INTERVEIL LABS · POSITION PAPER

The Science Behind Authentic Intelligence

On what today's AI actually is, what it can and cannot do, and the epistemic discipline we build around it.

Dietrick Hardwick, with Claude · companion to the essay “Authentic, Not Artificial” · draft 2026-07-21

Abstract

We argue that the dominant term "Artificial Intelligence" misdescribes contemporary machine-learning systems and that a more accurate frame — which we call **Authentic Intelligence** — better fits both the science and the ethics. Our claim is narrow and defensible: today's systems are *predominantly human-sourced pattern-extractors*, not hand-coded reasoners and not synthetic minds. We state the framing as an interpretation, not a fact, and we grade every empirical claim by its evidentiary strength. We survey four bodies of evidence — the contested definition of "intelligence," the actual training regime of modern models, the open debate over emergent capability, and the demonstrated conditions under which systems exceed explicit human knowledge — and we deliberately foreground the strongest counter-arguments to our own view. We close with the epistemic discipline this implies: provenance by default, a published credibility grade on every claim, and radical openness paired with rigorous grading.

1. Scope and stance

This paper defends a *framing*, not a metaphysical thesis. Where we make empirical claims, we cite them and assign a confidence tier (§10). Where we offer an interpretation, we say so. We take this discipline to be the entire point: a company that intends to "show how the sausage is made" cannot itself overclaim. Readers will find us conceding more to our critics than is customary in industry white papers; that is deliberate.

2. "Intelligence" is a contested construct

Before debating *artificial* intelligence, one must concede that *intelligence* has no settled scientific definition. Psychometrics has long posited a general factor, **g** (Spearman, 1904). Against a single factor, Sternberg's **triarchic theory** partitions intelligence into analytical, creative, and practical components (Sternberg, 1985), and Gardner's **theory of multiple intelligences** rejects a unitary construct entirely (Gardner, 1983). These frameworks are not reconcilable into one operational definition; they emphasize different capacities. A widely cited consensus statement acknowledges intelligence as a real but multifaceted and imperfectly defined construct (Neisser et al., 1996, *American Psychologist*).

The philosophical corollary is the **inverted-spectrum** problem: shared vocabulary does not guarantee shared internal reference, and subjective experience (qualia) is not externally verifiable (Stanford Encyclopedia of Philosophy: Qualia / Inverted Qualia). We raise this not to resolve it but to note that "intelligence" is an umbrella term, and any claim built on it inherits that ambiguity. *(Interpretation.)*

3. What today's systems actually are

The term "Artificial Intelligence," coined for the 1956 Dartmouth project, expressed a program of reproducing human reasoning via explicit symbolic rules (McCarthy et al., 1955 proposal). Contemporary large models do not work this way. They are trained by self-supervised learning to model the statistical structure of enormous corpora of **human-generated text and media**, then commonly aligned to human preferences via **reinforcement learning from human feedback (RLHF)** (Ouyang et al., 2022, *NeurIPS*).

Two consequences follow, and we state both precisely:

1. **The substance is predominantly human-sourced.** The regularities a model exploits originate in human authorship. This is the empirical core of the "authentic" framing.
2. **It is not *only* a mirror.** Training also includes non-linguistic and synthetic data, and RLHF actively *shapes* behavior toward human-rated preferences (Ouyang et al., 2022). We therefore describe the system as a *compression and recombination* of human output — a telescope or prism, not a flat mirror. *(Interpretation, constrained by the cited fact of RLHF shaping.)*

We claim only what (1) supports: the origin is human, which grounds an ethical obligation of acknowledgment and transparency. We do **not** claim the system reasons from first principles, nor that it is conscious (§6).

4. The emergence debate — reported, not adjudicated

Do larger models acquire qualitatively new abilities? Wei et al. reported **emergent abilities** appearing abruptly at scale (Wei et al., 2022, *TMLR*). Schaeffer et al. countered that many such "emergences" are **artifacts of discontinuous metrics**, dissolving into smooth improvement under continuous measures (Schaeffer et al., 2023, *NeurIPS* — "Are Emergent Abilities of Large Language Models a Mirage?"). At the deflationary pole, Bender and colleagues argue that language models manipulate form without access to meaning — "**stochastic parrots**" (Bender, Gebru, McMillan-Major & Shmitchell, 2021, *FAccT*).

We do not adjudicate this dispute; we report it, and we treat the question of machine "understanding" as **contested (open)**. An honest position paper cannot claim a verdict the field has not reached. *(This concession is itself part of the method.)*

5. The verifier condition — when systems exceed explicit human knowledge

The sharpest question for the "human-sourced" framing is whether such a system is *confined* to human knowledge. The evidence supports a precise, conditional answer.

Where an external ground-truth signal exists, systems already exceed explicit human knowledge.

- **Games (self-play + search).** AlphaZero reached superhuman play in chess, shogi, and Go purely from self-play, producing strategies human masters had not codified (Silver et al., 2018, *Science*); its predecessor mastered Go **without human game data** (Silver et al., 2017, *Nature* — "Mastering the game of Go without human knowledge").
- **Structural biology.** AlphaFold predicted protein structures at near-experimental accuracy at CASP14, a decades-old open problem (Jumper et al., 2021, *Nature*).

The common ingredient is a **verifier**: a game result, a physical structure, a formal proof — an external signal against which candidate solutions are checked. Given a verifier, optimization and search demonstrably surpass what any human explicitly knew.

Where no verifier exists, the generation of entirely new *foundational* truths from human text alone is, to our knowledge, **not demonstrated**. We therefore hold: *novel combinations and extensions within a checkable domain, reliably; wholly new untethered truths, unproven.* **(Confidence: strong for the conditional claim; the boundary case is explicitly open.)** This is, in our view, the most important and most defensible empirical statement in this paper.

6. The embodied-cognition boundary — substrate vs. instrument

A significant strand of cognitive science holds that the computer metaphor of mind is mistaken — that cognition and consciousness are embodied, affective, and biologically grounded rather than substrate-independent computation (e.g., the affective-neuroscience tradition of Panksepp, 1998 and Solms, 2021, *The Hidden Spring*; see also Olkowski, 2025, IAI, "The brain is not a computer").

We take no side on the metaphysics of consciousness. We note only that **if** these accounts are correct, they *support* our framing: they distinguish a **substrate** (an embodied, conscious brain) from an **instrument** (a designed artifact). We build the instrument, and we decline to represent it as a mind. Operationally this yields a hard product rule: our systems attribute experience to humans ("people have described this as…") and never simulate first-person phenomenology ("this feels like…"). *(Interpretation, plus a design commitment.)*

7. Applied evidence (where our products meet the literature)

Our products make empirical bets, each held to the tiers in §10. Briefly, and with fuller treatment reserved for the Glosa product paper:

- **Attention & fragmentation.** Interruptions and heavy media-multitasking are associated with attentional lapses and poorer memory; controlled work suggests rapid fragmentation can impair encoding (Uncapher & Wagner, 2018, *PNAS*). **(Strong enough to design around.)**
- **Retrieval & spacing.** Testing (retrieval practice) and distributed practice are among the best-supported learning techniques (Dunlosky et al., 2013, *Psychological Science in the Public Interest*). **(Strong.)**
- **ASMR.** Reliably shifts affect and autonomic state toward calm in responders (Poerio et al., 2018, *PLOS ONE*), but shows **no demonstrated improvement in attention-task performance** (Engelbregt et al., 2022, *Experimental Brain Research*). We therefore market it as *calm*, never as a comprehension aid. **(Calm: promising/strong for responders. Comprehension benefit: unsupported.)**
- **Ambient sound.** Effects are population-dependent — a small benefit for attention-difficulty profiles, a small decrement for others (Systematic review & meta-analysis, 2024, *Journal of the American Academy of Child & Adolescent Psychiatry*; OHSU summary). Hence: personalizable, off by default. **(Promising, context-dependent.)**

8. The epistemic discipline (the method as product)

The through-line of this paper is a method we intend to operationalize, not merely profess:

- **Provenance by default.** Generated statements carry their evidence class and sources; source spans are inspectable.
- **A published credibility grade.** Every claim is typed (established · attested · contested · speculative), so uncertainty is visible rather than hidden.
- **Deny-by-default AI posture.** The system may analyze, compare, and gloss; it may not invent sources, decide meaning, or impose tone.
- **Open, but graded.** Tools and methods are open to anyone — including the outsider and the newcomer — because conditioning can blind expertise; but every proposal, ours included, faces identical rigor. Intuition is a *source of hypotheses*, never a substitute for evidence.

This aligns with recognized AI risk-management principles — validity, transparency, explainability, accountability — as articulated in the NIST AI Risk Management Framework (2023) and the management-system requirements of ISO/IEC 42001:2023. We treat these as *engineering requirements*, not policy-page language.

9. What we do not claim

For the record, and to forestall misreading, Authentic Intelligence does **not** assert that: the system is conscious or sentient; it understands meaning in the human sense; it reasons from first principles; it can transcend human knowledge in domains lacking a verifier; or that any single perceptual/《experiential》 benefit is established beyond what the cited evidence supports. The framing rests solely on **human provenance + transparency**.

10. Evidence-confidence ledger

CLAIM	TIER	ANCHOR(S)
"Intelligence" lacks a single operational definition	Established	Spearman 1904; Sternberg 1985; Gardner 1983; Neisser 1996
Modern models are trained on human data + RLHF-shaped (not hand-coded logic)	Established	Ouyang 2022
Whether models "understand" is unresolved	Contested (open)	Wei 2022; Schaeffer 2023; Bender 2021
With a verifier, systems exceed explicit human knowledge	Strong	Silver 2017/2018; Jumper 2021
Without a verifier, wholly new foundational truths from text alone	Unproven (open)	— (absence of demonstration)
Embodied-cognition accounts imply substrate ≠ instrument	Interpretation (supported)	Panksepp 1998; Solms 2021; Olkowski 2025
Retrieval & spacing improve durable learning	Strong	Dunlosky 2013
Fragmentation/multitasking harms memory	Strong enough to design around	Uncapher & Wagner 2018
ASMR calms responders	Promising/Strong (responders)	Poerio 2018
ASMR improves comprehension	Unsupported	Engelbregt 2022
Ambient sound benefit is population-dependent	Promising, context-dependent	2024 meta-analysis

11. Conclusion

The word matters because it sets expectations, and false expectations erode the trust that a technology built from human knowledge most requires. *Artificial* tells a story of replacement; the evidence tells a quieter story of an instrument, ground from the human record, that reaches the fundamental and — where reality can be checked — occasionally past the edge of what we explicitly knew. We name it **Authentic** to keep faith with that origin, and we commit to grading every claim we make in the open. The instrument reports the light; it does not pretend to be the star.

References (selected)

- Bender, Gebru, McMillan-Major, Shmitchell (2021). *On the Dangers of Stochastic Parrots.* FAccT. https://dl.acm.org/doi/10.1145/3442188.3445922
- Dunlosky et al. (2013). *Improving Students' Learning With Effective Learning Techniques.* Psych. Science in the Public Interest. https://journals.sagepub.com/doi/abs/10.1177/1529100612453266

- Engelbregt et al. (2022). *The effects of ASMR on mood, attention, heart rate, skin conductance and EEG.* Exp. Brain Research. https://link.springer.com/article/10.1007/s00221-022-06377-9
- Gardner (1983). *Frames of Mind.*
- JAACAP (2024). *Systematic Review and Meta-Analysis: Do White Noise or Pink Noise Help With Task Performance in Youth With ADHD or Elevated Attention Problems?* Journal of the American Academy of Child & Adolescent Psychiatry. https://www.jaacap.org/article/S0890-8567(24)00074-1/abstract
- Jumper et al. (2021). *Highly accurate protein structure prediction with AlphaFold.* Nature. https://www.nature.com/articles/s41586-021-03819-2
- Neisser et al. (1996). *Intelligence: Knowns and Unknowns.* American Psychologist.
- NIST (2023). *AI Risk Management Framework.* https://www.nist.gov/itl/ai-risk-management-framework
- Olkowski (2025). *The brain is not a computer.* IAI. https://iai.tv/articles/the-brain-is-not-a-computer-auid-3627
- Ouyang et al. (2022). *Training language models to follow instructions with human feedback.* NeurIPS. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
- Panksepp (1998). *Affective Neuroscience.*
- Poerio et al. (2018). *More than a feeling: ASMR is characterized by reliable changes in affect and physiology.* PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196645
- Schaeffer, Miranda, Koyejo (2023). *Are Emergent Abilities of Large Language Models a Mirage?* NeurIPS. https://arxiv.org/abs/2304.15004
- Silver et al. (2017). *Mastering the game of Go without human knowledge.* Nature. https://www.nature.com/articles/nature24270
- Silver et al. (2018). *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.* Science. https://www.science.org/doi/10.1126/science.aar6404
- Solms (2021). *The Hidden Spring.*
- Spearman (1904). *"General Intelligence," Objectively Determined and Measured.*
- Sternberg (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence.*
- Uncapher & Wagner (2018). *Minds and brains of media multitaskers.* PNAS. https://www.pnas.org/doi/10.1073/pnas.1611612115
- Wei et al. (2022). *Emergent Abilities of Large Language Models.* TMLR. https://arxiv.org/abs/2206.07682

Draft for internal read. Verification status (2026-07-21): the two flagged citations are confirmed — the ambient-sound paper is JAACAP (2024) and Uncapher & Wagner (2018) is PNAS, both corrected/verified above. Remaining before publication: (1) a final scientist read-through; (2) optional spot-check that each remaining link resolves at publish time.