Kemman, M., Kleppe, M., & Scagliola, S. (2014). Just Google it: Digital research practices of humanities scholars. In C. Mills, M. Pidd, & E. Ward (Eds.), Proceedings of the Digital Humanities Congress 2012. HRI Online Publications.

Kirschner, P. A., & De Bruyckere, P. (2017). The myths of the digital native and the multitasker. Teaching and Teacher Education, 67, 135–142. https://doi.org/10.1016/j.tate.2017.06.001

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI, an experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160, 108352. https://doi.org/10.1016/j.chb.2024.108352

Krakowski, S. (2025). Human-AI agency in the age of generative AI. Information and Organization, 35(1), 100560. https://doi.org/10.1016/j.infoandorg.2025.100560

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. Nature Human Behaviour, 8, 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (Article 1121, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778

Lehmann, M., Cornelius, P. B., & Sting, F. J. (2025). AI meets the classroom: When do large language models harm learning? [Working paper]. SSRN. https://doi.org/10.2139/ssrn.4941259

Li, S., Liu, J., & Dong, Q. (2025). Generative artificial intelligence-supported programming education: Effects on learning performance, self-efficacy and processes. Australasian Journal of Educational Technology, 41(3), 1–25. https://doi.org/10.14742/ajet.9932

Linton, D. L., Pangle, W. M., Wyatt, K. H., Powell, K. N., & Sherwood, R. E. (2014). Identifying Key Features of Effective Active Learning: The Effects of Writing and Peer Discussion. CBE—Life Sciences Education, 13(3), 469–477. https://doi.org/10.1187/cbe.13-12-0242

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., & Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (pp. 6107–6117). Association for Computing Machinery. https://doi.org/10.1145/3711896.3736569

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., & Clune, J. (2026). Towards end-to-end automation of AI research. Nature, 651(8107), 914–919. https://doi.org/10.1038/s41586-026-10265-5

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences, 97(8), 4398–4403. https://doi.org/10.1073/pnas.070039597

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mei, P., Brewis, D. N., Nwaiwu, F., Sumanathilaka, D., Alva-Manchego, F., & Demaree-Cotton, J. (2025). If ChatGPT can do it, where is my creativity? Generative AI boosts performance but diminishes experience in creative writing. Computers in Human Behavior: Artificial Humans, 4, 100140. https://doi.org/10.1016/j.chbah.2025.100140

Merton, R. K. (1968). The Matthew effect in science. Science, 159(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627, 49–58. https://doi.org/10.1038/s41586-024-07146-0

MIT Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training. (2026). Report of MIT's Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training. Massachusetts Institute of Technology. https://web.mit.edu/files/AI-Committee-Final-Report.pdf

Mynlieff, M., Manogaran, A. L., St. Maurice, M., & Eddinger, T. J. (2014). Writing Assignments with a Metacognitive Component Enhance Learning in a Large Introductory Biology Course. CBE—Life Sciences Education, 13(2), 311–321. https://doi.org/10.1187/cbe.13-05-0097

Mogull, S. A. (2017). Accuracy of cited "facts" in medical research articles: A review of study methodology and recalculation of quotation error rate. PLOS ONE, 12(9), e0184727. https://doi.org/10.1371/journal.pone.0184727

Morrison, A. B., & Richmond, L. L. (2020). Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. Cognitive Research: Principles and Implications, 5, Article 1. https://doi.org/10.1186/s41235-019-0201-4

Nericcio, L., & Jenkins, R. (2026). Justifying and evaluating automation: A general method for quantifying the impacts of machine learning on human flourishing. AI and Ethics, 6, Article 346. https://doi.org/10.1007/s43681-026-01118-y

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

OECD & European Commission. (2026). Empowering learners for the age of AI: An AI literacy framework for primary and secondary education. OECD Publishing. https://doi.org/10.1787/65cd27d4-en

OpenAlex. (n.d.). Open access trends by year. Retrieved August 12, 2026, from https://developers.openalex.org/guides/recipe-oa-trends

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Pavlovic, V., Weissgerber, T., Stanisavljevic, D., Pekmezovic, T., Milicevic, O., Milin Lazovic, J., Cirkovic, A., Savic, M., Rajovic, N., Piperac, P., Djuric, N., Madzarevic, P., Dimitrijevic, A., Randjelovic, S., Nestorovic, E., Akinyombo, R., Pavlovic, A., Ghamrawi, R., Garovic, V., & Milic, N. (2021). How accurate are citations of frequently cited papers in biomedical literature? Clinical Science, 135(5), 671–681. https://doi.org/10.1042/CS20201573

Pearson, J., Dror, I. E., Jayes, E., Whordley, G.-R., Mason, G., & Nightingale, S. (2026). Examining human reliance on artificial intelligence in decision making. Scientific Reports, 16, Article 5345. https://doi.org/10.1038/s41598-026-34983-y

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. Royal Society Open Science, 12(4), 241776. https://doi.org/10.1098/rsos.241776

Pritchard, D. (2009). Knowledge, understanding and epistemic value. Royal Institute of Philosophy Supplements, 64, 19–43. https://doi.org/10.1017/S1358246109000046

Pritchard, D. (2026). Why technology doesn't normally make you dumber, but agentic AI will. International Journal of Human–Computer Interaction. Advance online publication. https://doi.org/10.1080/10447318.2026.2631678

Radüntz, T. (2020). The effect of planning, strategy learning, and working memory capacity on mental workload. Scientific Reports, 10, Article 7096. https://doi.org/10.1038/s41598-020-63897-6

Rapanta, C., Bhatt, I., Bozkurt, A., Chubb, L. A., Erb, C., Forsler, I., Gravett, K., Koole, M., Lintner, T., Örtegren, A., Petricini, T., Rodgers, B., Webster, J., Xu, X., Falgren Christensen, I.-M., Bonderup Dohn, N., Weilgaard Christensen, L. L., Zeivots, S., & Jandrić, P. (2025).

Critical GenAI literacy: Postdigital configurations. Postdigital Science and Education, 7, 1296–1333. https://doi.org/10.1007/s42438-025-00573-w

Reid, E. (2026, May 19). A new era for AI Search. Google. https://blog.google/products-and-platforms/products/search/search-io-2026/

Retkowsky, J., Hafermalz, E., & Huysman, M. (2024). Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. Business Horizons, 67(5), 511–523. https://doi.org/10.1016/j.bushor.2024.04.009

Reynolds, J. A., Thaiss, C., Katkin, W., & Thompson, R. J. (2012). Writing-to-Learn in Undergraduate Science Education: A Community-Based, Conceptually Driven Approach. CBE—Life Sciences Education, 11(1), 17–25. https://doi.org/10.1187/cbe.11-08-0064

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. Information Fusion, 126, 103599. https://doi.org/10.1016/j.inffus.2025.103599

Scherbakov, D., Hubig, N. C., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. Journal of the American Medical Informatics Association, 32(6), 1071–1086. https://doi.org/10.1093/jamia/ocaf063

Schulzke, E. (2026, August 22). As students embrace AI, universities scramble to keep up. Deseret News. https://www.deseret.com/magazine/2026/08/22/how-ai-is-changing-higher-education/

Shakarian, P. (2026). Toward artificial metacognition. Proceedings of the AAAI Conference on Artificial Intelligence, 40(48), 41000–41005. https://doi.org/10.1609/aaai.v40i48.42135

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y

Spearing, E. R., Gile, C. I., Fogwill, A. L., Prike, T., Swire-Thompson, B., Lewandowsky, S., & Ecker, U. K. H. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. Royal Society Open Science, 12(6), Article 242148. https://doi.org/10.1098/rsos.242148

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. Science Advances, 9(26), eadh1850. https://doi.org/10.1126/sciadv.adh1850

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. Computers in Human Behavior, 160, 108386. https://doi.org/10.1016/j.chb.2024.108386

Stolpe, K., Larsson, A., & Johansson Falck, M. (2026). Discipline-specific AI literacy (DiSAIL): A theoretical framework for situated engagement with generative AI in education. International Journal of Technology and Design Education, 36, 1917–1932. https://doi.org/10.1007/s10798-026-10060-3

Sun, D., Boudouaia, A., Zhu, C., & Li, Y. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. International Journal of Educational Technology in Higher Education, 21, Article 14. https://doi.org/10.1186/s41239-024-00446-5

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Article 680, pp. 1–24). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642902

Tatel, C. E., & Ackerman, P. L. (2025). Procedural skill retention and decay: A meta-analytic review. Psychological Bulletin, 151(6), 696–736. https://doi.org/10.1037/bul0000481

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

UNESCO. (2024). AI competency framework for students. https://doi.org/10.54675/JKJB9835

UNESCO. (2026, May 21). Digitization projects. https://www.unesco.org/en/archives/digitizationprojects

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. Computers & Education, 215, 105031. https://doi.org/10.1016/j.compedu.2024.105031

van Dalen, H. P. (2021). How the publish-or-perish principle divides a science: The case of economists. Scientometrics, 126, 1675–1694. https://doi.org/10.1007/s11192-020-03786-x

Vasconcelos, S., & Marušić, A. (2025). Gen AI and research integrity: Where to now? The integration of generative AI in the research process challenges well-established definitions of research integrity. EMBO Reports, 26(8), 1923–1928. https://doi.org/10.1038/s44319-025-00424-6

Wahn, B., & Schmitz, L. (2024). A bonus task boosts people's willingness to offload cognition to an algorithm. Cognitive Research: Principles and Implications, 9, Article 24. https://doi.org/10.1186/s41235-024-00550-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, Article 160018. https://doi.org/10.1038/sdata.2016.18

Williamson, E. R. (2026). Digitized archives, content providers, and slow scholarship: Why archival researchers should care about digital provenance. Digital Scholarship in the Humanities, 41(3), 1705–1723. https://doi.org/10.1093/llc/fqag062

Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radically collaborative research. Studies in History and Philosophy of Science Part A, 46, 16–23. https://doi.org/10.1016/j.shpsa.2013.11.007

Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. Scientific Reports, 10, 3632. https://doi.org/10.1038/s41598-020-60382-y

Wulff, P., & Kubsch, M. (2025). Learning against the machine: The double edged sword of (Gen)AI in STEM education. International Journal of STEM Education, 12, Article 66. https://doi.org/10.1186/s40594-025-00588-6

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. Nature Human Behaviour, 8, 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Yin, J., Xu, H., Pan, Y., & Hu, Y. (2025). Effects of different AI-driven chatbot feedback on learning outcomes and brain activity. npj Science of Learning, 10, Article 17. https://doi.org/10.1038/s41539-025-00311-8

Ying, Q., Dong, W., & Fabrikant, S. I. (2024). How do in-car navigation aids impair expert navigators' spatial learning ability? Annals of the American Association of Geographers, 114(7), 1483–1504. https://doi.org/10.1080/24694452.2024.2356858

Yu, S., Cheng, M., Jabbar, A., Sucholutsky, I., Collins, K. M., Jurafsky, D., & Hawkins, R. D. (2026). Cognitive offloading and the speedup illusion in human-AI interaction [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.23177

Zhu, Q., Li, X., Dong, Y., Chang, P., & Fan, M. (2026). Not all cognitive offloading is equal: Distinguishing dependent and autonomous offloading to generative AI. Frontiers in Psychology, 17, 1878629. https://doi.org/10.3389/fpsyg.2026.1878629

Zyphur, M. J. (2026). Responsible AI in academic research: A competency framework for research training. Instats Policy Series. https://doi.org/10.61700/t310y23grr