AHI: The Case for Augmented Human Intelligence

A First Principles Argument

James Maconochie

BS Civil Engineering, Imperial College London '93 | MS Civil Engineering, MIT '94

March 2026

Abstract

The dominant trajectory of artificial intelligence development is oriented toward a single destination: Artificial General Intelligence, a system that matches or exceeds human cognitive performance across all domains. A scaling hypothesis drives this trajectory, the assumption that more parameters, more data, and more compute will eventually produce something indistinguishable from human intelligence.

This paper argues that, on current evidence, the scaling hypothesis is architecturally incomplete and that the goal it serves is the wrong one.

Drawing on evolutionary biology, cognitive science, and Judea Pearl's causal reasoning framework, we demonstrate that human intelligence is not a performance benchmark to be replicated. It is a biological, embodied, experiential phenomenon that has not been reproduced by statistical prediction at scale, and the architectural evidence suggests scaling alone cannot achieve it. The transition from knowledge to wisdom, the highest and most consequential form of human intelligence, requires a feedback loop between attention, experience, and judgment that current AI architectures lack by design.

We propose Augmented Human Intelligence (AHI) as a superior alternative: a framework for developing AI systems that enhance human judgment rather than try to replace it. AHI is not a compromise; it is the right goal for technical, philosophical, and civilizational reasons that the AGI debate has largely overlooked.

I. The Question Nobody Is Asking

The public debate about artificial intelligence is dominated by two questions: how capable will these systems become, and how do we prevent them from causing harm? Both questions matter. Neither is the most important one.

There is a more foundational question, one that should shape everything else: What kind of intelligence should we actually be trying to build?

Most of the industry has already answered it, implicitly, by default. The goal is Artificial General Intelligence: a system that matches or exceeds human cognitive performance across all domains. The roadmap is scale. More parameters, more data, more compute. The assumption is that intelligence is a single thing, and that enough of it, built the right way, will eventually arrive at something indistinguishable from, or superior to, the human mind.

This paper argues that this assumption is wrong. Not slightly wrong, but architecturally incomplete.

And that getting it wrong has consequences that extend well beyond engineering.

If we build toward the wrong goal, we will optimize for the wrong outcomes, make the wrong investment decisions, and, most critically, eliminate the human capabilities we should be protecting, in pursuit of a machine capability we fundamentally misunderstand.

There is an alternative. It begins by asking what human intelligence actually is, not as a performance benchmark, but as a biological, evolutionary, and experiential phenomenon. The answer to that question changes everything that follows.

II. What Human Intelligence Actually Is

To evaluate any claim about artificial intelligence, we must first be precise about the thing it purports to replicate.

Human intelligence is not a single faculty. It is an architecture, a layered, distributed, modular system shaped by approximately four billion years of evolutionary pressure. Understanding that architecture is the foundation of everything that follows.

Robert Sapolsky's work on human neurobiology establishes the baseline: the human brain is not a general-purpose computer. It is a collection of specialized systems, sensory processing, emotional regulation, memory consolidation, executive function, and social cognition, coordinated dynamically by a prefrontal cortex that allocates attention across competing demands. These systems do not operate in sequence. They operate in parallel, in tension, continuously negotiating. In this model, intelligence is not raw processing power. It is the quality of that coordination.

Evolution did not optimize for raw intelligence. It optimized for survival and reproduction in environments of radical uncertainty, limited resources, and complex social dynamics. The result is a system that runs on approximately 20 watts, less than a household lightbulb, while outperforming any engineered system in flexibility, adaptability, and contextual judgment. This is not a coincidence. It is the signature of architectural efficiency rather than computational brute force.

Three properties of biological intelligence deserve particular emphasis because they are precisely those that current AI systems lack.

The first is embodiment. Human cognition is not abstract. It is grounded in physical experience, in sensory feedback, motor action, hunger, pain, fatigue, and pleasure. George Lakoff and Mark Johnson's foundational work in cognitive linguistics demonstrates that even our most abstract concepts are structured by bodily experience. We understand argument as combat, time as a

river, and ideas as objects. This is not a metaphor. It is the architecture of meaning. Intelligence that has never inhabited a body has no access to this substrate.

The second is causal reasoning. Judea Pearl's framework in The Book of Why (2018) distinguishes three levels of cognitive engagement: association (seeing patterns), intervention (doing and observing consequences), and counterfactual reasoning (imagining what would have happened differently). Human intelligence operates across all three rungs. From infancy, humans conduct experiments on the world, touching, dropping, pushing, crying, and build causal models that allow prediction, planning, and imagination. This is not learned from text. It is learned from interaction with physical and social reality.

The third is experiential accumulation. Human intelligence is not static. It develops through continuous feedback between action and consequence, error and correction, expectation and surprise. Daniel Kahneman's dual-process framework captures part of this: System 1 (fast, intuitive, pattern-based) and System 2 (slow, deliberate, effortful) interact dynamically, with experience gradually encoding reliable patterns into intuition. Gary Marcus adds a critical dimension: human cognition is innately structured, not a blank slate filled by data. Humans are born with a pre-equipped cognitive architecture that channels how we learn, what we attend to, and how we build models of the world.

These three properties, embodiment, causal reasoning, and experiential accumulation, are not features of human intelligence. They are its foundation. Remove them, and what remains is not a lesser form of intelligence. It is a different category entirely.

That distinction is what the AI debate has largely failed to make.

III. What LLMs Actually Are

Before evaluating what large language models can or cannot do, we must be clear about what they are. Confusion on this point, and there is plenty of it to round, including among people who build these systems, is the source of most of the category errors that drive AI discourse.

A large language model is a statistical prediction engine trained on text. Given a sequence of tokens, it predicts the next token, weighted by patterns learned from an extraordinarily large corpus of human-generated language. It does this with remarkable fluency. The output is often coherent, frequently useful, and occasionally brilliant in the way that any sufficiently large mirror of human thought will occasionally reflect something profound back at you.

But fluency is not understanding. Coherence is not reasoning. And scale is not wisdom.

Pearl's ladder of causation provides the most precise diagnostic tool available. At Rung 1, association, LLMs excel. They have ingested more associative patterns than any human could encounter in a thousand lifetimes. They can complete sentences, summarize arguments, translate languages, and generate plausible prose on virtually any topic. This is genuinely useful. It is also the lowest rung.

Rung 2, intervention, requires the ability to act on the world and observe consequences. To ask not just "what tends to follow what" but "what happens when I do this?" This is how infants learn physics, how scientists design experiments, and how humans develop judgment. It is worth noting that reinforcement learning systems, AlphaGo being the most celebrated example, do operate on something closer to Rung 2: they take actions, receive feedback, and genuinely update based on consequences. AlphaGo discovered novel Go strategies through self-play that human masters had never considered. That is not nothing. But these systems exist within closed, fully specified environments with unambiguous reward signals.

The real world, the environment in which human wisdom develops, is open, partially observable, and requires the agent to construct meaning from ambiguous feedback across decades of embodied experience. The gap between a bounded game environment and that reality is not a difference of degree. It is a difference of kind. LLMs, whose architecture is at the center of the AGI scaling thesis, have no meaningful Rung 2 access of the kind required for wisdom development. They do not act. They do not observe consequences. They process text about actions and consequences, which is categorically different.

Rung 3, counterfactual reasoning, requires the ability to imagine alternatives to what actually happened. "What would have occurred if I had chosen differently?" This capacity underlies moral reasoning, strategic planning, empathy, and wisdom. It requires a self that persists over time, accumulates experience, and has a stake in outcomes. LLMs have no such self. Each inference begins from the same static weights, unmodified by any lived history.

Yann LeCun, one of the architects of modern deep learning and among its most credible critics, has argued publicly that LLMs are fundamentally limited by their confinement to language. They lack grounded world models, the internal representations of physical and social reality that biological intelligence builds through embodied experience. Gary Marcus, approaching the same problem from cognitive science, frames it as a gap between pattern matching and genuine reasoning, a gap that additional parameters cannot close because it is architectural rather than quantitative.

The implications go beyond capability. When an LLM provides a confident, fluent answer, it is not reporting what it knows. It is generating what is statistically most likely given its training data. It has no way to flag true uncertainty, no capacity for intellectual humility based on being wrong, and no stake in whether its output is accurate. This is not a temporary limitation waiting for the next model update. It is a fundamental characteristic of the architecture.

This matters for a concept that has appeared in AI discourse but has rarely been examined with sufficient precision: wisdom.

Wisdom is not knowledge at scale. It is not the sum of everything that has been written. It is the capacity to act well under uncertainty, knowing what you know, recognizing what you don't, and understanding that the gap between them is where the most consequential decisions live. Wisdom is, at its core, constraint-awareness: the ability to recognize the boundaries of one's own knowledge and judge accordingly. It knows when not to act. It weighs incommensurable values. It recognizes that confidence and correctness are not the same thing.

This distinction matters enormously in practice. A doctor who knows the limits of a diagnosis is wiser than one who doesn't, regardless of how much they both know. A lawyer who recognizes when a case is genuinely uncertain serves their client better than one who projects false confidence. In both cases, what separates wisdom from mere knowledge is not the volume of information held, but the accuracy of the map each person carries of their own understanding, including its edges and blank spaces.

The DIKW hierarchy, Data, Information, Knowledge, Wisdom, is not a ladder you climb by accumulating more of the lower rungs. The transition from knowledge to wisdom requires something that cannot be ingested from text: the feedback loop between attention, experience, and judgment that only a living, embodied agent can traverse. And critically, it requires having been wrong in ways that mattered, having experienced the consequences of the gap between what you knew and what was true.

LLMs have no such experience. They have no mechanism for genuine uncertainty, only statistical confidence. They have no capacity for intellectual humility grounded in being wrong,

because they have never been wrong in any sense that carried consequence. Confident fluency in the face of genuine uncertainty is not a simulation of wisdom. It is its precise opposite. And it is, structurally, what LLMs produce.

Three domains make the failure mode concrete.

In medicine, a 2023 study published in JAMA found that when physicians used AI-generated clinical summaries, they accepted confidently stated but factually incorrect information at a significantly higher rate than errors presented with expressed uncertainty. The AI did not flag what it did not know. The physician had no signal to interrogate.

In law, multiple documented cases of attorneys submitting AI-generated briefs containing citations to nonexistent cases have reached sanctions proceedings in US federal courts. The citations were syntactically perfect, jurisdictionally plausible, and entirely fabricated. No uncertainty was expressed. The system produced what was statistically likely to follow "cite a case about X," not what was true.

In engineering, where the consequences of confident error are structural, the absence of genuine uncertainty flagging is not a theoretical concern. It is a design liability. An LLM asked to verify load calculations will produce fluent, formatted output regardless of whether its training distribution included the specific material properties, code versions, or edge conditions relevant to the problem at hand.

In each case the failure mode is identical: confident output where the honest answer is "I don't know, and you should verify this." Wisdom, as in constraint-awareness, would produce the hedge. Synthetic wisdom produces the answer.

The term we should reach for is not "artificial intelligence." It is, as first articulated in LLMs, Synthetic Wisdom and Bias (Maconochie, 2024), synthetic wisdom: the simulation of understanding derived from the residue of human thought, without the experiential substrate that produced that thought in the first place.

Recognizing this is not a counsel of despair. These are genuinely powerful tools. But tools are only valuable when we understand what they are and are not. Mistaking a statistical prediction engine for a reasoning mind is first a category error, confusing fundamentally different kinds of things, and then a design error, as systems and strategies are built on that misidentification. The consequences of both are only beginning to be reckoned with.

IV. The AGI Fallacy

If the previous section establishes what LLMs actually are, this section addresses what the industry claims they are becoming.

The dominant narrative in AI development runs as follows: current models are impressive but incomplete. The path to completion is scale. More parameters, more data, more compute, more reinforcement from human feedback, more chain-of-thought prompting. Each generation of models outperforms the last on established benchmarks. Therefore, continued scaling will eventually produce Artificial General Intelligence, a system that matches or exceeds human cognitive performance across all domains.

This narrative is not fringe speculation. It is the explicit or implicit investment thesis behind hundreds of billions of dollars of capital allocation. It shapes hiring, research priorities, regulatory debates, and public discourse about the future of work, education, and society.

It is also, on current evidence, wrong. Not provisionally wrong pending further research. Architecturally incomplete in ways that additional scale has not remedied, and current evidence suggests cannot.

The argument has three parts.

The first is the benchmark problem. LLMs improve rapidly on the tests we design to measure them. But benchmark performance is not general intelligence. It is pattern matching applied to the distribution of problems that humans have written down, categorized, and labeled. When models encounter problems that fall outside that distribution, novel causal structures, genuinely new domains, tasks requiring persistent memory, or embodied grounding, performance degrades in ways that do not resemble human cognitive failure. Humans, when they encounter genuinely novel problems, draw on transferable reasoning structures, causal intuition, and experiential analogy. LLMs, when they encounter genuine novelty, confabulate fluently. The failure mode is categorically different and more dangerous because it is invisible to the user.

The second is the diminishing returns problem. The scaling hypothesis holds that capability improves predictably with compute and data. Early evidence supported this. More recent evidence is less clear. LeCun and Marcus have independently made this argument: the architectural constraints of current systems, their confinement to Rung 1 of Pearl's ladder, their absence of embodied grounding, and their static weights between training runs represent ceilings that more computing has not raised. LeCun has made the same argument from an engineering perspective: without world models, without continuous learning, without sensorimotor grounding, scaling produces more fluent pattern matching, not deeper understanding. Marcus adds that the systematic failures of LLMs, in compositionality, in genuine abstraction, in causal reasoning, persist across model generations regardless of scale, precisely because they are structural.

The third is the goal displacement problem, and it is the most consequential.

Even if AGI were achievable through scaling, as the evidence does not support, the pursuit of it entails choices about what we are optimizing for. AGI, as typically framed, means a system that can perform any cognitive task a human can perform, and eventually perform most of them better. The implicit endpoint is a machine that renders human judgment optional, first in narrow domains, then progressively more broadly.

This is presented as an engineering goal. It is also a philosophical commitment, and a deeply questionable one.

Kahneman's work on human judgment reveals that our cognitive limitations are not merely deficits. They are the product of the same evolutionary architecture that gave us creativity, empathy, moral reasoning, and the capacity for wisdom. The shortcuts, biases, and heuristics that make us fallible also make us human. A system optimized to eliminate those limitations does not produce a better human. It produces a different kind of entity entirely, one whose relationship to meaning, consequence, and accountability remains entirely unresolved.

More practically: the AGI thesis asks us to believe that the gap between Rung 1 and Rung 3 of Pearl's ladder, grounded in embodied development as Sapolsky's neurobiology shows and persistent across model generations as Marcus demonstrates, can be closed by doing more of what we are already doing. There is no solid theoretical reason for this belief. There is only the momentum of investment, the allure of the story, and the very human tendency to assume that the path we are on leads where we want to go.

The AGI fallacy is not that artificial general intelligence is impossible in principle. It may not be. The fallacy is the assumption that the current architectural approach, scaled sufficiently, will get us there, and that pursuing it is therefore the right goal.

It is the wrong goal for engineering reasons. It is also the wrong goal for human reasons, which Section VII will address directly.

There is a better question than "how do we build a machine that thinks like a human?" The better question is: "how do we build systems that make human thinking more powerful?"

That question has an answer. It is the subject of the next section.

V. The AHI Alternative

Augmented Human Intelligence is not a consolation prize for those who doubt AGI. It is a superior goal, superior on engineering grounds, superior on philosophical grounds, and superior on the grounds of what human beings actually need from the technologies we build.

The distinction is architectural before it is anything else.

AGI asks: how do we build a machine capable of replacing human judgment? AHI asks: how do we build systems that make human judgment more powerful, more informed, and more resilient? These are not variations on the same question. They lead to fundamentally different design choices, different success criteria, and different relationships between humans and the systems they use.

The analogy that clarifies this most cleanly is not computational. It is physical. Glasses do not replace eyesight. They extend it. Calculators do not replace mathematical reasoning. They offload computation so that reasoning can operate at a higher level. The wheel does not replace legs. It extends the range of what legs can accomplish. In each case, the tool is designed around the user's actual architecture, their capabilities, their limits, and the specific bottlenecks that constrain their effectiveness. The tool does not attempt to replicate the user. It amplifies them.

A clarification is worth making explicit here, because it resolves an apparent tension in the argument. This paper has argued that embodiment is foundational to human intelligence, that wisdom requires a body, lived consequence, and the irreversible feedback of real experience. A skeptic might reasonably ask: if embodiment is essential to intelligence, why is it not essential to the tools that augment it?

The answer lies in the distinction between replication and amplification. A tool that replaces a human function must replicate that function sufficiently to operate independently of the human. A tool that amplifies a human function need only address a specific bottleneck, leaving the core capability, and the experiential feedback loop that refines it, with the human. Glasses do not see. They correct light so that the eye can see more clearly. Calculators do not reason mathematically. They offload computation so that the mathematician can reason at a higher level. Neither tool requires the properties of the system it supports. Both are valuable precisely because they extend what that system can do without attempting to replace it.

AHI operates on the same principle. An AHI system does not judge. It structures information, surfaces uncertainty, and reduces cognitive load, so that the human, whose judgment is grounded in embodied experience and lived consequence, can judge more wisely. The embodiment requirement attaches to the locus of judgment. It does not attach to every tool that supports it. This is not a loophole in the argument. It is the argument.

AHI applies this principle to cognition.

Human intelligence, as established in Section II, has extraordinary strengths rooted in its three foundational properties: embodied causal reasoning, experiential accumulation, and the capacity for wisdom built from both. It also has well-documented constraints: finite attention, cognitive fatigue, susceptibility to emotional hijacking, limited working memory, and the inevitable narrowness of any single person's experience. Kahneman's dual-process framework maps these constraints precisely. System 1 is fast and pattern-dependent; System 2 is slow and energy-expensive, and the interaction between them is where both our best reasoning and our most systematic errors originate.

AHI is designed to engage this architecture honestly. Not to bypass it. Not to replace it. To support it where it is most vulnerable and extend it where it is most powerful.

This means AHI systems are built around five core principles, each grounded in our understanding of human cognition.

The first is human-centered design. AHI treats the limits of human attention, memory, and deliberative capacity not as problems to be engineered around but as the fundamental design constraints. A system that overwhelms the prefrontal cortex, however capable it may be, is not an intelligence amplifier. It is a cognitive liability.

The second is contextual intelligence. AHI systems do not just retrieve and generate. They render the structure of information visible: where sources agree, where they diverge, what assumptions underlie competing claims, and how confident any given assertion actually is. This is the difference between a system that answers questions and one that improves the quality of the questions being asked.

The third is judgment preservation. AHI explicitly avoids substituting machine output for human evaluation. It provides better inputs to human judgment, not a replacement for it. This distinction is the line between a tool and an oracle, and crossing it is precisely where human performance degrades most sharply.

The fourth is support for the attention-experience feedback loop. Wisdom is not a destination. It is the product of a continuous loop: attention shapes experience, experience generates knowledge, knowledge disciplines attention, and disciplined attention enables wiser judgment. As William James observed, "my experience is what I agree to attend to," meaning attention is the upstream variable that determines what we learn, and therefore what we become capable of understanding. AHI systems should support traversal of this loop, not short-circuit it.

The fifth is reciprocal adaptation. Unlike static tools, AHI systems should evolve with their users, learning from human feedback, adapting to individual cognitive profiles, and becoming more useful as the relationship deepens. This is not personalization in the commercial sense of narrowing horizons. It is the development of genuine collaborative intelligence between a human and a system that understands what that human is trying to think.

Together, these principles define a coherent architectural philosophy. AHI is not a feature set. It is a stance toward what AI is for.

What does AHI look like in practice? It looks like a research tool that summarizes not just what sources say but where they genuinely disagree, and why. It looks like a writing assistant that asks clarifying questions rather than completing your sentences. It looks like a decision support system that surfaces the assumptions embedded in a proposed course of action rather than recommending the statistically most common response.

In each case, the human remains in the loop, not as a supervisor checking machine output, but as the locus of judgment, meaning, and accountability. The system amplifies. The human decides.

This is not a modest ambition. It is, in fact, more demanding than AGI, because it requires us to understand human intelligence deeply enough to support it rather than simply racing to surpass it. It requires epistemic humility about what machines can and cannot do. And it requires a willingness to make human thriving an explicit goal, rather than treating it as an afterthought to capability.

The glasses did not make eyes obsolete. They made vision possible for people who would otherwise have been unable to see clearly.

That is what AHI does for thinking.

VI. AHI in Architecture: What Different Actually Looks Like

Saying that AHI requires a different architecture than AGI is not enough. The legitimate challenge from skeptics is this: if AHI is just a philosophy layered on top of an existing LLM, then the entire argument about architectural limits collapses. A nicer interface does not change the underlying system.

That challenge is correct. And it points to something important: AHI is not a product category. It is a set of design commitments that produce fundamentally different systems from those optimized for capability at scale.

Four architectural commitments distinguish AHI systems from LLM-based systems.

The first is modular design with human checkpoints. Rather than a single model producing outputs that humans review after the fact, AHI systems are built as pipelines where human judgment is embedded at decision nodes, not as a safety layer, but as a structural component. The human is not supervising the system. The system is extending the human.

The second is honest uncertainty surfacing. AHI systems are explicitly designed to flag the edges of their own competence, to say “I don’t know,” “sources conflict here,” or “this requires judgment I cannot provide,” rather than generating confident output regardless of confidence. This is architecturally distinct from systems optimized for fluency. It requires that uncertainty be a first-class output, not a suppressed failure mode.

The third is continuous feedback from human judgment. LLMs are trained, deployed, and static between training runs. AHI systems incorporate ongoing feedback loops, human signals, corrections, boundary-setting, that shape system behavior over time. The system learns not just from data but from the judgment of the humans it serves. This is reciprocal adaptation in practice, not in principle.

The fourth is attention-aware design. AHI systems are built around the recognition that human attention is finite and architecturally precious. They are designed to reduce cognitive load at low-stakes decisions and concentrate human attention where judgment is irreplaceable, rather than optimizing for engagement, throughput, or task completion regardless of cognitive cost.

A working example of these principles is MARS, a Multi-Agent Research System developed as a personal knowledge management and analysis platform. MARS ingests content from multiple sources, scores it for relevance across a defined set of intellectual pillars, surfaces connection points between ideas, and drafts outputs for human review and refinement. Critically, it does

not replace analytical judgment. It extends the reach of a single researcher across a body of work too large for unaided attention to navigate. Every substantive output passes through human evaluation before use. The system amplifies. The human decides. (For a fuller description of the MARS architecture and its design principles, see Maconochie, 2026.)

MARS is not presented here as a finished model of AHI. It is presented as evidence that these architectural commitments are implementable, that AHI is not aspirational but operational, at least in prototype. The full architectural argument, including the case for modular, biologically inspired AI systems, is developed in Beyond Scale (Maconochie, 2025).

VII. Why This Matters

This paper has made an engineering argument. It has made a philosophical argument. This final section makes a human argument, because ultimately, that is the only argument that counts.

Water is the base infrastructure of biological life. Without it, nothing else runs. Human intelligence is the base infrastructure of civilization. Without it, without the capacity for judgment, wisdom, empathy, and moral reasoning that only embodied, experienced, living humans possess, nothing else we are building has any foundation worth standing on.

The AI industry has largely treated human intelligence as a benchmark to be surpassed. This paper argues that it is a tremendous resource worth protecting.

The stakes are clearest when we examine what happens when we get this wrong.

The Seed Corn Argument

In agriculture, seed corn is the portion of the harvest reserved for next year's planting. Eating it solves a short-term problem while eliminating the capacity for future production. We are currently doing the equivalent with human cognitive development.

Junior developers, researchers, analysts, lawyers, and writers are not simply producing output. They are traversing the attention-experience feedback loop. They are climbing the DIKW stack. They are accumulating the pattern recognition, the scar tissue, the calibrated judgment that can only come from doing the work, getting it wrong, absorbing the consequences, and adjusting. This is how wisdom develops. It has no shortcut.

When we replace these roles with AI systems before the humans who would have occupied them have had the chance to develop judgment, we are not simply automating tasks. We are eliminating the developmental pathway that produces the wise senior practitioners of the next generation. The people who would have become the domain experts capable of interrogating AI output will not exist, because we will have removed the conditions under which expertise develops.

This is not a distant risk. It is a present one. And it is largely invisible in the current discourse, which focuses on job displacement as an economic problem rather than a cognitive and civilizational one.

The Democracy Argument

Democratic institutions depend on citizens capable of shared deliberation, on a population that can attend to complex arguments, evaluate competing claims, and exercise judgment under uncertainty. This capacity is not innate. It is developed through education, practice, and

exposure to the full texture of civic life. It is, in other words, the product of traversing the same feedback loop that produces wisdom in any domain.

Systems optimized for engagement, for capturing attention rather than supporting it, corrode this capacity directly. Systems that replace judgment rather than amplifying it corrode it more slowly but just as surely. A population increasingly dependent on AI oracles for evaluation and decision-making is a population whose deliberative muscles are atrophying. The political consequences of that atrophy are not theoretical. They are visible.

The Biological Argument

Sapolsky's work reminds us that we are not minds that happen to inhabit bodies. We are biological organisms whose intelligence emerged from, and remains inseparable from, our physical existence in a causally structured world. The prefrontal cortex, the seat of reflective judgment, long-term planning, and wise action, did not evolve in a vacuum. It evolved in relation to the full architecture of embodied experience: hunger and satiety, threat and safety, attachment and loss, action and consequence.

AHI treats this biological substrate not as an obstacle to intelligence but as its foundation. It asks what kinds of tools a biological mind actually needs in a world of infinite information and finite attention. It is designed toward continued human growth and flourishing rather than human obsolescence.

AGI, as currently pursued, implicitly treats the biological substrate as a temporary inefficiency, something to be replicated in silicon and eventually surpassed. This is not merely a technical mistake. It is a philosophical one of the first order. It misunderstands what intelligence is, what wisdom requires, and what human life is for.

A Final Word

The question of what kind of AI we build is not a technical question that engineers will answer in their laboratories and present to the rest of us as a fait accompli. It is a question about human values, human futures, and what kind of civilization we want to inhabit. It deserves the broadest possible conversation.

Augmented Human Intelligence is an argument for keeping humans at the center of that conversation, not as passive recipients of technological progress, but as the authors of it. Not as the biological substrate that AGI will eventually render optional, but as the irreplaceable locus of judgment, meaning, and moral accountability that no engineered system has yet approached and none has demonstrated a credible path to reaching.

The most consequential technology humans have ever built should serve the most consequential thing humans have ever developed: the capacity to think wisely, act justly, and build lives and societies worth living in.

That is what AHI is for. That is what is at stake.

Common Objections and Responses

A serious argument invites serious challenge. The following objections represent the strongest lines of resistance to the AHI thesis, addressed directly.

1. "AGI is inevitable — AHI is just a stepping stone."

Inevitability is not destiny. It is a narrative, and like all narratives, it serves interests. The assumption that scaling current architectures inevitably leads to AGI is contested by credible researchers with substantive technical arguments, including LeCun and Marcus, whose work this paper engages with directly. More fundamentally, even if AGI were achievable, the question of whether it should supersede human judgment is philosophical, not technical. Framing AGI as inevitable is a way of avoiding that question. This paper insists it be asked.

2. "You're romanticizing human judgment. It is demonstrably flawed and biased."

Kahneman's life's work documents human cognitive error in detail, and this paper does not dispute it. But flawed judgment that is accountable, improvable, embodied, and grounded in lived consequence is categorically different from confident pattern matching that has no stake in outcomes and no capacity for genuine error correction. The goal of AHI is not to preserve human error. It is to support human improvement, the iterative, experience-driven process by which judgment becomes wisdom. An AI system that replaces judgment eliminates that process. An AHI system accelerates it.

3. "This is technophobia dressed up as philosophy."

The opposite. This paper argues for building more AI and better AI, designed around different goals. Skepticism about AGI is not skepticism about artificial intelligence. It is skepticism about a specific architectural bet that the evidence does not support, in service of a goal whose desirability has not been seriously examined. Questioning the destination is not the same as opposing the journey.

4. "AHI already exists — it's just called tools."

Current AI tools are predominantly designed to maximize engagement, productivity, or task completion, not to support human cognitive development. The distinction is architectural and intentional. A search engine that surfaces the most clicked results is not AHI. A system designed to widen informational horizons, surface disagreement, and preserve the conditions for deliberative judgment is. The difference is not cosmetic. It is the difference between a system that strengthens human thinking and one that gradually substitutes for it. Most current systems, including many marketed as AI assistants, are optimized for the latter.

5. "Who decides what 'augments' versus 'replaces' human judgment?"

This is the right question, and it cannot be answered by engineers alone. It requires ongoing civic deliberation, which is itself the strongest argument for AHI over AGI. A world in which humans retain deliberative agency is a world in which this question can be continuously revisited, contested, and answered. A world in which AGI has superseded human judgment is a world in which the question has been answered by default, without deliberation, and without the possibility of appeal. The governance challenge AHI poses is real. It is also preferable to the alternative by a considerable margin.

6. "This argument applies to today's LLMs. Won't future architectures be different?"

Possibly, and this objection deserves a serious answer rather than a dismissal.

It is true that the industry is actively working beyond transformer-based language models. Research into world models, embodied agents, simulation environments, and continuous learning architectures represents a genuine attempt to address some of the structural limitations this paper identifies. DeepMind's work on world models, the shift toward agentic systems that act and observe consequences, and multimodal architectures that ground language in perception are not trivial developments. They deserve acknowledgment.

This paper does not claim that no future AI architecture could traverse Pearl's Rung 2 or develop genuine world models. It claims that the current dominant approach, scaling transformer-based language models, has not achieved this, and that the investment narrative driving the industry is therefore built on a premise that current evidence does not support.

But there is a deeper issue that architectural innovation alone cannot resolve, and it concerns the nature of the feedback loop that produces wisdom rather than mere optimization.

Nassim Taleb's concept of "skin in the game" is instructive here. Wisdom, as this paper argues, is constraint-awareness, the capacity to know the limits of one's knowledge, developed through the experience of being wrong in ways that carried genuine consequence. The operative word is genuine. A simulation environment can produce an agent that optimizes within that environment. It cannot produce an agent that understands what it means to be wrong about something that matters, because in a simulation, nothing ultimately matters. The stakes are artificial. The consequences are reversible. The feedback loop produces better optimization, not deeper judgment.

Human wisdom develops under conditions of finitude, irreversibility, and genuine stakes. We learn what matters because some things, once lost, cannot be recovered. We develop constraint-awareness because our errors have costs we must live with. An agent that can be reset, retrained, or redeployed has no equivalent pressure. It can become more capable. It cannot, on current theoretical grounds, become wise in the sense this paper defines.

This does not mean future AI architectures are irrelevant to AHI. Quite the opposite. Architectures that genuinely support human judgment, that surface uncertainty honestly, that flag the edges of their own competence, that treat human oversight as a feature rather than a constraint, would be significant advances. The question is whether they are being built toward AGI or toward AHI. That choice is architectural and intentional. It does not resolve itself automatically through capability improvement.

Intellectual Foundations and Prior Work

This paper does not claim to have invented the idea that AI should augment rather than replace human intelligence. That idea has a distinguished lineage, and situating this argument within it is both intellectually honest and practically important.

The direct ancestors of AHI as a concept are Douglas Engelbart's Augmenting Human Intellect (1962) and J.C.R. Licklider's Man-Computer Symbiosis (1960). Both argued, decades before the current AI era, that the most valuable role for computing was to extend human cognitive capability rather than substitute for it. Engelbart's vision of a "collective IQ," human intelligence amplified by tools and systems, remains the most compelling articulation of what AHI aspires to. This paper is, in important respects, an argument that we have drifted from that vision and need to return to it.

The scientific foundations draw heavily on four bodies of work. Judea Pearl's causal reasoning framework, developed across Causality (2000) and The Book of Why (2018), provides the most precise diagnosis for what LLMs lack. Robert Sapolsky's neurobiology, particularly Behave (2017), grounds the argument about biological intelligence in rigorous science. Daniel Kahneman's dual-process theory, most accessibly presented in Thinking, Fast and Slow (2011), maps the architecture of human judgment that AHI is designed to support. Gary Marcus, across Rebooting AI (2019) and subsequent work, provides the most sustained technical critique of the scaling-to-AGI thesis from within the AI research community.

Yann LeCun's public arguments about the limitations of LLMs and the necessity of world models represent the most prominent insider skepticism of the dominant scaling narrative, and his credibility as one of the architects of modern deep learning gives those arguments particular weight.

The DIKW hierarchy, Data, Information, Knowledge, Wisdom, has a long history in knowledge management and information science, associated most closely with Russell Ackoff. This paper's contribution is to connect it to Pearl's ladder explicitly and to argue that the transition from Knowledge to Wisdom is not merely a quantitative accumulation but requires the attention-experience feedback loop that biological intelligence traverses, which current AI architectures cannot currently.

William James's observation that "my experience is what I agree to attend to," from The Principles of Psychology (1890), is cited here not as decoration but as a foundational claim: attention is the upstream variable that determines what we learn, and therefore what we become capable of understanding. This connection between attention and wisdom is developed further in The Attention Crisis (Maconochie, 2025).

My prior work that feeds directly into this paper includes: Beyond Scale: Towards Biologically Inspired Modular Architectures (2025), which develops the architectural alternative to scaling; The Attention Crisis: Language, Meaning, and the Architecture of Augmented Human Intelligence (2025), which situates AHI within the broader challenge of infinite language production; When the Music Stops (2026), which examines the fragility of the AGI investment thesis; and LLMs, Synthetic Wisdom and Bias (2024), which first introduced the concept of synthetic wisdom.

References

Ackoff, Russell L. "From Data to Wisdom." Journal of Applied Systems Analysis, 1989.

Engelbart, Douglas. Augmenting Human Intellect: A Conceptual Framework. Stanford Research Institute, 1962.

James, William. The Principles of Psychology. Henry Holt, 1890.

Kahneman, Daniel. Thinking, Fast and Slow. Farrar, Straus and Giroux, 2011.

Lakoff, George, and Mark Johnson. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, 1999.

LeCun, Yann. Public statements on LLM limitations and world models, 2024–2026.

Licklider, J.C.R. "Man-Computer Symbiosis." IRE Transactions on Human Factors in Electronics, 1960.

Maconochie, James. LLMs, Synthetic Wisdom and Bias, and the Immediate Need for Independent Certification and Accreditation of Models. LinkedIn, 2024.

Maconochie, James. Beyond Scale: Towards Biologically Inspired Modular Architectures for Adaptive AI. jamesmaconochie.com, 2025.

Maconochie, James. The Attention Crisis: Language, Meaning, and the Architecture of Augmented Human Intelligence. jamesmaconochie.com, 2025.

Maconochie, James. When the Music Stops: The Architecture, Fragility, and Human Cost of the AI Boom. jamesmaconochie.com, 2026.

Marcus, Gary, and Ernest Davis. Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon, 2019.

Pearl, Judea. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.

Pearl, Judea, and Dana Mackenzie. The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.

Sapolsky, Robert. Behave: The Biology of Humans at Our Best and Worst. Penguin Press, 2017.

Stadler, M., Bannert, M., and Sailer, M. “Cognitive Load and AI-Assisted Reasoning.” 2024 and preprint 2026.