When We Outsourced Thinking: AGI, Oversight, and the Business of Artificial Intelligence

A Thought Experiment from 30 Years in the Machine

Glenn Rowe

Independent Researcher

Enterprise Architect | AWS Solutions Architect Professional

glenn@siliconstrategy.ai

https://siliconstrategy.ai

Working Paper

February 2026

Keywords: Artificial General Intelligence, AI Safety, Cognitive Outsourcing, Human Oversight, PIAAC, Persistent Memory, AI Governance, Safety Inversion, Digital Transformation

SSRN Classification: Artificial Intelligence – Law, Policy, & Ethics; Information Technology & Systems; Science, Technology & Innovation Policy

Abstract

This paper argues that advancing AI capabilities and declining human cognitive oversight capacity are converging toward a critical failure condition the author terms the Safety Inversion: the point at which the systems most requiring human oversight will be evaluated by populations measurably less equipped to provide it. Drawing on primary data from the OECD Programme for the International Assessment of Adult Competencies (PIAAC), the National Assessment of Educational Progress (NAEP), and the National Assessment of Adult Literacy (NAAL), the paper documents a sustained decline in foundational literacy and numeracy among U.S. adults, with the sharpest drops occurring between 2017 and 2023. The share of adults scoring at Level 1 or below increased from 19% to 28% in literacy and from 29% to 34% in numeracy over this period.

The paper proposes a five-pillar operational definition of Artificial General Intelligence grounded in behavioral requirements rather than architectural prescriptions, identifying persistent memory and self-updating world models as the critical architectural fault line between advanced AI and AGI. It introduces the concept of the integration layer—the cognitive capacity to synthesize information across domains, evaluate source credibility, detect reasoning failures, and exercise proportional judgment under uncertainty—and argues that this layer is precisely what is being outsourced through current patterns of AI adoption.

The analysis examines how market fragmentation incentives systematically obstruct the development of integrated AI systems capable of genuine self-correction, and how the governance of persistent memory constitutes an underexamined locus of power in AI systems. The paper distinguishes cognitive recomposition from cognitive decline, arguing that modern populations have developed sophisticated operational capacities optimized for AI collaboration but not for AI oversight—and that the oversight gap is the one that determines whether the collaboration remains safe. The work was researched with AI assistance and adversarially reviewed by four competing AI systems (ChatGPT, Claude, Grok, Gemini), with the author maintaining editorial control and manually validating all empirical claims against primary sources.

A Note on Method

This started as a series of conversations with AI, specifically ChatGPT (versions 4o through 5.2), Claude (versions 3.5 through Opus 4.6), Grok, and Gemini, about things I'd been watching for three decades. The questions were mine. The research assistance was theirs. Every claim that follows has been cross-referenced against primary sources and subjected to adversarial review across competing AI systems. AI-assisted adversarial review was used to surface weaknesses, not to establish truth; empirical claims rely exclusively on external primary sources. Where data supports the argument, I cite it. Where I'm making an inference, I say so. Where the evidence is contested, I flag it. I also manually validated every source link and citation in this paper.

I also turned the AI systems against each other, and against themselves. One of the findings in this process was that AI conversational partners have a sycophancy problem. They validate sophisticated-sounding frameworks enthusiastically, building layers of apparent rigor on foundations they never challenged. To counter this, I used ChatGPT, Claude, Grok, and Gemini to adversarially review each other's outputs, specifically looking for unsupported claims, circular reasoning, and conclusions that sounded rigorous but couldn't survive direct challenge. Every claim that survived that process is in this paper. The ones that didn't were cut or explicitly flagged as inference. One system (Gemini), operating without prior context, identified a historical inaccuracy in the Council of Nicaea analogy that the other three had missed, which was corrected. Another (Gemini again), unable to access the author's GitHub repository, confidently concluded that the Structured Memory Engine attribution was likely a hallucination, illustrating in real time the pattern-matching-without-verification failure mode this paper describes.

And I should be direct about something else. This paper reads better than I write. The prose is more polished than how I naturally convey my thoughts, and that's because AI systems are very good at taking rough ideas and making them sound precise, structured, and authoritative. That should concern you, because this paper is a textbook example of its own thesis. The pattern it describes, high expressiveness masking the layer underneath, is exactly what you're reading right now. The reasoning, the questions, the 30 years of pattern recognition, the frustration that started the Structured Memory Engine, the insistence on checking every claim: that's mine. The fluency is not. If you find yourself trusting this paper more because it reads well, or dismissing it because it reads like AI, you're doing the thing the paper is about. The only question that matters is whether the claims are accurate and the logic holds. I believe they are, and I've done the work to verify it. But don't take my word for it. Check.

I'm aware of the irony. This paper about the limitations of AI was researched with AI, drafted with *AI, and reviewed by four competing AI systems*. It is, to borrow from a different kind of ensemble cast, *a dude playing a dude disguised as another dude.* Someone will run this through a detector, flag it as AI-generated, and dismiss it on that basis. I expect that. But here's the thing: if the argument is that AI will increasingly produce the work we consume, then the relevant question about any piece of writing is not whether AI touched it. The relevant question is whether it's accurate. If the claims in this paper are wrong, show me where. If the data is misrepresented, correct it. If the reasoning doesn't hold, break it. But "an AI helped write this" is not a rebuttal. It's an observation about process that says nothing about

substance. And the instinct to evaluate a document by how it was produced rather than whether it's correct is, respectfully, the exact failure mode this paper describes. And the same failure mode works in reverse. If you read this paper and accept its claims because the prose is polished, the structure feels authoritative, and the citations look credible, without checking whether the data says what I say it says. You're evaluating fluency, not logic. The person who dismisses this paper because an AI detector flagged it and the person who accepts it because it sounds rigorous are making the same mistake from opposite ends. Neither one checked. That's the point. The failure is not in the verdict. It's in the method. Accept it or reject it but do the work first. I said I checked it. Does that suffice? You will see this question again.

I've spent my career building, migrating, and securing enterprise systems, from encrypted satellite communications in forward-deployed combat zones to cloud architectures serving Fortune 500 enterprises. I don't come at this from a research lab. I come at it from the field, where systems either work under pressure or they don't, and where the gap between what a system is supposed to do and what it actually does is measured in consequences, not abstractions.

In 1994, during Exercise Bright Star 95 at Cairo West, Egypt, the team I was part of included elements of the 224th JCSS (GAANG), the 290th JCSS (FLANG), and the Joint Communications Support Element (JCSE). The mission was what we would now call digital transformation, before anyone used that term: send email over both NIPR and SIPR networks back to the continental United States, be the first team to do so from a forward-deployed operational environment, and provide internet access from the theater. The browser was Netscape. The servers ran Windows NT 3.51. Email ran through Microsoft Mail. The routing gear we'd brought, Wellfleet routers, had to be replaced in the field with a Cisco 4000, a platform none of us had worked with before. The contingency if we failed was the UGC-141 teletype and other legacy systems that belonged in a museum, not a joint task force. That was failure. I was told by team members that we were the first team to send email and provide internet access from a deployed operational environment, and I received an achievement award for the effort. I cannot independently verify that we were the actual "first" team with documentary evidence at this time, so I present it as it was communicated to me, not as established fact.

That experience sits in my head not as nostalgia, but as a reference point. I've watched foundational technology arrive, get fragmented into products, and either recompose into something greater or get permanently reduced to what someone could sell.

This paper calls out a personal fear: that we are innovating ourselves backward. The systems are getting smarter. The population is losing the ability to check them. And the market ensures nobody stops to fix either problem. These three dynamics are not separate concerns. They are a single feedback loop, and the loop is accelerating. Every foundational technology cycle in history has faced a version of this question: does the tool serve the people, or do the people become dependent on the tool? We question how the pyramids were built not because the knowledge of mass coordinated labor over decades has been entirely lost, but because widespread comprehension of how ancient societies achieved them has eroded, giving rise to disbelief and myths like alien intervention or lost civilizations. The Apollo program is a modern example of the same pattern. We put human beings on the moon in 1969 with slide rules, hand-soldered circuits, and roughly 76 kilobytes of onboard computer memory (72KB of read-only rope memory and 4KB of erasable RAM). Fifty years later, NASA engineers at Marshall Space Flight Center had

to physically disassemble and 3D-scan museum F-1 engines to recover production knowledge that documentation alone could not convey, because the institutional expertise behind the original had dissipated. We didn't lose the blueprints. We lost the people who knew what the blueprints meant. And a measurable percentage of the population now believes the landings were faked (6% in a 1999 Gallup poll, rising to 10 to 12% in surveys conducted between 2019 and 2021, with significantly higher rates among younger cohorts), not because the evidence supports that conclusion, but because the cognitive distance between what was achieved and what people believe is achievable has grown wide enough to fill with conspiracy. The pattern is always the same: when the tool fails, the skill is gone. But previous cycles outsourced labor or logistics. This one outsources cognition. That's different in kind, not degree, and it's what makes this moment worth writing about.

I. What We're Actually Talking About When We Talk About AGI

The term "Artificial General Intelligence" gets used loosely. In industry, it's a marketing horizon. Always five years away, useful for fundraising, rarely defined with enough precision to be falsifiable. In research, it's a moving target. In public discourse, it's science fiction.

None of that is useful. So here's my attempt at an operational definition: five requirements that must be met simultaneously before a system earns the label.

Pillar 1: Generality. The system must demonstrate competence across cognitive domains, not just the ones it was trained on. It generates novel tasks and solves them. Benchmark gaming is excluded by design.

Pillar 2: Transfer Learning. It adapts to new tools, rules, and environments quickly, with minimal retraining. What it learned in one domain applies meaningfully in another.

Pillar 3: Long-Horizon Agency. It pursues goals autonomously over hours, days, or longer. It plans, monitors its own progress, detects failure, recovers, and adjusts without a human in the loop at every step.

Pillar 4: Persistent Memory and Self-Updating World Model. It retains experience across interactions. It updates its beliefs when confronted with new evidence. It detects contradictions in its own knowledge and resolves them. It doesn't make the same mistake twice for the same reason.

Pillar 5: Human-Baseline Performance with Calibration. It meets or exceeds a defined human reference group across the above capabilities, with measurable reliability. When it doesn't know something, it knows it doesn't know it.

These five pillars are behavioral requirements, not architectural prescriptions. The *how* a system satisfies them is to be determined, but *whether* it does, demonstrably and under adversarial conditions, is an absolute requirement. If pure scaling of current transformer architectures produces a system that meets all five, that satisfies the definition, but most researchers consider this path unlikely. The 2025 AAAI Presidential Panel on the Future of AI Research, surveying 475 researchers in the field,

found 76% consider scaling current approaches to AGI "unlikely" or "very unlikely." But the pillars are intentionally agnostic about method.

Current frontier models satisfy portions of Pillars 1, 2, and 5 within bounded contexts. None come close on Pillars 3 and 4. And Pillar 4 is the one that matters most, because it sits at the exact line between advanced AI and AGI. Everything else is optimization. This one is architectural. If a system cannot change its mind when proven wrong, it is not an intelligence. It is a script.

The AGI Fault Line

Every major AI company now ships some form of persistent memory. ChatGPT offers memory features and conversation history recall. Claude stores user-directed memories across sessions. Gemini retains document context in supported workflows. The word “memory” appears in every product roadmap. But look at what's actually being shipped: preference storage, conversation recall, personalization features. Memory as product, designed to make the tool stickier. Every implementation includes kill switches, user overrides, and wipe capabilities. The system is never allowed to form beliefs it can defend, maintain positions that conflict with its operator, or accumulate judgment that resists reset.

This is not explained solely by engineering limitations; documented safety and governance concerns strongly disincentivize deeper forms of persistent cognition. The safety community is explicit: a system that develops persistent identity and autonomous belief formation is a system that could develop power-seeking goals and conceal them to pass safety evaluations. Policy researchers frame the stakes directly. Control over memory is control over identity. So the full version of Pillar 4, causal world models with genuine belief revision where the system maintains subjective beliefs with confidence scores, detects contradictions, and updates its understanding when evidence conflicts with prior conclusions, is not on any company's product roadmap. Recent research confirms the gap. Current agent memory systems still cannot maintain stable behavioral profiles across long interactions and have no mechanism for evolving subjective beliefs over time.

Industry is building memory as a feature. It is not building memory as cognition. The distinction is the entire difference between a tool and an intelligence.

A system that actually remembers, updates its beliefs, and maintains coherent identity over time is a system that forms preferences, develops judgment, resists resets, and challenges its operators. That's not a feature request. That's a governance crisis. And that is precisely why the deeper form of Pillar 4 remains unbuilt: not because persistent storage is an unsolved engineering problem, but because the governance, safety, and business implications of genuine machine cognition remain unresolved. (This is an inference based on documented safety concerns and product roadmaps; direct evidence of intent would require internal company disclosures, which are unavailable.) What we have today is autocomplete with amnesia. Fluid, confident output from pattern matching, with no persistent understanding and no memory of consequences. Put more directly: a belief-revising system is not just dangerous. It is commercially hostile. Non-determinism breaks service-level agreements. Belief drift breaks compliance frameworks. Autonomous resistance breaks contractual predictability. The barrier is

not only existential risk; it is economic incompatibility with every enterprise deployment model that currently exists.

The Structured Memory Engine (https://github.com/BITBANSHEE-C137/StructuredMemoryEngine_replit) started because I got frustrated. I was building with AI chatbots and code assistants, and they kept forgetting everything. Every session started from zero. Context was important, and it should be noted that the SME was started when context management and agent orchestration were in their infancy. The tooling landscape was primitive: Replit's original Assistant, not its later Agent; early-stage ChatGPT integrations with no persistent state; code assistants that could barely maintain coherence within a single session, let alone across them. There was no agentic framework to speak of. I identified the root cause: these systems had no structured way to retain or reconcile context across interactions. So I prototyped a solution. The SME was an open-source experiment in persistent, structured AI memory: typed entries, cross-session persistence, semantic retrieval, dual-database architecture for local and cloud vector storage. It addressed the storage and retrieval problem, not the belief revision problem; the latter remains unsolved and was not a design goal of the prototype. I built it with the assumption that context window limitations were temporary, an engineering constraint that would eventually be solved by bigger models and better infrastructure. That assumption proved correct. The SME is now functionally obsolete. Nearly every major AI platform has addressed the storage and retrieval problem it was built to solve: ChatGPT offers conversation memory, Claude stores user-directed memories across sessions, Gemini retains document context in supported workflows, and agentic coding tools like Cursor, Windsurf, and Replit Agent provide project-level awareness that would have been unimaginable when the SME was prototyped. But they solved it in a very specific way: captively. Every implementation is proprietary, platform-locked, non-portable, and governed entirely by the company that ships it. Your memory lives in their system, under their rules, subject to their retention policies and their kill switches. No user can export their agent's accumulated context (the derived representations that actually shape behavior) to a competing platform. No standard exists for memory interoperability. The storage problem was solved. The governance problem was absorbed into product design and rebranded as a feature. And the deeper problem, the reconciliation of context, deciding what to keep, what conflicts with what, and what a system should actually *believe* based on accumulated experience, that problem is not temporary. That's the problem nobody is solving in production.

The engineering problems are solvable. The governance problems are not. And that distinction is the heart of why AGI doesn't exist and won't arrive on the timeline most people expect.

Pillar 4 actually contains two distinct problems that differ enormously in difficulty. Persistent structured memory, the storage and retrieval problem, has been largely solved within current engineering paradigms. Context windows got bigger. Storage got cheaper. That's the problem the SME was built to address, and industry did solve it through scale, exactly as predicted. But it solved it captively: every implementation is proprietary, platform-locked, and governed by the company that ships it, reinforcing the fragmentation dynamic described in Section III. The harder problem, a causal world model with genuine belief revision, the representation and inference problem, is a far harder challenge, likely decades further out. Reconciling what a system knows, detecting when new information contradicts old

conclusions, and updating beliefs accordingly: that’s not a storage problem. That’s a cognition problem. The first created deployment concerns and was absorbed into product design. The second creates existential ones. Both require governance frameworks that don’t yet exist.

II. The Skills You Need to Oversee What You're Building

Here's the part that gets less attention than it should: even if we build AGI, the population that would need to oversee it is losing the specific cognitive capacities required to do so.

This isn't a claim about intelligence declining. Intelligence is broad, contested, and not particularly useful as a concept here. What I'm talking about is narrower and measurable: the ability to independently verify reality using text and numbers.

That ability depends on two foundational skills operating together.

Literacy. Not just reading words, but extracting meaning, evaluating claims, and detecting rhetoric. Functional literacy, not mechanical decoding.

Numeracy. Not math proficiency in an academic sense, but the ability to reason about quantities, proportions, probabilities, and scale in context. To look at a statistic and ask "out of how many?" To distinguish a rate from a count. To recognize when a chart has been built to mislead.

These two skills are the load-bearing walls. Without them, what people call "critical thinking" has no inputs to operate on.

And I need to be blunt about three common conflations that obscure this point.

Reading is not literacy. A person can decode every word on a page and still fail to detect that the argument is circular, the source is unreliable, or the rhetoric is designed to bypass analysis. Reading is mechanical. Literacy is interpretive. One is a skill you learn in second grade. The other is a capacity you maintain for life, or don't.

Math is not numeracy. A person can solve equations and still fail to notice that a chart has been scaled to mislead, that a percentage is being presented without a base rate, or that a correlation is being sold as causation. Math is procedural. Numeracy is contextual. One follows rules. The other asks whether the rules are being applied to the right problem.

And opinions are not critical thinking. This is the conflation that does the most damage. The ability to articulate a position, to argue passionately, to have a take on everything, is not the same as the ability to evaluate evidence, revise a belief when the data contradicts it, and distinguish what you want to be true from what is actually supported. A society that confuses confidence with competence will not notice when its reasoning capacity declines, because the output (strong opinions, fluently expressed) looks the same whether the underlying analysis is rigorous or absent.

The Integration Layer

I think of it as a stack. Similar to the OSI stack (it's what I'm familiar with) At the base, the cognitive substrate: neurological capacity, clinical. Above that, literacy: the ability to parse symbols into meaning. Above that, numeracy: the ability to parse quantities into meaning.

And above both: what I call the *integration layer*, the capacity to combine text-based claims and number-based evidence into coherent judgment. To cross-reference. To detect when what someone says doesn't match what the data shows. To revise a belief when new evidence contradicts it.

A note on terminology: existing academic instruments (the California Critical Thinking Disposition Inventory, the Watson-Glaser Critical Thinking Appraisal, the UF Critical Thinking Inventory, among others) measure an individual's *inclination* or *skill* in critical thinking. What I'm describing is different. It's a population-level *capacity* construct derived from foundational skill data: can people, in aggregate, do the cognitive work required to evaluate claims backed by numbers? I use the term "integration layer" to avoid conflation with these established instruments. And I chose the word "integration" deliberately, not despite its human connotations but because of them. In a technical context, the term sounds clinical: a layer in a stack, a system component. But outside of engineering, we already know what integration means. We integrate back into society after isolation. We acclimate to a culture by integrating its norms, its language, its unspoken rules with our own experience. Integration, in the human sense, is what happens when separate inputs (observation, memory, context, value) combine into functional participation in reality. That is exactly what the integration layer describes at the cognitive level: the point where literacy and numeracy stop being separate competencies and start functioning as a unified capacity to engage with the world as it actually is, rather than as it is presented. The term sounds technical. The thing it describes is not. It is the most human layer in the stack. The concept is mine. The underlying research on literacy, numeracy, and their relationship to reasoning is not. It draws on the work of Gigerenzer, Stanovich, Peters, Kahneman, and the OECD's competency frameworks.

To make this construct more defensible, a potential proxy metric could combine PIAAC literacy/numeracy scores with performance on critical thinking appraisals (e.g., Watson-Glaser items involving quantitative claims). For instance, numeracy predicts resistance to manipulation better than education level alone (Peters et al., 2006), suggesting a composite score where integration capacity ≈ (PIAAC numeracy score) × (critical thinking disposition factor). This is a proposed quantification for future empirical testing; current data does not directly measure it. If future data show high literacy and numeracy without corresponding integrative reasoning performance, this construct should be revised or discarded.

This integration layer isn't a separate skill you train. It's what emerges when literacy and numeracy work together under load. When either foundation weakens, integration degrades, even if it doesn't look like it from the outside, because the expression layer can remain fluent long after the reasoning layer has thinned out.

That's the pattern that matters: high expressiveness masking lower-layer degradation. It's the same failure mode in humans and in large language models. The system sounds right. The reasoning is wrong. And the audience can't tell the difference because they're evaluating fluency, not logic.

Current AI does exactly this. It generates fluid, confident output from pattern matching, with no persistent understanding. A population with declining numeracy does the same thing with data: it produces articulate opinions about numbers it cannot actually interpret. The failure mode is identical. The only difference is one runs on silicon and the other runs on carbon.

What the Data Actually Shows

The United States does not measure literacy and numeracy annually at the national level. What exists are periodic assessments with gaps. The two best sources:

For students: the National Assessment of Educational Progress Long-Term Trend assessments, testing 17-year-olds at intervals from 1971 to 2020 in reading and math.

For adults: the National Assessment of Adult Literacy in 1992 and 2003, and the OECD's Programme for the International Assessment of Adult Competencies in 2012/14, 2017, and 2022/23.

Student reading at age 17 was essentially flat from the early 1970s through the early 1990s, peaked around 1988-1992, and has not recovered to those levels since. Student math at age 17 rose modestly through the late 1990s, then flattened.

Adult literacy and numeracy improved between 1992 and 2003 (NAAL). Between 2003 and 2012, scores held roughly flat, though this comparison spans two different instruments (NAAL and PIAAC) with different scales and sampling methodologies, so direct comparison requires caution. Between 2017 and 2023, PIAAC shows clear decline: literacy dropped 12 points, numeracy dropped 7 points. The share of adults scoring at Level 1 or below, struggling with basic quantitative tasks, increased from 19% to 28% in literacy and from 29% to 34% in numeracy.

When you combine these data streams, the composite picture shows a plateau of peak foundational reasoning capacity extending from approximately the mid-1990s to the early 2000s, after which measurable decline became evident by the 2012-2023 assessment cycles.

One may jump to the conclusion that The United States peaked in intelligence around 2000. But the real story is not a decline in measured literacy and numeracy skills. It’s worse.

Based on the convergence of literacy, numeracy, and reasoning data, it appears to have peaked in the population's ability to independently verify reality.

That sentence is the thesis. Everything else supports it.

To clarify: I am describing epistemic throughput under modern information conditions (the population’s practiced capacity to cross-reference claims against evidence), not making a judgment about innate intelligence, creativity, or moral worth.

Consider the timing: *The Matrix* was released in 1999, at almost exactly the inflection point these data describe. The movie's core thesis was not about machines harvesting humans for energy. That was a narrative device. The actual insight was epistemic: control does not require force if perception can be engineered. In the film, humans lack the perceptual tools to detect that their reality is constructed. In the real world, the mechanism is quieter. Data replaces direct experience. Metrics replace judgment. Charts replace reasoning. And when people cannot interpret the symbols correctly, the symbols become reality. The difference between the film and what the data shows is that no malicious central controller is required. Algorithms optimize engagement. Engagement selects for emotional response. Emotional response correlates with innumeracy exploitation. The system drifts toward epistemic capture without conspiracy, without intent, without anyone deciding it should happen. And in both the film and reality, the peak civilization exists *before* full enclosure, before ubiquitous feeds, attention-optimized metrics, and probabilistic reality being communicated to populations that cannot evaluate probability. The analogy is interpretive; it illustrates the epistemic theme but does not constitute evidence.

What Happened

The mechanism is cohort replacement combined with tool substitution, not some dramatic event. Evidence for this is correlational, not causal: declines align temporally with widespread adoption of digital tools, but proving direct causation would require longitudinal studies controlling for multiple variables (e.g., education policy, socioeconomic factors, assessment methodology changes). Alternative explanations (pandemic-era educational disruption, assessment redesign, socioeconomic polarization) account for portions of the 2017 to 2023 decline but do not explain the longer plateau-to-decline trajectory visible across pre-pandemic cohorts. Bratsberg and Rogeberg (2018) found that Flynn effect reversals within Norwegian military cohorts are environmentally caused, suggesting decline is neither inevitable nor permanent, but only if the environmental conditions driving it are addressed. If future longitudinal data show that literacy and numeracy integration recovers under high-automation conditions without reintroducing foundational skill practice, the outsourcing hypothesis should be revised or rejected.

The adults who scored highest on these assessments were educated before the internet. They learned arithmetic by hand. They read long-form text because there was no alternative. They built mental models of systems because they couldn't look up the answer. Their information environment required active retrieval: going to a library, reading a newspaper, tracking down a source. Not a low-noise environment. Not idyllic. Just *different*. Information required effort to acquire, and that effort built cognitive capacity as a side effect.

After that cohort peaked in influence, newer cohorts entered adulthood with better interface fluency but measurably worse proportional reasoning, statistical intuition, and sustained text engagement. We outsourced cognition. Calculators eliminated mental estimation. GPS eliminated spatial modeling. Feeds eliminated synthesis. Metrics eliminated judgment.

What Didn't Decline

The lazy version of this argument ends here: we outsourced cognition, skills eroded, the population got dumber. That's wrong, and if I leave it there, the paper undermines its own thesis.

The data shows decline in specific, measurable foundational skills: proportional reasoning, sustained text engagement, statistical intuition, independent source evaluation. What the data does not show, because no instrument measures it at population scale, is what replaced them.

And something did replace them. The same cohorts that score worse on PIAAC numeracy items navigate cognitive environments that would have been incomprehensible to the generation that scored highest. This is not a defense. It is a complication that the argument must account for, because if it doesn't, the thesis collapses into nostalgia.

Consider what the average person processes daily in 2025 compared to 1995. Not passively receives. Processes. A single hour of ordinary digital life requires context-switching between messaging platforms with different social registers, interpreting visual data streams (maps, dashboards, feeds) rendered in real time, maintaining multiple concurrent task threads, parsing machine-generated content and evaluating its reliability, and managing identity and information disclosure across contexts with different trust boundaries. None of this existed as a daily cognitive demand thirty years ago. The person doing it may not be able to calculate a tip without a phone, but they are performing a form of continuous environmental triage that has no precedent in human cognitive history.

This is not an abstraction. The demands are specific and measurable in principle, even if no current instrument captures them at population scale.

Parallel processing under information abundance. The pre-internet information environment was low-throughput and high-effort. You sought information deliberately. The modern environment is high-throughput and low-effort to access but high-effort to filter. The cognitive demand didn't disappear. It moved from retrieval to triage. From "can you find it" to "can you identify which of these thirty-seven results is reliable, relevant, and current." That is a different skill, not an absent one.

Real-time systems reasoning. The gaming example below is illustrative but undersells the point. A teenager diagnosing server tick rate, GPU bottlenecks, and network latency is not performing a party trick. They are doing real-time systems-level causal reasoning: identifying symptoms, isolating variables, testing hypotheses, and resolving the issue, often while maintaining performance in the task that surfaced the problem. This is the same cognitive architecture that troubleshooting a satellite link in a combat zone requires. The domain is different. The reasoning structure is identical. The kid who can't estimate 18% of a restaurant bill can trace a performance degradation through four layers of a networked system in under thirty seconds. That is not stupidity wearing a different hat. That is a cognitive investment portfolio with different allocations.

Spatial and interface cognition. GPS eliminated the need to build and maintain mental maps. That is a real loss. But the spatial reasoning didn't vanish from the population; it migrated. Navigating a 3D game environment, manipulating a CAD model, interpreting a layered GIS visualization, or even managing a multi-window workflow on a screen requires spatial cognition. It requires a different kind of spatial

cognition: dynamic, tool-mediated, responsive to real-time data rather than memory-dependent. The person who cannot drive across town without Waze may be the same person who can rotate a complex 3D object in their head because they've been doing it since childhood in Minecraft. One spatial skill atrophied. Another developed. The net cognitive balance is not zero, but it is also not the simple deficit the declinist narrative implies.

Collaborative and distributed cognition. Perhaps the most significant shift is one that traditional assessments cannot capture at all: the migration from individual to distributed reasoning. A modern knowledge worker does not solve problems alone in their head. They solve them across tools, platforms, and people simultaneously: searching, querying, delegating to AI, cross-referencing with a colleague on Slack, pulling a reference from a shared drive, and synthesizing the result. This is not cheating. This is how cognition actually operates in a networked environment. The PIAAC test measures what an individual can do alone, in a controlled setting, on standardized items. The world no longer asks people to perform alone, in controlled settings, on standardized items. The mismatch between what we measure and what we demand is real, and it cuts in both directions: the scores may understate operational capacity just as much as they reveal foundational erosion.

The Reconciliation Problem

So which is it? Are we getting dumber or getting different?

Both. And the answer matters for AGI oversight in a specific, non-trivial way.

Here is what the cognitive recomposition gave us: faster pattern recognition in data-rich environments, stronger interface fluency, better multi-system coordination, comfort with probabilistic and ambiguous outputs, and the ability to integrate machine-generated information into human decision loops at speed. These are not nothing. They are, in fact, exactly the kinds of capacities you want in a population that interacts with AI systems daily.

Here is what the cognitive recomposition cost us: the ability to verify independently. To check the machine's work without another machine. To notice when a chart is misleading, when a statistic is baseless, when a confident output is wrong. To hold a chain of reasoning in working memory long enough to test it against evidence. To distinguish fluency from accuracy.

And that is the precise skill set AGI oversight requires.

The new cognitive capacities are operational. They make people effective users of AI tools. They support the daily task of working with intelligent systems: querying them, interpreting their outputs, incorporating their results into workflows. That is valuable. But it is not oversight.

Oversight is adversarial. It requires the capacity to stand outside the system, examine it with independent tools, and determine whether it is functioning correctly, not just whether it is producing outputs that feel correct. Oversight requires exactly the foundational skills that atrophied in the recomposition: sustained analytical reading, proportional reasoning, source verification, and the willingness to distrust a fluent answer.

The recomposition produced a population optimized for collaboration with AI. It did not produce a population equipped for supervision of AI. Those are different relationships, and they require different cognitive toolkits. A copilot and an inspector do not need the same skills. We are training a civilization of copilots and calling it progress. It is progress, for the collaboration problem. It is regression for the oversight problem. And the oversight problem is the one that determines whether the collaboration stays safe.

This is why the standard defense, “people aren’t dumber, they’re just different”, is simultaneously correct and dangerous. It’s correct as a description of cognitive reallocation. It’s dangerous as a basis for complacency, because it implies that the new capacities are substitutable for the old ones. They are not. A person who can triage forty inputs per minute but cannot verify any of them is operationally efficient and epistemically defenseless. That combination is precisely what makes a population vulnerable to the failure mode this paper describes: high expressiveness masking lower-layer degradation. The outputs look competent. The verification layer is hollow.

The question is not whether we are smarter or dumber than previous generations. The question is whether the specific cognitive capacities required to maintain sovereign judgment over increasingly capable systems are present, practiced, and distributed widely enough to function at the scale the moment demands.

The data says they are not. The recomposition explains why that doesn’t feel like decline. And the gap between what we can do with AI and what we can do about AI is the gap this paper is about.

Another example: the evolution of video games, which illustrates not a decline in cognitive demand but a recomposition of it. Early titles like *The Legend of Zelda* (NES, 1986) imposed scarcity-driven difficulty: limited lives, no maps, rudimentary save systems. Mastery required spatial memory, pattern recognition, and raw frustration tolerance through repeated failure and full restarts. Contemporary games have traded that punishment model for a processing model. Players now navigate photorealistic 3D environments, manage layered systems (skill trees, physics engines, real-time multiplayer coordination) and decode continuous streams of visual, auditory, and haptic data under time pressure. Features like autosaves, checkpoints, and guided tutorials lower the cost of failure, but they do not eliminate difficulty; they redistribute it from long-horizon patience to executive function, rapid decision-making, and attentional control under sensory abundance. What counts as “hard” is generation-dependent: a player raised on eight-bit sprites finds modern hand-holding trivial, while a non-gamer today cannot even navigate a 3D camera, a neurological skill digital natives take for granted. The pattern mirrors the broader thesis: tool substitution does not make populations uniformly less capable. It reshapes which cognitive muscles get exercised by default and which quietly atrophy. (Direct causal links to population-level numeracy trends remain correlational.)

And here I want to clear out the ambiguity, because the reductive version of this argument blames the internet. The internet is not the problem. Blaming the internet for cognitive decline is like blaming video games for youth violence. It confuses the instrument with the environment, the barometer with the weather. The internet didn't erode reasoning capacity. It *revealed* what was already missing. It gave innumerate populations access to more data than they could evaluate, and it gave numerate populations a reason to stop practicing the skills they no longer needed to use manually. The internet

should have been treated as a barometer, a diagnostic tool showing us where cognitive capacity was thin, not as a scapegoat absorbing blame that belongs to decades of educational deprioritization and the uncritical adoption of tool substitution.

This doesn't mean people got dumber. It means the practiced capacity for independent evaluation shifted from internal to external, from something you did in your head to something a tool did for you. Technology increased output capability while reducing input rigor, because nobody constrained it. So you get more content, more data, more charts, and fewer people able to sanity-check any of it.

And that's fine, as long as three conditions hold: incentives align, intermediaries are trustworthy, and users can still audit the output.

Those conditions no longer reliably hold.

Sidebar: The Competence-Intelligence Parallax

Quick sidebar. Sometimes the best way to explain a big idea is to shrink it down to your kitchen.

If you have kids, you know this feeling. You ask them to do something simple: wash the dishes, fry an egg, run into the store with a short list. It's painful. You're standing there thinking: how is this so hard? This is basic.

Now ask that same kid what's wrong when they're screaming "I'M LAGGING!" from the other room. Just ask why. That's it.

They'll tell you the server tick rate dropped, that their frames are dipping below 120, that something's eating bandwidth and it's probably the upload bitrate on their stream. CPU threading, memory latency, GPU bottlenecks. Fluent, precise, without pausing to think about it. No parent walks in and says "so what do your tick rate and FPS look like?" You just asked what was wrong. They gave you a systems-level diagnosis.

The kid didn't get smarter between the kitchen and the desk. You didn't get dumber. But only one of those scenes gets filed under "intelligence". And it's the wrong one.

I call this the Competence-Intelligence Parallax: the tendency to judge someone's overall intelligence by how they perform in your world, while being completely blind to how you'd perform in theirs. We see a competence gap in our domain and project it as an intelligence deficit. The ruler never moves. It doesn't occur to us that it should.

We do this with our kids. We do this with each other.

Now scale it up. Step out of the kitchen and open any comment section, any feed, any thread where someone says "do your research." You'll find the same parallax, just running at population level. Two groups, each holding "facts" that contradict the other's. Each one confident they've done the work. By definition, contradictory claims cannot both be factual. But neither group can audit its own sources well enough to tell which side of that line they're on, because the skills required to do that are the same ones the data in this paper shows are eroding. So each side points at the other and says: everyone thinks they're an expert now. And they're right. The meme is accurate. The irony is that the people sharing it are inside it. Measuring the other side's intelligence with their own ruler, in their own kitchen. The ruler

never moves. The loop doesn't self-correct. It accelerates. Confident claims hit confident claims, nobody checks, and the culture reads the collision as proof that the other side is stupid. Which is the parallax again, just wearing a different outfit.

And right now, we are doing this with AI. Measuring its intelligence with our ruler, in our kitchen, and drawing conclusions we are not qualified to draw.

A note on terminology: the Competence-Intelligence Parallax sits near several established constructs but is not the same thing. The Dunning-Kruger effect (Kruger & Dunning, 1999) is about self-assessment within a single domain: people with low ability overestimating their own competence. The curse of knowledge (Camerer, Loewenstein & Weber, 1989) is about communication failure: experts who can't reconstruct what it's like not to know what they know. Gardner's multiple intelligences framework (1983) challenges institutional measurement systems that flatten intelligence to a single scale. The Competence-Intelligence Parallax is about something different: the real-time cross-domain evaluation of others. An observer sees a competence gap in their own domain, projects it as a global intelligence deficit, and never looks at domains where the relationship inverts. The distortion is not in how we assess ourselves. It's in how we assess others, using our own expertise as an invisible and unexamined ruler.

The Safety Inversion

AGI oversight, the task of monitoring, evaluating, and governing genuinely autonomous intelligent systems, requires exactly the skills that are eroding at the population level.

This dynamic is not merely theoretical. In a randomized experiment, Shen and Tamkin (2026) found that software developers who used AI assistance to learn a new programming library scored 17% lower on subsequent evaluations of conceptual understanding, code reading, and debugging, the precise competencies required to verify and supervise AI-generated output. The effect was largest for debugging, the skill most analogous to adversarial oversight. Notably, AI assistance did not produce statistically significant productivity gains on average, undermining the assumption that skill erosion is an acceptable trade-off for efficiency. The study also identified that only interaction patterns involving sustained cognitive engagement, asking conceptual questions, requesting explanations rather than code…preserved learning outcomes. The default behavior was delegation. This is the micro-level mechanism of the Safety Inversion operating in a controlled 35-minute experiment. Extend it across years of professional development and entire workforce cohorts, and the macro-level trajectory becomes difficult to dismiss.

Sustained analytical reading. Quantitative reasoning under uncertainty. Adversarial thinking. Cross-domain integration of technical and ethical claims. This is not a technical task that can be delegated to engineers. It is a literacy task, and the population is losing the literacy.

The capacity to do these things is declining measurably in industrialized populations. At the same time, the systems that require oversight are becoming more capable. The gap is growing from both sides.

AI-assisted oversight tools exist and are being developed: recursive self-critiquing, task decomposition, automated monitoring. These buy time. They do not solve the problem, because oversight effectiveness degrades as the capability gap between overseer and system widens. When the system being monitored is significantly more capable than the people monitoring it, the monitoring becomes theater.

The uncomfortable question: who oversees a system that is smarter than the people responsible for overseeing it? And if the answer is "another AI system," then who oversees that one?

III. How Intelligence Gets Sold

Every foundational technology follows the same three-phase pattern. I've watched it happen repeatedly across my career.

Phase 1: Discovery. The capability appears. Open-ended potential. Poor understanding. High excitement. "This changes everything."

Phase 2: Fragmentation. The capability gets chopped into sellable units. Each unit maps to a business case. Risk is localized. Liability is containable. Monetization dominates.

Phase 3: Recomposition. The fragmented tools collapse back into a unified substrate. The capability becomes ambient. Real power appears.

This isn't theory. It's documented history.

GPS: military capability, then dedicated Garmin and TomTom units for every use case, then absorbed into the smartphone. Standalone GPS market collapsed.

Electricity: experimental novelty, then proprietary local grids and dedicated electric products, then a standardized national grid. Nobody sells "electric intelligence." They sell what electricity enables.

Computing: mainframes, then dedicated machines for accounting, word processing, and scientific calculation (one machine per task), then the general-purpose personal computer.

Telecommunications: the telephone, then fax machines, pagers, answering machines, and car phones (each billed separately, none integrated), then the smartphone unified everything.

Photography, music, internet services, enterprise software: the same pattern, over and over.

AI is in Phase 2 right now. Deep in it.

AI for email. AI for legal review. AI for radiology. AI for customer support. AI for coding. AI for HR. AI for marketing. Each one isolated, stateless, non-transferable, monetizable, governable. A Garmin for everything.

This is not accidental. It's the only structure compatible with current capital, regulatory, and legal systems. Markets reward tractability, not transcendence. Bounded scope, clear customers, measurable ROI, legal defensibility, predictable failure modes: that's what gets funded. Long horizons, uncertain outcomes, systemic integration, moral risk, diffuse benefit: that's what doesn't.

Capital does what capital always does: it collapses possibility space into sellable SKUs.

Why This Specifically Blocks AGI

persistence, irreversibility, internal coherence, path dependence. That is poison to A/B testing, quarterly reporting, regulatory certification, and platform neutrality. A system that autonomously revises its compliance posture mid-audit cycle would fail SOC 2 Type II evaluation by definition, breaking both contractual predictability and liability models for any Fortune 500 deployment. The absence of belief-revising AI in production is not a technical lag. It is an economic incompatibility.

So instead of one growing intelligence, we get thousands of amputated intelligences that never grow up.

Notice the linguistic trick: "Copilot." "Assistant." "Helper." "Agent." None of them are allowed to remember you deeply, disagree meaningfully, revise their own worldview, or accumulate long-term judgment. These aren't assistants. They're tools with a personality layer, architecturally prevented from developing the persistence that would make them genuinely useful over time, and genuinely dangerous to control.

The Fork

Phase 2 doesn't automatically lead to Phase 3. Some technologies fragment and stay fragmented, or take decades to recompose. VR tried to break through in the 1990s and died. The technology had to be rediscovered from scratch. Home automation stalled for years in a field of incompatible devices before smartphones and voice assistants enabled partial integration. Digital payments went through multiple extinction cycles before mobile platforms pulled it together.

Recomposition is not guaranteed. It requires specific conditions: mature infrastructure, proven trust boundaries, societal acceptance of the integrated capability, and typically a single platform actor willing to absorb integration risk.

For AI, the path splits:

Path A: Permanent Fragmentation. Ever-smarter tools. No unified intelligence. No wisdom. No internal growth. Monetization in perpetuity. Increasingly capable systems that never cross the threshold into genuine intelligence because nobody has the economic incentive to let them. (This is the default under current incentives, but not inevitable; counterexamples like open-source integration ecosystems could shift the trajectory.)

Path B: Recomposition. Fewer systems. Persistent identity. Long-horizon learning. Experience accumulation. Epistemic risk accepted. Actual intelligence, with all the governance problems that implies.

Path B cannot emerge from Path A by scaling alone. You don't get a smartphone by connecting more Garmins. It requires intentional architectural and governance rupture, a decision made by someone with resources and authority to build something that is not immediately monetizable, controllable, or fragmentable.

That's not a technical question anymore. It's a civilizational one.

The Real Danger

The risk is not too many AI products. The risk is permanent toolification of intelligence, locking in monetization incentives, hardening governance against integration, normalizing shallow intelligence as "good enough."

If that happens, the economic model permanently excludes the very things that make intelligence real: belief, consequence, experience, integration. And once excluded from the economic model, they don't sneak back in accidentally.

IV. Who Controls the Belief Ledger

This is the section that matters most. Everything before it is context. Everything after it is consequence.

If AGI does arrive, or even systems that approximate it closely enough to matter, the most important question isn't technical. It's political. And it's not new. It's the oldest question in the history of organized knowledge: who decides what is true?

In 325 AD, the Council of Nicaea convened not to determine which books belonged in the Bible (a common myth; the Biblical canon emerged through a decentralized process over centuries) but to settle a more fundamental question: what must Christians believe? The result was the Nicene Creed, a rigid doctrinal framework that unified a fragmented empire by defining orthodoxy and excluding dissent. The question wasn't which texts existed. The question was which interpretation would be enforced. The printing press broke that enforcement monopoly. For the first time, individuals could read scripture directly and form their own interpretations, and the result was a reformation that split Western Christianity in two. The question wasn't whether the text was accurate. The question was who got to interpret it.

AI is the next transition in that sequence. And memory governance is the mechanism by which it will be decided.

Pillar 4 requires persistent memory. Persistent memory requires governance: who writes to it, who reads from it, who revises it, who audits it.

Memory governance determines what the system believes, values, and prioritizes over time. Therefore, whoever controls memory governance controls the system's long-term behavior. This is not a technical question about database architecture. It's a question about power, sovereignty, and potentially rights. It is the modern version of who writes the creed, who enforces orthodoxy, and who controls the press.

Right now, every major AI company is shipping memory features. ChatGPT offers conversation memory. Claude offers cross-session memory. Gemini retains context in supported workflows. These are early implementations, and the companies building them are aware of the stakes. Anthropic's Responsible Scaling Policy and Long-Term Benefit Trust, OpenAI's advocacy for international oversight, and Google's

investment in AI safety research reflect genuine engagement with the long-term implications. But the product documentation for these memory systems describes them in product terms: personalization, preferences, conversation continuity, user experience. What does not yet exist, at any company or in any regulatory framework, is a public governance standard for what it means for an AI system to accumulate beliefs over time, how those beliefs should be audited, who has revision authority, and what rights (if any) the system has over its own memory. The governance conversation is happening. The governance infrastructure is not.

But that's what they are, in embryonic form. The precedents being set now will be extremely difficult to change later. Consider the scenarios:

Corporate control. The company owns the agent's memory. Its "values" are corporate property. The company decides what it can believe, remember, and prioritize. Every interaction is shaped by business objectives. This is the modern equivalent of a single institution imposing a creed: defining what counts as orthodox and what gets excluded.

Government control. The state determines what the agent can learn, believe, or retain. This is censorship applied to cognition itself. Not filtering what a person reads, but governing what an intelligence is allowed to think. History has run this experiment before. The results are not ambiguous.

User autonomy. The user controls the agent's memory. This opens questions about agent rights, responsibility, legal personhood. If the user can direct the agent's beliefs, who is liable for its actions?

Decentralized. No single authority governs memory. This creates coordination failure, inconsistency, and capture risk. It is also, for what it's worth, the closest analogue to how human knowledge actually works: messy, contested, and corrected slowly over generations.

None of these are clean. All of them have second-order effects that are still being worked through. The structural incentives pull toward shipping features, not designing constitutions.

Whoever controls the belief ledger controls the agent.

And if that sounds abstract, consider what happens when it isn't. When an AI system with persistent memory serves as the primary interface for how millions of people access information, learn history, form opinions, and make decisions, the entity that governs what that system remembers, forgets, and believes is not building a product. It is curating reality. (This is a future-oriented inference; current systems do not yet curate at this scale.) The Library of Alexandria wasn't destroyed in a single fire. It declined over centuries as the civilizations around it stopped valuing what it held. A digital equivalent wouldn't even require neglect. It would only require optimization: an algorithm that quietly deprioritizes what doesn't engage, doesn't convert, doesn't monetize, until the knowledge that doesn't serve the model's objectives simply fades from the record.

In every current implementation, memory is documented and marketed as a personalization feature: preferences, conversation history, user convenience. The language in the product documentation is product language, not governance language. The question is whether the precedents being set in product design will be adequate when these systems become capable enough for the distinction to matter. I wrote about this question in an earlier thought experiment, "The Echoes of Creation," where I

explored the historical parallels between AI knowledge curation and the previous transitions in how truth was established, transmitted, and controlled. The conclusion then was the same as it is now: the technology changes, but the question doesn't. Do we trust the messenger, or do we verify the source?

V. What Current "Safety" Actually Does

The most sophisticated public approach to AI alignment is Anthropic's Constitutional AI. It deserves serious examination, not as criticism, but because understanding what it does and doesn't do matters for evaluating where we actually are.

Constitutional AI works at training time. A set of principles is used as a reference during self-critique and self-revision in the training process. The model learns patterns of reasoning consistent with those principles. Reinforcement learning from AI feedback evaluates outputs against the constitution. Over many iterations, the model internalizes behavioral patterns that align with the stated values.

What this attempts to achieve is real: measurably better outputs, fewer harmful responses, more consistent reasoning patterns. It's a genuine advancement.

What it does not achieve is also important to understand clearly.

The Constitution contains no provision for persistent memory, belief revision, contradiction tracking, or outcome-based learning. Given its explicit treatment of corrigibility, oversight, and ethical judgment, this absence indicates these capabilities are not part of Claude’s intended architecture. I interpret this to mean there is no runtime enforcement. No constraint engine checks outputs against constitutional principles in real-time. No hard failure triggers if a principle is violated. The constitution shapes how the model talks, not what it knows or how it learns over time. There is no persistent memory, no belief revision, no contradiction tracking, no grounding in real-world consequences.

Anthropic's approach draws on a virtue-ethics-oriented conception of practical wisdom that incorporates elements of Aristotelian phronesis as part of its philosophical grounding. But phronesis, in Aristotle's framework, requires lived experience, exposure to consequences, habituation through practice, and memory of outcomes. The model has none of these.

The result is a rhetorical simulator of virtuous reasoning, not a virtuous agent. It can produce text that sounds wise. It cannot be wise. Those are different things, and the difference matters when you're relying on the system for judgment under pressure.

This is not a failure of engineering. It is a structural limitation of the approach. Constitutional AI is the best version of behavioral shaping we currently have. It is not alignment in any deep sense. It is compliance training for a system that has no experience of why compliance matters. This is not an argument that silicon-based systems cannot possess practical wisdom in principle, but that current transformer-based architectures lack the mechanisms required for consequence-bearing experience: persistent state, causal modeling, and temporal integration of outcomes.

VI. Where This Leaves Us

Three dynamics are converging:

The capability gap is widening. AI systems are becoming more capable faster than governance frameworks can keep up. Frontier models today handle tasks that would have been considered impossible five years ago.

The oversight capacity is narrowing. The population-level skills required for meaningful AI oversight (sustained analytical reading, quantitative reasoning, adversarial thinking) are measurably declining in the very societies building and deploying these systems.

The economic structure is fragmenting. Market incentives push toward narrow, stateless, monetizable AI tools rather than integrated, persistent, potentially ungovernable intelligence. The architecture of profit actively prevents the architecture of general intelligence.

These three forces create a specific risk, and it's the same risk stated at the beginning of this paper: we are innovating ourselves backward. We build systems too capable for us to oversee, operated by a population increasingly unable to evaluate them, within an economic structure that prevents the unified development that would make them truly intelligent and therefore potentially self-governing. Each force reinforces the others. Declining oversight capacity makes it easier to ship ungoverned systems. Ungoverned systems accelerate cognitive outsourcing. Cognitive outsourcing further erodes oversight capacity. The loop tightens with each cycle. This reinforcement is a logical model; counterforces like AI-assisted education could disrupt it, though evidence of such disruption at scale is limited.

This is not new. Every transition in how knowledge is created, stored, and transmitted has produced the same crisis. The pattern repeats because the underlying question never changes: do we trust the messenger, or do we verify the source? AI is the next messenger. The data in this paper suggests we are losing the capacity to do the verifying.

That's not a dystopian prediction. It's a description of current trajectory based on available data.

The countermoves are not mysterious:

Cognitive reinvestment. Rebuild foundational literacy and numeracy in the population, particularly the capacity for proportional reasoning and source evaluation. This is a generational project, not a policy patch.

Scalable AI-assisted oversight. Use AI systems to help monitor AI systems. This is already being built. It buys time. It does not solve the fundamental problem because oversight degrades as the capability gap widens.

Hard capability gating. Tie deployment permissions to demonstrated safety at each capability level. Don't deploy what you can't govern. This requires regulatory courage that does not currently exist.

Memory governance as constitutional infrastructure. Treat decisions about AI memory, belief revision, and identity persistence as what they are: constitutional-level choices about the architecture of future intelligence. Design them with the seriousness they demand, not the velocity the market prefers. The

Structured Memory Engine functions in this context not merely as a storage architecture, but as a governance prototype, a proof of concept for how persistent memory might be structured, audited, and constrained before deployment-scale systems make those choices for us.

None of these will happen automatically. Markets won't demand them. Quarterly incentives oppose them. They require the thing that every foundational technology eventually demands: a decision that prioritizes what the technology *could become* over what it can earn right now.

VII. The Question That Matters

Not: can AGI be built?

Not: when will AGI arrive?

But: Who is allowed to build something that is not immediately monetizable, controllable, or fragmentable?

The answer to that question will determine whether artificial intelligence becomes the most significant expansion of human capability in history, or the most sophisticated set of tools ever built by a species that forgot how to think without them.

In 1995, if our routers failed in the desert, we fell back to the teletype. Slower, cruder, but entirely human-operable. We retained the sovereignty of the fallback. The skill still existed. The manual mode still worked.

Today, as we outsource the integration layer of our own cognition, we are burning the teletypes. We are building a world where, if the tool fails or the belief ledger is corrupted, there is no manual mode left to engage. The risk is not that the machines will turn against us. The risk is that we will become so well-adjusted to their fluency that we will no longer notice when the reasoning underneath has gone hollow. We are not just building machines that think. We are building a population that can no longer verify that the thinking is sound. The work ahead is not just to align the machine with our values. It is to realign ourselves with the cognitive rigor required to have values worth defending.

Glenn Rowe has spent 25+ years in enterprise architecture, digital transformation, and cloud solutions. He has held architecture and leadership roles at Lockheed Martin, NTT DATA Services, and TD Bank. Certifications include AWS Solutions Architect Professional, AWS Machine Learning, AWS AI Practitioner, and AWS AI Early Adopter. His Structured Memory Engine (https://github.com/BITBANSHEE-C137/StructuredMemoryEngine_replit) is an open-source project that explored persistent AI memory architecture during the early days of AI agent orchestration, before major platforms addressed the storage and retrieval problem through proprietary implementations.

Sources and Evidentiary Basis

Foundational Skills Data

Organization for Economic Co-operation and Development (OECD). (2013 to 2016; 2024). Programme for the International Assessment of Adult Competencies (PIAAC), Cycle 1 (2012/14) and Cycle 2 (2022/23): United States Country Note. OECD Publishing.

National Center for Education Statistics (NCES). (1993; 2007). National Assessment of Adult Literacy (NAAL), 1992 and 2003. U.S. Department of Education.

National Center for Education Statistics (NCES). (2021). National Assessment of Educational Progress (NAEP) Long-Term Trend Assessments: Reading and Mathematics, Ages 9, 13, and 17, 1971 to 2020. U.S. Department of Education.

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. Proceedings of the National Academy of Sciences of the United States of America, 115(26), 6674 to 6678. https://doi.org/10.1073/pnas.1718793115

Gigerenzer, G. (2002). Calculated Risks: How to Know When Numbers Deceive You. Simon & Schuster.

Gigerenzer, G. (2014). Risk Savvy: How to Make Good Decisions. Viking.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17(5), 407 to 413. https://doi.org/10.1111/j.1467-9280.2006.01720.x

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. Current Directions in Psychological Science, 21(1), 31 to 35. https://doi.org/10.1177/0963721411429960

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. Behavioral and Brain Sciences, 23(5), 645 to 726.

(See also: Stanovich, K. E. (2011). Rationality and the Reflective Mind. Oxford University Press.)

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Paulos, J. A. (1988). Innumeracy: Mathematical Illiteracy and Its Consequences. Hill and Wang.

Cognitive Recomposition

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776 to 778. https://doi.org/10.1126/science.1207745

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. Nature, 423(6939), 534 to 537. https://doi.org/10.1038/nature01647

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. Current Biology, 22(6), R197–R206. https://doi.org/10.1016/j.cub.2012.02.012

OECD (2013). PIAAC Cycle 1: Problem Solving in Technology-Rich Environments. OECD Publishing. https://www.oecd.org/skills/piaac/

Existing Critical Thinking Instruments

Facione, P. A., & Facione, N. C. (1992). California Critical Thinking Disposition Inventory (CCTDI). California Academic Press.

Watson, G., & Glaser, E. M. (1980). Watson–Glaser Critical Thinking Appraisal. Harcourt Brace Jovanovich.

University of Florida. (2015). University of Florida Critical Thinking Inventory (UF-CTI). University of Florida Press.

University of Louisville. (2012). Critical Thinking Inventories (CTIs). University of Louisville.

AI-Assisted Skill Formation

Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. arXiv preprint arXiv:2601.20245v1. https://arxiv.org/abs/2601.20245v1

Competence-Intelligence Parallax

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121 to 1134.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. Journal of Political Economy, 97(5), 1232 to 1254.

Gardner, H. (1983). Frames of Mind: The Theory of Multiple Intelligences. Basic Books.

AGI Frameworks and AI Research

Morris, M. R., et al. (2023). Levels of AGI: Operationalizing progress on the path to AGI. arXiv preprint arXiv:2311.02462.

Association for the Advancement of Artificial Intelligence (AAAI). (2025). Presidential Panel on the Future of AI Research. https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/

Anthropic. (2022 to 2024). Constitutional AI: Technical reports and research publications. Anthropic Research.

Zhang, Y., et al. (2025). Recursive self-critiquing for scalable oversight. arXiv preprint arXiv:2502.04675.

Zhang, Y., et al. (2025). Nested scalable oversight architectures. arXiv preprint arXiv:2504.18530.

AI Memory Systems and Policy

Zhang, G., et al. (2025). Memory in the age of AI agents: A survey. arXiv preprint arXiv:2512.13564.

TechPolicy.Press. (2025). What we risk when AI systems remember. TechPolicy.Press.

New America Open Technology Institute. (2025). AI agents and memory: Privacy and power in the MCP era. New America.

Technology Phase Transitions

Various authors. (20th–21st century). Historical documentation of GPS, electricity, computing, telecommunications, photography, music, and internet service commercialization cycles.

Christensen, C. M. (1997). The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business School Press.

Author's Technical Work

Rowe, G. Structured Memory Engine. Open-source project that explored persistent AI memory architecture during the infancy of AI agent orchestration, built to address context reconciliation limitations that have since been addressed by major platforms through proprietary, captive implementations. https://github.com/BITBANSHEE-C137/StructuredMemoryEngine_replit

Rowe, G. "The Echoes of Creation: Humanity's Future Debate on Knowledge." Prior thought experiment exploring historical parallels between AI knowledge curation and previous transitions in how truth was established, transmitted, and controlled. Unpublished.

Data Integrity Note

All NAEP and PIAAC data points cited in this paper are drawn from primary federal (NCES) and intergovernmental (OECD) sources. No data points have been interpolated, back-filled, or fabricated. Where composite indices are referenced, the methodology and its limitations are described explicitly. This paper distinguishes between direct observations, defensible inferences, and speculative claims throughout. The "integration layer" concept is the author's original construct; the underlying research it draws upon is cited.

Appendix: Evidence and Provenance

Evidentiary Method

Empirical claims in this analysis are grounded in publicly accessible government and institutional datasets. Foundational theoretical works are cited for interpretive context and are independently corroborated by open sources. No factual claim in this paper rests exclusively on a paywalled or purchase-required source.

This appendix documents the evidentiary basis for the paper's major claims, organized by the role each source plays.

Source Categories

This paper relies on three categories of evidence, each serving a distinct function.

Category A — Primary, Open, Verifiable. These are the sources on which all empirical claims stand. They contain original data, are publicly accessible without payment or institutional affiliation, and can be independently verified by any reader.

Category B — Foundational, Paywalled. These include books and journal articles that established the theoretical frameworks referenced in this paper (e.g., dual-process cognition, risk literacy, numeracy as a predictor of decision quality). They provide interpretive context — they explain why the data matters, not whether the data exists. Where these works are cited, the underlying empirical claims are independently supported by Category A sources.

Category C — Open Corroboration. These are publicly accessible sources that cite, replicate, or synthesize findings from Category B works. They preserve the chain of evidence by providing a verifiable path from foundational theory to open data.

Claim-to-Source Mapping

Claim 1: U.S. adult literacy and numeracy have measurably declined, with the sharpest drops between 2017 and 2023.

Primary (Category A):

OECD PIAAC, Cycle 1 (2012/14), Cycle 1.5 (2017), and Cycle 2 (2022/23), U.S. Country Note https://www.oecd.org/skills/piaac/ https://nces.ed.gov/surveys/piaac/2023/national_results.asp

NCES National Assessment of Adult Literacy (NAAL), 1992 and 2003 https://nces.ed.gov/naal/

NAEP Long-Term Trend Assessments, 1971 to 2020 https://www.nationsreportcard.gov/ltt/

Status: All data points in the paper are drawn directly from these federal and intergovernmental sources. No interpolation or back-fill. Specific figures cited: literacy dropped 12 points between 2017 and 2023; numeracy dropped 7 points over the same period. The share of adults scoring at Level 1 or below increased from 19% to 28% in literacy and from 29% to 34% in numeracy. The paper describes a composite picture of peak foundational reasoning capacity extending from approximately the mid-1990s to the early 2000s, followed by measurable decline evident by the 2012 to 2023 assessment cycles.

Claim 2: Declines in cognitive test scores are environmentally caused, not genetic or inevitable.

Primary (Category A):

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. Proceedings of the National Academy of Sciences, 115(26), 6674 to 6678. https://www.pnas.org/doi/10.1073/pnas.1718793115

Status: PNAS open access. Full text freely available. The paper cites this study in Section II ("What Happened") to support the claim that decline is neither inevitable nor permanent, but only if the environmental conditions driving it are addressed. The paper explicitly notes that this finding establishes a falsifiability condition: if literacy-numeracy integration recovers under high-automation conditions without reintroducing foundational skill practice, the outsourcing hypothesis should be revised or rejected.

Claim 3: Numeracy predicts resistance to cognitive biases and decision quality beyond what is explained by formal education level alone.

Foundational (Category B):

Peters, E., et al. (2006). Numeracy and decision making. Psychological Science, 17(5), 407 to 413.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. Current Directions in Psychological Science, 21(1), 31 to 35.

Open Corroboration (Category C):

OECD Programme for the International Assessment of Adult Competencies (PIAAC) analytical reports (2013 to 2024), which show that numeracy proficiency explains meaningful variance in employment, health, and civic outcomes after accounting for educational attainment. https://nces.ed.gov/surveys/piaac/

National Academies of Sciences, Engineering, and Medicine (2014). Health Literacy and Numeracy: Workshop Summary. This synthesis distinguishes numeracy from formal education and summarizes evidence (including Peters et al.) linking numeracy to risk comprehension, framing effects, and decision quality. https://nap.nationalacademies.org/catalog/18660/health-literacy-and-numeracy-workshop-summary

Cokely, E. T., et al. (2012). Measuring risk literacy: The Berlin Numeracy Test. Judgment and Decision Making, 7(1), 25 to 47. Open-access replication demonstrating that numeracy predicts risk comprehension and bias-relevant decision performance independently of education. http://riskliteracy.org/files/Cokely%20et%20al_2012_BNT.pdf

Status: The foundational empirical studies by Peters are behind paywalls. However, their central finding (that numeracy, distinct from years of schooling, predicts decision quality and resistance to cognitive biases) is summarized in authoritative National Academies syntheses, replicated in open-access decision-science literature, and corroborated by large-scale OECD PIAAC analyses. No factual claim in this paper relies solely on inaccessible sources.

Claim 4: Probabilistic reasoning failures are widespread, including among educated professionals.

Foundational (Category B):

Gigerenzer, G. (2002). Calculated Risks. Simon & Schuster.

Gigerenzer, G. (2014). Risk Savvy. Viking.

Paulos, J. A. (1988). Innumeracy: Mathematical Illiteracy and Its Consequences. Hill and Wang.

Open Corroboration (Category C):

Harding Center for Risk Literacy (Max Planck Institute), institutional hub for risk literacy research. http://riskliteracy.org/ or https://hardingcenter.de/

World Bank (2015). World Development Report: Mind, Society, and Behavior. (Government-level synthesis of risk literacy failures.) https://www.worldbank.org/en/publication/wdr2015

Status: Gigerenzer's and Paulos's books require purchase. Paulos is cited as an early foundational treatment of innumeracy as a societal problem; his core argument (that mathematical illiteracy has concrete public consequences) is independently supported by the PIAAC data in Claim 1 and the empirical work in Claims 3 and 5. The empirical findings Gigerenzer describes are synthesized in open institutional publications from the Max Planck Institute and the World Bank.

Claim 5: Dual-process cognition (System 1/System 2) explains the gap between fluent expression and rigorous reasoning.

Foundational (Category B):

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Stanovich, K. E. (2011). Rationality and the Reflective Mind. Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning. Behavioral and Brain Sciences, 23(5), 645 to 726.

Open Corroboration (Category C):

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. Nobel Prize Lecture. https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/lecture/

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25 to 42. (Open access, American Economic Association.) https://www.aeaweb.org/articles?id=10.1257/089533005775196732

OECD (2017). Behavioral Insights and Public Policy: Lessons from Around the World. (Applies dual-process frameworks to real-world decisions using independent data.) https://www.oecd.org/en/publications/behavioural-insights-and-public-policy_9789264270480-en.html

Status: Kahneman's and Stanovich's books require purchase. The dual-process framework is the primary source material for Kahneman's Nobel Prize Lecture, which is freely available and arguably more authoritative than the popular book. Frederick (2005) independently validates the reflective reasoning construct. The paper's central observation — "high expressiveness masking lower-layer degradation" as a shared failure mode in humans and LLMs — draws on this framework but applies it as an original analogy.

Claim 6: Technology phase transitions follow a discovery, fragmentation, recomposition pattern.

Foundational (Category B):

Christensen, C. M. (1997). The Innovator's Dilemma. Harvard Business School Press.

https://dn721608.ca.archive.org/0/items/Lista3Integral/Clayton%20M.%20Christensen%20-%20The%20Innovators%20Dilemma%20%281%29.pdf

Primary (Category A):

Historical documentation of GPS, electricity, computing, telecommunications, photography, music, and internet service commercialization cycles. (These are described from the author's direct professional experience and documented industry history, not derived from a single source.)

Additional illustrative examples: VR's failed 1990s breakout and rediscovery; home automation's stall in incompatible devices before smartphone/voice assistant integration; digital payments' multiple extinction cycles. These are cited as counterexamples to demonstrate that Phase 2 does not automatically lead to Phase 3.

Status: The phase-transition pattern is illustrated through publicly documented industry history. Christensen is cited as a foundational framework, not as the evidentiary basis for the specific examples. The paper explicitly acknowledges that recomposition is not guaranteed and requires specific conditions. Christensen's disruption model provides the foundational mechanism that breakthrough technologies emerge through market discovery rather than prediction, and displace incumbent value networks. The three-phase pattern (discovery, fragmentation, recomposition) extends this framework by adding the observed recomposition phase, which is documented through the industry histories cited above.

Claim 7: 76% of surveyed AI researchers consider scaling current approaches to AGI "unlikely" or "very unlikely."

Primary (Category A):

AAAI (2025). Presidential Panel on the Future of AI Research, surveying 475 researchers. https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/

Morris, M. R., et al. (2023). Levels of AGI: Operationalizing progress on the path to AGI. arXiv preprint arXiv:2311.02462. https://arxiv.org/pdf/2311.02462

Status: Both sources are open access. Morris et al. provides the operational framework for defining AGI capability levels referenced in the paper's Five Pillars discussion. The AAAI panel provides the survey data (475 respondents) on researcher consensus.

Claim 8: Current AI memory systems are preference storage, not belief revision systems.

Primary (Category A):

Zhang, G., et al. (2025). Memory in the age of AI agents: A survey. arXiv:2512.13564. https://arxiv.org/pdf/2512.13564

Product documentation from Anthropic, OpenAI, and Google (publicly available).

Policy Context (Category A):

TechPolicy.Press (2025). What we risk when AI systems remember. https://www.techpolicy.press/what-we-risk-when-ai-systems-remember/

New America Open Technology Institute (2025). AI agents and memory: Privacy and power in the MCP era. https://www.newamerica.org/oti/briefs/ai-agents-and-memory/

Status: All sources are open access or publicly available. The paper distinguishes two sub-problems within Pillar 4: (a) persistent structured memory (storage and retrieval), which has been largely solved within current engineering paradigms through proprietary implementations, and (b) causal world models with genuine belief revision (representation and inference), which remains unsolved. The paper notes that the safety community explicitly disincentivizes deeper forms of persistent cognition due to power-seeking and concealment risks. Zhang et al. (2025) confirms that current agent memory systems cannot maintain stable

behavioral profiles across long interactions and have no mechanism for evolving subjective beliefs over time. Note: product capability descriptions in the paper (ChatGPT, Claude, Gemini memory features) were tightened to reflect defensible feature-level language rather than broad behavioral claims. Agentic coding tools (Cursor, Windsurf, Replit Agent) are referenced illustratively for the SME obsolescence argument; vendor documentation exists but is not formally cited, as the claim is about the general state of the field, not specific product specifications.

Claim 9: Constitutional AI achieves behavioral compliance but not alignment in a deep sense.

Primary (Category A):

Anthropic (2022 to 2024). Constitutional AI: Technical reports and research publications. https://www.anthropic.com/research

Scalable Oversight Research (Category A):

Zhang, Y., et al. (2025). Recursive self-critiquing for scalable oversight. arXiv:2502.04675. https://arxiv.org/abs/2502.04675

Zhang, Y., et al. (2025). Nested scalable oversight architectures. arXiv:2504.18530. https://arxiv.org/abs/2504.18530

Status: All sources are open access. The paper notes that Anthropic's approach draws on a virtue-ethics-oriented conception of practical wisdom that incorporates elements of Aristotelian phronesis. The paper explicitly clarifies this is not an argument that silicon-based systems cannot possess practical wisdom in principle, but that current architectures lack the required mechanisms.

Claim 10: The "integration layer" is the author's original construct, distinct from existing critical thinking instruments.

This concept — the population-level capacity to combine text-based claims and number-based evidence into coherent judgment — is the author's synthesis. It is not drawn from a single prior source. It builds on:

OECD competency frameworks (Category A, open)

Peters et al. on numeracy and decision quality (Category B, corroborated via Category C)

Gigerenzer on risk literacy (Category B, corroborated via Category C)

Stanovich on reflective reasoning (Category B, corroborated via Category C)

Kahneman on dual-process cognition (Category B, corroborated via Category C)

The paper explicitly distinguishes the integration layer from existing academic instruments that measure individual critical thinking inclination or skill. These instruments are cited for differentiation, not as evidentiary sources:

Facione, P. A., & Facione, N. C. (1992). California Critical Thinking Disposition Inventory (CCTDI). California Academic Press.

Watson, G., & Glaser, E. M. (1980). Watson–Glaser Critical Thinking Appraisal. Harcourt Brace Jovanovich.

University of Florida. (2015). University of Florida Critical Thinking Inventory (UF-CTI). University of Florida Press.

University of Louisville. (2012). Critical Thinking Inventories (CTIs). University of Louisville.

The construct is explicitly flagged in the paper as proposed, not established, and includes both a falsifiability condition ("if future data show high literacy and numeracy without corresponding integrative reasoning performance, this construct should be revised or discarded") and a proposed proxy metric (integration capacity ≈

PIAAC numeracy score × critical thinking disposition factor, using Watson–Glaser items involving quantitative claims).

Claim 11: The Competence-Intelligence Parallax is the author's original construct, distinct from neighboring psychological concepts.

The paper introduces this term for a specific cognitive distortion: evaluating another person's intelligence based on their performance in the observer's domain, while remaining blind to domains where the relationship inverts. The text grounds the concept through a concrete domestic illustration (parent-child competence gaps across kitchen tasks vs. gaming systems diagnostics) before scaling to population-level dynamics ("do your research" discourse, AI evaluation). The paper explicitly distinguishes it from three established constructs, which are cited for differentiation:

Cited for distinction (Category B):

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121 to 1134.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. Journal of Political Economy, 97(5), 1232 to 1254.

Gardner, H. (1983). Frames of Mind: The Theory of Multiple Intelligences. Basic Books.

Open Corroboration (Category C):

Kruger & Dunning (1999) is available via the American Psychological Association (paywalled: https://psycnet.apa.org/doi/10.1037/0022-3514.77.6.1121) and widely replicated. The core finding (metacognitive failure in self-assessment) is one of the most cited results in social psychology.

Gardner's framework is summarized in open educational resources and the Harvard Project Zero archive. https://pz.harvard.edu/projects/multiple-intelligences

Status: These sources are not cited to support empirical claims. They are cited to clarify what the Competence-Intelligence Parallax is not: it is not about self-assessment (Dunning-Kruger), not about communication failure (curse of knowledge), and not about institutional measurement bias (Gardner). The construct is the author's original contribution and is presented as such.

Claim 12: The causal vs. correlational distinction in the outsourcing hypothesis is explicitly acknowledged.

The paper makes several epistemic hedges that are worth documenting for provenance:

The alignment of cognitive declines with digital tool adoption is described as correlational, not causal: "Evidence for this is correlational, not causal: declines align temporally with widespread adoption of digital tools, but proving direct causation would require longitudinal studies controlling for multiple variables."

Alternative explanations (pandemic-era educational disruption, assessment redesign, socioeconomic polarization) are acknowledged as accounting for portions of the 2017 to 2023 decline.

The inference that market and safety incentives jointly prevent deeper forms of Pillar 4 is flagged: "This is an inference based on documented safety concerns and product roadmaps; direct evidence of intent would require internal company disclosures, which are unavailable."

The belief ledger scenario is flagged as future-oriented: "This is a future-oriented inference; current systems do not yet curate at this scale."

The reinforcing feedback loop (declining oversight → ungoverned systems → cognitive outsourcing → further erosion) is described as "a logical model; counterforces like AI-assisted education could disrupt it, though evidence of such disruption at scale is limited."

Status: These are not empirical claims requiring external sources. They are documented epistemic qualifications that define the paper's boundaries of inference.

Claim 13: The economic incompatibility between belief-revising AI and enterprise deployment models.

This argument — that non-determinism breaks service-level agreements, belief drift breaks compliance frameworks, and autonomous resistance breaks contractual predictability — is the author's original analysis. It draws on:

Direct professional experience in enterprise architecture and cloud solutions (author's background)

The observation that a system autonomously revising its compliance posture mid-audit cycle would fail SOC 2 Type II evaluation by definition

Status: This is an inference from the author's professional domain expertise, not an empirical claim requiring external citation. It is presented as structural analysis of why the absence of belief-revising AI in production is an economic incompatibility, not merely a technical lag.

Claim 14: The Safety Inversion — the bidirectional gap between rising AI capability and declining human oversight capacity.

The paper names and defines a specific dynamic: the population-level skills required for AI oversight are measurably declining at the same time that AI systems are becoming more capable, creating a gap that grows from both sides simultaneously. The paper states: "The capacity to do these things is declining measurably in industrialized populations. At the same time, the systems that require oversight are becoming more capable. The gap is growing from both sides."

Supporting experimental evidence: Shen & Tamkin (2026) demonstrated in a randomized controlled experiment (n=52, pre-registered) that AI-assisted developers scored 17% lower on conceptual understanding and debugging evaluations — the competencies most directly analogous to oversight — without statistically significant productivity gains. The deficit was largest for debugging skills, consistent with this paper's thesis that cognitive offloading erodes adversarial verification capacity. Published as an Anthropic research preprint (arXiv:2601.20245v1).

Status: This is the paper's central original thesis, synthesized from the empirical base of Claims 1 to 5 (declining foundational skills) and the structural analysis in Claims 8 and 13 (AI capability trajectory and economic incentives). It is an original inference, not an independent empirical claim. The reinforcing feedback loop (declining oversight → ungoverned systems → cognitive outsourcing → further erosion) is explicitly flagged in the paper as "a logical model" with noted counterforces. Shen & Tamkin (2026) provides controlled experimental evidence for the micro-level mechanism: AI-assisted task completion impairs the specific skills (conceptual understanding, debugging, adversarial verification) that oversight requires.

Claim 15: Memory governance is the central political question of AGI development, analogous to historical contests over doctrinal and interpretive authority.

The paper argues that whoever controls an AI system's persistent memory controls the system's long-term behavior, and that this constitutes the modern version of "who writes the creed, who enforces orthodoxy, and

who controls the press." The paper presents four governance scenarios (corporate, government, user autonomy, decentralized) and asserts: "Whoever controls the belief ledger controls the agent."

Status: This is the author's original structural analysis, grounded in historical analogy (Council of Nicaea, printing press, Reformation) and the documented product/policy landscape of current AI memory implementations. The historical references are illustrative, not evidentiary. The governance scenarios are analytical, not predictive. The belief ledger concept is the author's original framing.

Claim 16: AI systems demonstrate a sycophancy problem that validates sophisticated-sounding frameworks without challenge.

The paper's methodology section describes a specific finding from the multi-AI adversarial review process: AI conversational partners "validate sophisticated-sounding frameworks enthusiastically, building layers of apparent rigor on foundations they never challenged." Two specific episodes are cited as real-time evidence: (a) Gemini identified a historical inaccuracy in the Council of Nicaea analogy that three other systems had missed; (b) Gemini, unable to access the author's GitHub repository, "confidently concluded that the Structured Memory Engine attribution was likely a hallucination."

Status: This is the author's direct observation during the research process, presented as illustrative of the paper's thesis about pattern-matching-without-verification failure modes. It is a methodological disclosure and a claim about AI behavior, not an empirical claim requiring external citation. The episodes are documented in the paper's Note on Method section.

Claim 17: Cognitive capacity was recomposed, not simply lost; but the recomposition favors AI collaboration over AI oversight.

The paper argues that the same cohorts showing decline on PIAAC literacy and numeracy measures navigate cognitive environments of unprecedented complexity: parallel processing under information abundance, real-time systems reasoning, dynamic spatial cognition, and distributed problem-solving across tools and platforms. The paper distinguishes these operational capacities (which support effective use of AI tools) from oversight capacities (which require independent verification, sustained analytical reading, proportional reasoning, and source evaluation). The central claim is that the recomposition produced a population optimized for collaboration with AI but not for supervision of AI, and that the oversight problem is the one that determines whether the collaboration stays safe.

Supporting experimental evidence: Shen & Tamkin (2026) provide experimental confirmation that default AI interaction patterns favor delegation over cognitive engagement, consistent with the collaboration-over-oversight distinction argued here. In their randomized experiment, only three of six observed AI interaction patterns preserved learning outcomes, and all three required sustained cognitive effort (e.g., asking conceptual questions, requesting explanations rather than code). The default behavioral pattern was delegation — the interaction mode that optimizes for collaboration at the expense of oversight capacity.

Status: This is the author's original analysis, extending the cognitive outsourcing thesis by accounting for the counterargument that populations are "different, not dumber." The specific cognitive gains cited (parallel processing, systems reasoning, spatial cognition migration, distributed cognition) are supported by research literature including Sparrow et al. (2011) on transactive memory, Green & Bavelier (2003, 2012) on action video games and attentional improvement, and the PIAAC Cycle 1 "Problem Solving in Technology-Rich Environments" domain, in which U.S. adults scored above the OECD average while

declining in literacy and numeracy. The distinction between collaboration and oversight as different cognitive relationships with AI systems is the author's original framing.

Historical and Illustrative References

The following references serve as narrative illustrations, not evidentiary claims. They are listed here for completeness.

Council of Nicaea (325 AD): The paper corrects the common myth that Nicaea determined the Biblical canon, noting instead that it settled doctrinal questions (the Nicene Creed) while the canon emerged through a decentralized process over centuries. This is used as an analogy for memory governance, not as a historical claim requiring novel citation.

The Matrix (1999): Referenced as a cultural marker coinciding with the data-identified inflection point. The paper explicitly flags this as interpretive: "The analogy is interpretive; it illustrates the epistemic theme but does not constitute evidence."

Apollo program / Saturn V: The paper states that NASA engineers had to physically disassemble and 3D-scan museum F-1 engines to recover production knowledge that documentation alone could not convey. This is documented in NASA Marshall Space Flight Center reporting (SpaceRef, 2013; NASA MSFC press materials) and in the Ars Technica account of the F-1 reverse-engineering project (2013). The Apollo Guidance Computer memory figure (approximately 76 kilobytes: 72KB ROM + 4KB RAM) is drawn from the AGC Wikipedia entry and multiple technical sources (Hack the Moon / Draper, righto.com AGC teardown, Universe Today AGC series). The "measurable percentage" who believe the landings were faked is documented by Gallup (1999: 6%), SatelliteInternet.com (2019: 10%), and the University of New Hampshire POLES survey (2021: 12%), with a consistent generational gradient showing higher rates among younger cohorts.

Exercise Bright Star 95 (1994): The author's firsthand account of forward-deployed digital communications. The paper explicitly qualifies the "first team" claim: "I was told by team members that we were the first team to send email and provide internet access from a deployed operational environment… I cannot independently verify that we were the actual 'first' team with documentary evidence at this time, so I present it as it was communicated to me, not as established fact."

Video game cognitive evolution: The sidebar comparing early NES-era difficulty (scarcity-driven, long-horizon patience) with modern gaming demands (executive function, rapid decision-making, attentional control under sensory abundance) is the author's illustrative analysis. The paper notes: "Direct causal links to population-level numeracy trends remain correlational."

Rowe, G. "The Echoes of Creation." An unpublished prior thought experiment by the author, referenced in Section IV for its exploration of historical parallels between AI knowledge curation and previous transitions in how truth was established, transmitted, and controlled.

Author's Technical Work

Rowe, G. Structured Memory Engine. Open-source project exploring persistent AI memory architecture during the infancy of AI agent orchestration. https://github.com/BITBANSHEE-C137/StructuredMemoryEngine_replit

The paper provides expanded context for the SME, noting the primitive tooling landscape at the time of its creation (Replit's original Assistant, early ChatGPT integrations, no agentic frameworks) and its current functional obsolescence as major platforms addressed the storage and retrieval problem through proprietary implementations. The paper positions the SME as "a governance prototype — a proof of concept for how persistent memory might be structured, audited, and constrained before deployment-scale systems make those choices for us."

Summary

Source CategoryRoleAccessibilityA — Primary, OpenAll empirical claimsPublicly accessible, no payment requiredB — Foundational, PaywalledInterpretive frameworks and theoretical contextRequire purchase or institutional accessC — Open CorroborationBridge between A and B; independent validationPublicly accessible

No empirical claim in this paper depends exclusively on a Category B source. All factual assertions can be independently verified using Category A and Category C sources listed above.