The Methodology Problem

Why Research on AI and Cognition Confounds Technology with Ungoverned Use

Basil C. Puglisi, MPA

Human-AI Collaboration Strategist and AI Governance Consultant

The Research Flaw: Testing Consumption, Not Engagement

Every study claiming that AI use erodes critical thinking quietly shares the same design flaw. They are not measuring governed AI use. They are measuring unstructured prompt in, answer out workflows that ask nothing of the user beyond consumption. By "governance" I mean structured interaction protocols that require users to engage in metacognitive processes, including planning, monitoring, and evaluating, before accepting AI outputs. This definition draws on established educational research showing that such processes drive learning and retention (Flavell, 1979; Schraw & Dennison, 1994).

The Microsoft and Carnegie Mellon survey of 319 knowledge workers collected 936 self reported examples of generative AI use and then analysed how people felt about their own critical thinking and confidence (Lee et al., 2025). Nowhere in that protocol is evidence that participants were required to verify sources, cross check contradictions, compare multiple systems, or document decisions against explicit checkpoints. The study observes cognitive offloading in an environment that never asked for cognitive engagement in the first place.

The MIT Media Lab EEG study took a similar approach (Kosmyna et al., 2025). Participants were divided into three groups for SAT style essay writing: one with an LLM, one with a search engine, and one "brain only." The LLM condition simply instructed people to use ChatGPT for essay drafting. There was no governance structure that required them to interrogate the model's output, justify their reasoning, or provide evidence trails before submission. It is not surprising that the group with the easiest path to copy and paste displayed the lowest neural engagement and the most formulaic writing. The design tested ungoverned LLM use, then inferred conclusions about AI and cognition in general. Note that this study remains a preprint awaiting peer review as of November 2025, with the authors acknowledging limitations including small sample size and context specificity to essay writing.

This is equivalent to asking whether reading makes people smarter by observing a cohort that passively skims text without annotation, questioning, or synthesis, then generalising the result to all reading. The experimental frame confounds technology with methodology. What is actually being studied is a specific behaviour: cognitive offloading into a single unstructured system that never pushes back.

Overlooked Dissent: Structure Determines Outcome

Within the same research ecosystem there is already evidence that structured AI use can enhance, rather than erode, critical thinking. It appears in the footnotes and sidebars rather than in the headline claims.

Microsoft researcher Lev Tankelevitch explicitly notes that AI can "synthesize ideas, enhance reasoning, and encourage critical engagement" when it is used as a thought partner and when users are required to validate outputs rather than accept them at face value (Diaz, 2025). That is a governance statement nested inside a cognitive concern.

The Nigerian after school pilot, supported by the World Bank, reports learning gains equivalent to nearly two years of typical schooling compressed into six weeks (Crawfurd et al., 2025). Those results did not emerge from children left alone with a chatbot. Teachers framed prompts, anchored content to curriculum, and provided context and follow up tasks. The AI acted as tutor within a designed structure, not as a shortcut around effort.

The Chinese EFL intervention study in the European Journal of Education found statistically significant improvements in critical thinking for students who used AI tools in literature classes (Liu & Wang, 2024). Those gains did not come from free form answer generation. They came from structured tasks where AI supported question generation, interactive debates, and guided practice. Scaffolding, not spontaneity, produced the benefit.

Across these cases the pattern is consistent. In the most cited cases, negative outcomes appear in unstructured access that enables users to outsource thinking without friction. Positive outcomes cluster where structure is imposed that forces metacognitive processing. The variable is not AI presence. The variable is governance framework presence. Yet no major study has directly tested this by running the same AI tools under governed and ungoverned conditions side by side.

HAIA-RECCLIN as Missing Experimental Condition

Governance frameworks change the interaction model in a fundamental way. HAIA-RECCLIN (Human-AI Interaction Architecture with Researcher, Editor, Coder, Calculator, Liaison, Ideator, Navigator roles) represents one such framework designed to encode metacognitive checkpoints into AI collaboration workflows.

In the standard research condition, the workflow looks like this:

> *Prompt → AI generates answer → User glances at output and accepts or lightly edits → Task complete*

The user can move from question to answer without ever surfacing their reasoning or interrogating the system's logic. Critical thinking becomes optional effort.

In a HAIA-RECCLIN governed condition, the workflow is categorically different:

> *Prompt with explicit role assignment → AI provides sources, facts, conflicts, confidence scores, and recommendations → User evaluates evidence quality → User examines contradictions across sources or systems → User records an explicit decision and justification → Task complete*

Planning appears in the role selection and prompt design. Monitoring appears in source review and conflict analysis. Evaluation appears in the requirement to justify decisions before completion. These are the exact metacognitive behaviours that educational research associates with growth in critical thinking, now embedded as process requirements rather than as aspirational advice. This structure aligns with Vygotsky's scaffolding theory, where temporary support enables learners to perform beyond their current capability (Vygotsky, 1978). It also reflects Sweller's cognitive load theory, optimizing germane load (learning relevant effort) while reducing extraneous load (confusion from unstructured information) (Sweller, 2010). The checkpoint requirements shift users from Kahneman's System 1 (fast, automatic) thinking toward System 2 (slow, effortful) processing (Kahneman, 2011).

The current research narrative tends to collapse these distinctions into a single conclusion: "AI reduces critical thinking." In reality, the evidence shows that unstructured AI access reduces critical thinking, while structured uses are beginning to show the opposite effect. The missing experimental condition is not a different model. It is a different governance architecture for the same model.

Visualizing the Three Conditions

The following framework illustrates the three experimental conditions that research should compare:

Human Alone	Ungoverned AI	Governed AI
Baseline cognitive engagement	Cognitive offloading; reduced neural engagement	Cognitive enhancement through structured checkpoints
Full metacognitive effort required	Metacognition optional; path of least resistance	Metacognition required by workflow design
Learning through struggle	Learning bypassed; skill atrophy risk	Learning embedded in governance structure

Current research tests only the first two conditions. The third remains the missing experimental variable.

Testable Framework, Not Proven Intervention

At this point the claim is not that HAIA-RECCLIN is already proven to enhance critical thinking. The claim is that it encodes the same metacognitive structures that educational research repeatedly associates with improvement. Planning appears in explicit role assignment and prompt design. Monitoring appears in structured source review and conflict analysis. Evaluation appears in required decision justification before completion.

The logic chain is straightforward. Metacognitive structures tend to enhance learning. HAIA-RECCLIN is a metacognitive structure. Therefore HAIA-RECCLIN should be treated as a testable enhancement framework, not as a rhetorical flourish.

The honest position is that current evidence is inferential rather than direct. The next step is not to celebrate a solution. The next step is to run the experiments that either validate or falsify the prediction. Two important caveats apply. First, governance can fail if implemented perfunctorily. Users who mechanically tick checkpoint boxes without genuine cognitive engagement will not experience enhancement; bad governance is as useless as no governance. Second, governance scaffolds expert

thinking but cannot create expertise from zero. A novice lacking domain knowledge may struggle to evaluate AI outputs even when required to review them. The framework supports those with foundational competence; it does not substitute for that competence.

Multi-AI as Synthetic Collaborative Learning

Collaborative learning research has shown for decades that exposure to diverse viewpoints, structured disagreement, and peer explanation strengthens critical thinking (Johnson & Johnson, 1999). Students who have to articulate and defend their reasoning outperform those who work entirely alone, precisely because disagreement forces them to refine and justify their thinking.

Multi-AI frameworks recreate that environment synthetically. When a user queries multiple systems such as ChatGPT, Claude, Gemini, and others in parallel, and then compares their divergent answers, AI no longer functions as a single authority. It becomes a generator of structured dissent. While no published study has yet directly compared single-AI versus multi-AI parallel querying for cognitive outcomes, the mechanism aligns with established collaborative learning research showing that exposure to competing perspectives drives deeper analysis.

If one system frames an answer around efficiency, another leans on ethical risk, and a third surfaces historical context, the user can no longer accept one response without question. They face a set of competing claims, each with different evidence and implicit values. The cognitive task shifts from "find the answer" to "evaluate which reasoning is stronger and why."

The Human Enhancement Quotient (HEQ) framework is built to measure exactly this effect. It asks whether AI interaction increases or decreases the amount and quality of cognitive work done by the human. Enhancement occurs when systems generate interpretable conflicts that push users into higher order synthesis. Decline occurs when a single system smooths away disagreement and presents narrowing narrative certainty that users simply accept.

How HEQ Measures Enhancement

Human Enhancement Quotient is not a vibe score. It is a measurement frame that compares the amount and quality of cognitive work a person performs before and after governed AI collaboration.

Baseline measurement begins with human only tasks where participants solve problems, write analyses, or make decisions without AI involvement while we record performance, time, error rates, and evidence traces.

Intervention measurement involves the same participants using governed multi-AI workflows that require role assignment, conflict review, and explicit decision logs.

Outcome assessment returns them to independent tasks that test knowledge retention, transfer to novel problems, and the ability to explain their reasoning without AI assistance.

HEQ sits on that comparison. It looks at changes in task quality, error detection, synthesis depth, and resilience when AI is removed again. Enhancement is not whether people feel smarter with AI. Enhancement is whether they think better when the systems are taken away.

In practice, HEQ scoring might evaluate: (1) depth of source engagement, measuring whether users cite and critique multiple sources or accept single answers; (2) conflict resolution quality, assessing how thoroughly users reconcile contradictory AI outputs; (3) justification strength, rating the logical coherence of documented decisions; and (4) independent transfer, measuring performance on novel problems after AI assistance is removed. HEQ can combine behavioural metrics, rubric based reasoning scores, and where appropriate neural or physiological data, but its core function is simple. It asks whether governed AI collaboration leaves the human more capable on independent work than they were at baseline.

From Critique to Research Design

If the research agenda is going to move from blaming technology to testing methodology, the experiments need to change shape.

The simplest design is a three arm trial. One group works with no AI assistance. One group works with unstructured AI access that mirrors current studies. One group works with governed multi-AI protocols such as HAIA-RECCLIN that require role selection, conflict review, and written decision justification. All three groups receive the same tasks and are evaluated not only on immediate performance, but on later independent tasks that test transfer and retention. The question is not which group types most words per minute. The question is which group shows deeper reasoning when the AI is gone. Based on structured intervention effect sizes in comparable contexts, such as the Nigerian pilot reporting $d = 0.31$, we predict governed AI conditions will show Cohen's $d \geq 0.3$ advantage on transfer tasks compared to ungoverned conditions. Educational interventions typically target this threshold for practical significance.

A second design extends this frame into workplace settings. Teams are randomly assigned to continue their current AI practices, to adopt a single assistant with light guidance, or to implement checkpoint based governance with multi-AI dissent. Over several months researchers can track not just output volume but error rates, rework cycles, incident reports, and the quality of documented reasoning in decisions that matter. Cognitive enhancement then becomes visible in fewer downstream failures, stronger justifications, and better recovery when systems are wrong.

These are not exotic protocols. They are standard experimental structures that simply treat governance as the primary variable. Once those trials run, the field can finally distinguish what AI does under ungoverned conditions from what it does when the interaction is designed to enforce critical thinking rather than bypass it.

Methodology Determines Learning, Not Technology

This methodological problem is not new. Early studies of computers in classrooms often reported little improvement in learning outcomes and concluded that educational technology failed to deliver on its promise (Cuban, 2001). Later work

revealed the key variable. When computers were used mainly for drill and practice, they added little. When they were embedded into project based, collaborative, and reflective activities, they amplified learning instead. The technology stayed the same. The pedagogy changed.

AI now sits at the same crossroads. Ungoverned access allows cognitive outsourcing because nothing in the workflow interrupts it. Users move from prompt to polished answer in a single step. As the MIT study shows, neural engagement drops when the easiest path is to copy and paste rather than think and write. As the Microsoft survey shows, self confidence rises while effort declines, which is a classic signature of over reliance on automation.

Structured governance frameworks such as HAIA-RECCLIN and Checkpoint-Based Governance invert that pattern. They introduce friction by design. Users must review sources, reconcile conflicts, and document their reasoning at defined checkpoints before any answer can be treated as complete. Governance becomes the pedagogy behind AI use, in the same way that instructional design became the pedagogy behind computers in classrooms.

Weak processes produce weak learning regardless of technology. Rigorous processes produce advancement with or without advanced tools. The emerging research record correctly documents that ungoverned AI use can produce cognitive decline and over reliance. It then overreaches by treating that pattern as an inherent property of AI itself, rather than as a predictable effect of unstructured interaction.

Stakes Beyond Academia

The implications extend beyond research methodology into domains where cognitive capability directly affects outcomes.

Education: If governance frameworks can encode the structural moves that effective teachers make, they could extend quality instruction to contexts where expert educators are scarce. The Nigerian pilot suggests this is possible. The question is whether frameworks like HAIA-RECCLIN can preserve enough of the teacher mediated benefit to matter at scale.

Workplace: In high stakes fields such as law, medicine, and finance, cognitive errors compound into costly failures. If governed AI use reduces rework cycles and improves documented reasoning, the return on governance overhead may exceed its cost within months. Organizations currently face a choice between ungoverned AI that feels fast but generates hidden downstream errors, and governed AI that feels slower but builds capability that compounds.

Democratic discourse: If citizens learn to interrogate AI outputs through structured evaluation, they become harder to manipulate with generated misinformation. Governance frameworks that teach users to question sources, identify conflicts, and justify conclusions may serve as inoculation against epistemic manipulation. The stakes here are not merely individual but collective.

Addressing Objections

Governance Overhead

A reasonable objection is that governance frameworks add friction. They slow people down. In fast moving environments, any structure that increases cognitive load can look like a tax on productivity.

That concern is valid if you measure only first pass speed. It looks different if you measure total cycle time. When decisions rely on unstructured AI answers that feel confident but rest on shallow reasoning, the cost does not show up at the moment of generation. The cost shows up in rework, corrections, compliance issues, and trust erosion downstream. One limitation of this argument is that overhead costs in real settings have not yet been systematically quantified. Future research should measure both the time investment of governance checkpoints and the downstream error reduction to calculate net productivity impact.

Checkpoint friction is front loaded effort that reduces back end damage. In Growth Operating System terms, ungoverned AI optimises for short term throughput at the expense of capability. Governed AI optimises for capability that compounds. The research question is not whether governance slows the first answer. The research question is whether governance reduces total cost of error over time.

Teacher Mediation

Another objection is that the strongest positive studies relied on teachers, not frameworks. The Nigerian pilot succeeded because educators guided prompts. The EFL intervention worked because instructors designed tasks and feedback. It is fair to say that human pedagogy sits at the centre of those gains.

The point is not to equate HAIA-RECCLIN with a teacher. The point is that HAIA-RECCLIN captures the structural moves those teachers made. It forces context setting, question framing, feedback comparison, and explicit reflection into the workflow, even when no instructor is present. Governance is automated pedagogy. It turns those pedagogical patterns into operational requirements.

The research task is to see how much of the teacher mediated benefit we can preserve by encoding structure directly into the interaction with AI, instead of hoping that every user behaves like an excellent educator by instinct.

Time on Task

A methodologist can also argue that any governed condition will show better cognitive outcomes simply because people spend more time thinking. On that view, time on task is a confounding variable that undermines claims about governance.

The key move is to stop treating time and structure as separate forces. Governance induced time on task is the hypothesised mechanism. The framework makes it harder to move forward without planning, monitoring, and evaluation, so people naturally invest more cognitive effort.

Rather than trying to subtract that effect out, experiments should measure it directly. Does extra time spent in unstructured wandering produce the same gains as time

spent in structured conflict review and justification, or not. If structure matters beyond raw minutes, that will show up in the quality of reasoning and transfer, not just in the clock.

Implementation Pathways

Governed AI use sounds compelling in theory, but real world adoption faces practical hurdles. User resistance, tool complexity, and organizational inertia can derail even well designed frameworks. Several strategies can address these barriers.

Incremental adoption: Start with lightweight checkpoints rather than full governance protocols. A simple requirement such as "list three sources you considered" or "note one point of disagreement between AI outputs" imposes minimal overhead while establishing the habit of verification. Full HAIA-RECCLIN implementation can follow once users experience the value of structured review.

Visible incentives: Make cognitive enhancement measurable and visible. If users can see their HEQ scores improving over time, governance shifts from imposed burden to tracked development. Gamification elements such as "your synthesis depth improved 20% this month" can motivate sustained engagement.

Tool integration: Partner with AI platforms to embed governance prompts directly into interfaces. A system that asks "before accepting this answer, how would you verify it?" normalizes evaluation as part of the workflow rather than as external compliance. The goal is to make governed use the path of least resistance, not an additional burden.

Inviting Falsification

Once those objections are on the table, the path forward becomes less about defending a framework by assertion and more about inviting falsification. Either governed multi-AI use produces measurable cognitive enhancement beyond unstructured access, or it does not. The responsible move is to treat HAIA-RECCLIN and HEQ as hypotheses in need of data, not as branding.

The missing move in the research agenda is straightforward. Stop asking "Does AI make us dumber or smarter?" and start asking "Under which governance conditions does AI enhance cognition, and under which does it erode it?"

Only when identical AI systems are tested under governed and ungoverned protocols will the field be able to distinguish the effect of the model from the effect of the method.

The research record documents a real phenomenon: unstructured AI access can atrophy critical thinking. The error is in generalising that finding to AI itself, rather than recognising it as the predictable outcome of ungoverned use. Governance frameworks like HAIA-RECCLIN and measurement instruments like HEQ offer a testable alternative. The invitation now is for the research community to run the experiments.

We are ready to share HAIA-RECCLIN protocols, HEQ measurement instruments, and experimental design specifications with any university lab, corporate research

team, or policy organization willing to run comparative trials. The frameworks exist. The hypotheses are specified. What remains is the data. Collaborators interested in testing governed versus ungoverned AI conditions can contact the author to discuss partnership structures that preserve research independence while enabling practical validation.

References

Crawfurd, L., Mills, M., Elks, P., Beg, S., Eze, P., & Ramachandran, D. (2025). AI tutoring in Nigerian after-school programs: Experimental evidence. World Bank Policy Research Working Paper.

Cuban, L. (2001). Oversold and underused: Computers in the classroom. Harvard University Press.

Diaz, J. (2025, March 10). Science shows AI is probably making you dumber. Luckily, there's a fix. Fast Company.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906-911.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. Theory Into Practice, 38(2), 67-73.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kosmyna, N., et al. (2025). This is your brain on ChatGPT: EEG analysis of cognitive engagement during LLM-assisted writing. arXiv preprint arXiv:2506.08872. [Preprint]

Lee, M., et al. (2025). The impact of generative AI on critical thinking: A survey of knowledge workers. Microsoft Research / Carnegie Mellon University.

Liu, W., & Wang, Y. (2024). The effects of AI tools on critical thinking in English literature classes: An intervention study. European Journal of Education, 59(4).

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19(4), 460-475.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational Psychology Review, 22(2), 123-138.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.