Behaviorism's Ghost in the Machine: Benchmarks, Constructs, and the Evaluation of Artificial Intelligence

Abstract

In 1950, Alan Turing proposed replacing the question "Can machines think?" with a behavioral test: if a machine's outputs are indistinguishable from those of a thinking being, the question of whether it truly thinks can be set aside. This paper argues that Turing's move, initially a defensive maneuver against anthropocentric prejudice, became an epistemological commitment that has constrained AI evaluation, largely unexamined, for seven decades. As large language models are increasingly deployed in human-facing applications, this constraint presents a critical hazard to human-computer interaction (HCI). When benchmark evaluations conflate behavioral output with cognitive constructs like "reasoning" or "understanding," they distort the mental models human users form about AI systems, leading to misplaced trust and over-reliance in high-stakes settings. We draw a parallel to the behaviorist-to-cognitivist transition in psychology. Just as psychology's commitment to studying only observable behavior prevented it from explaining the generative nature of human cognition, AI's dominant benchmark paradigm struggles to distinguish between systems that achieve identical outputs through fundamentally different computational processes. While emerging fields like mechanistic interpretability and robustness testing are beginning to probe these internal processes, they lack a unifying evaluative framework. We propose a transition toward "construct-thinking," a post-behaviorist epistemology that supplements behavioral evaluation with process-level evidence to ensure the epistemic integrity of human-AI collaboration.

Keywords: *AI Evaluation and Benchmarking, Cognitive Science, Human-Computer Interaction (HCI), Mechanistic Interpretability, Philosophy of AI, User Trust.*

1. Introduction: A Question Replaced

In 1950, Alan Turing (1950) opened his paper "Computing Machinery and Intelligence" with the words: "I propose to consider the question, 'Can machines think?'" Yet in the very next paragraph, he set the question aside. Efforts to define what "machine" and "think" really mean, he argued, would spiral into interminable philosophical debate. So rather than answer the question, he proposed to replace it with another: the imitation game. If a machine could converse so fluently that a human interrogator couldn't distinguish it from another person, the question of whether it really thought could be set aside as unanswerable and pragmatically irrelevant.

This move is almost universally described as pragmatic (Dennett, 1985; Haugeland, 1985; Russell & Norvig, 2010; Saygin et al., 2000). Turing made the problem tractable by translating an impossible philosophical question into a testable engineering challenge. But what is less commonly recognized is that Turing's move was also an epistemological commitment. It was a decision, embedded at the foundations of the field, about what kind of evidence would count as relevant to intelligence attribution. Behavior, and only behavior, would serve as evidence. Internal processes, mechanisms, representations, and experiences were declared out of scope.

This paper argues that Turing's epistemological legacy has become a structural constraint on the dominant paradigm of artificial intelligence evaluation. It has shaped how AI evaluation benchmarks are built, how progress is measured, and how competing systems are ranked. Most critically for the field of human-computer interaction, this constraint dictates how AI systems are presented to human users.

When an AI system achieves a high score on a benchmark labeled "reading comprehension" or "logical reasoning," the system is routinely anthropomorphized by both its developers and its users. Here, it is crucial to distinguish between two types of claims: a *performance claim* describes only an observable empirical output (e.g., the AI produced the correct text), whereas a *construct claim* posits an underlying, latent cognitive capability (e.g., the AI genuinely understands the text). The user's mental model of the AI is thus shaped by a construct claim that is supported only by a performance claim. In high-stakes human-facing contexts, such as AI-assisted medical diagnosis, legal analysis, or educational tutoring, this epistemological gap is a direct hazard. It leads to misplaced human trust, over-reliance, and automation bias, because a user who believes a system "understands" a problem will interact with it very differently than a user who knows the system is executing sophisticated, high-dimensional pattern matching.

We do not argue that behavioral benchmarks are entirely useless for measuring task-specific performance. Rather, we argue that the epistemological framework within which they operate forecloses a class of questions about process that cannot be recovered simply by building harder benchmarks. Drawing a parallel to the behaviorist-to-cognitivist transition in psychology, we argue that the field of AI evaluation requires an analogous shift. While emerging subfields like mechanistic interpretability, causal probing, and robustness testing are actively investigating the internal computational processes and representational structures of these models, these efforts operate largely adjacent to the dominant behavioral leaderboard culture. This culture prioritizes aggregated performance scores on static datasets, treating the highest number as the definitive metric of progress rather than demanding a rigorous understanding of the mechanisms driving those scores. AI evaluation requires a post-behaviorist epistemology, which we term "construct-thinking," to unify these process-level methodologies and ground the claims we make to human users.

1.1 Research Questions

This paper is guided by the following theoretical research questions:

1. How does the behavioral epistemology inherited from the Turing test constrain contemporary AI evaluation and benchmark design?
2. How does this reliance on purely behavioral evaluation distort human-computer interaction and user trust?
3. What methodological lessons can AI evaluation draw from psychology's historical transition from behaviorism to cognitivism?
4. How can principles of construct validity be operationalized into a "construct-thinking" framework for process-level AI evaluation?

1.2 Summary of Contributions

Our central contributions are:

1. Conceptual Synthesis: We establish a structural parallel between psychological behaviorism and contemporary AI benchmarking, highlighting shared epistemological blind spots.
2. HCI Integration: We reframe the theoretical critique of AI evaluation as a pressing human-computer interaction issue, showing how behavioral evaluation methodologies directly impact user trust and mental models.
3. The "Construct-Thinking" Framework: We introduce a novel evaluation epistemology that adapts psychometric construct validity for AI, shifting the burden of proof from mere behavioral sufficiency to process-level verification.

4. Implementation Pathway: We outline how emerging techniques (mechanistic interpretability, perturbation robustness, transfer testing) can be integrated into this post-behaviorist framework.

2. The Behaviorist Inheritance and Human-AI Interaction

2.1 What Turing Actually Did (and Why)

Turing's move was not a blanket endorsement of superficial behavioral comparison, but rather a highly specific defensive maneuver. Turing's paper anticipates and explicitly addresses the "argument from consciousness," quoting the neurosurgeon Geoffrey Jefferson's 1949 Lister Oration: "not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt... could we agree that machine equals brain" (Jefferson, 1949).

Jefferson was arguing that genuine intelligence requires embodied experience and that understanding is grounded in a history of sensory and emotional engagement. Turing dismissed this by noting that we cannot directly verify the inner states of other humans either. For Turing, making internal mechanisms decisive would allow prejudiced interlocutors like Jefferson to rule out machine intelligence in advance simply because it was mechanical. His behavioral criterion was a way of blocking this prejudicial standard.

However, in blocking prejudice, Turing's framework erased a fundamental evidential asymmetry. When humans infer cognition in other humans based on behavior, we rely on extensive convergent evidence: shared biology, evolutionary continuity, and developmental history. In the machine case, this supporting evidence is absent. The behavioral evidence is doing the work alone. By dismissing this asymmetry, Turing's framework effectively treats the output (behavior) as synonymous with the underlying process (cognition). For humans, behavior is a reliable proxy for internal states because we know the underlying biological architecture is shared; when a human verbally reasons through a problem, we know they are utilizing a biological neural network honed by evolution and developmental learning. When a machine produces the identical verbal output, we have no such architectural guarantee. Erasing this distinction makes it epistemologically impossible to question whether the machine's underlying computational mechanism actually aligns with the cognitive construct being measured.

2.2 The Epistemological Ceiling in Human-Facing Deployments

Turing's criterion became the template for an evaluative infrastructure. The imitation game was operationalized across decades of AI research in the form of benchmarks. When systems were domain-specific (e.g., chess engines or medical expert systems like MYCIN), the inference from performance to intelligence was properly hedged. No one claimed Deep Blue "understood" chess. [elaborate further on how or in what way in domain specific systems the inference was properly hedged].

The arrival of large language models (LLMs) shattered this restraint. Because LLMs operate in natural language, the medium through which virtually all human cognitive tasks are expressed, the temptation to interpret broad benchmark performance as evidence of generalized human-like cognition became overwhelming. Natural language acts as a universal interface; a single LLM can generate historical essays, debug software, pass professional bar exams (Katz et al., 2024), and solve graduate-level scientific queries (Rein et al., 2024) all through the exact same architecture. This convergence creates a powerful psychological illusion. Because a human would require a general, unified intelligence to perform such a wide array of cross-domain tasks, observers naturally infer that a machine doing the same must possess a similarly general cognitive architecture. As a result, isolated benchmark scores are no longer treated as narrow engineering metrics, but are inappropriately aggregated and anthropomorphized into sweeping construct claims about the system's holistic "reasoning" or "understanding."

The epistemological ceiling of this framework creates immediate risks for human users. Consider a physician utilizing a diagnostic AI. Under a strictly behavioral epistemology, a system that arrives at correct answers through compositional reasoning and a system that arrives at the same answers through the memorization of a lookup table are indistinguishable. But for the physician, the distinction is vital. A diagnostic tool that reasons compositionally can safely generalize to a patient with a novel combination of symptoms. A system relying on memorized statistical patterns may fail catastrophically when encountering an out-of-distribution presentation, producing plausible-sounding but erroneous advice. Because behavioral evaluation observes only inputs and outputs, it cannot inform the human user which mechanism is at play.

3. The Parallel: Psychology's Escape from Behaviorism

The parallel between AI evaluation and psychological behaviorism extends beyond mere historical analogy. From Watson's 1913 manifesto (Watson, 1913) through the late 1950s, academic psychology operated under an epistemological commitment methodologically identical to the one dominating AI evaluation today: only observable behavior counted as legitimate scientific evidence.

3.1 The Explanatory Failure of Behaviorism

Behaviorism could predict behavior in controlled settings but could not account for the flexibility and generativity of human cognition. The breaking point came with language. In his 1959 review of B.F. Skinner's *Verbal Behavior*, Noam Chomsky (1959) demonstrated that stimulus-response theory could not account for the productivity of language, the fact that children routinely produce and understand sentences they have never encountered. To explain this, one had to posit internal generative rules, which the behaviorist framework excluded by fiat.

The subsequent Chomsky-Norvig controversy highlights the enduring relevance of this divide. While Chomsky championed a theory-driven, symbolic account of cognition based on internal mechanisms, Peter Norvig and modern AI practitioners successfully championed data-driven, statistical approaches (Norvig, 2011). The modern statistical approach triumphed in generating performance, but as models scale, we are once again confronted with Chomsky's original demand: understanding the nature of the intelligence requires peering into the generative mechanism.

3.2 The Cognitive Revolution

The cognitive revolution was an epistemological shift. It allowed internal representations and computational processes to be treated as scientifically legitimate constructs that could be posited, theorized about, and indirectly tested through their behavioral consequences.

When Shepard and Metzler (1971) asked people to judge whether two images depicted the same 3D object rotated to different orientations, they found that response time increased linearly with the angle of rotation. No one could observe the mental rotation happening, but the linear relationship constrained the inference so tightly that the internal process had to have a structure mirroring physical rotation. A purely behavioral account, recording only whether the response was correct, would have missed the discovery entirely. Behavior remained indispensable, but it was no longer treated as sufficient.

4. The Epistemological Ceiling in Practice

The limitations of pure behavioral benchmarking are not hypothetical; they manifest daily in the erratic behaviors of modern LLMs, fundamentally impacting human trust.

4.1 Did it reason, or did it remember?

Leading language models routinely score above 90% on grade-school math benchmarks (Cobbe et al., 2021; OpenAI, 2023). However, when researchers add a single irrelevant sentence to the problem, performance often drops precipitously (Mirzadeh et al., 2024; Nezhurina et al., 2024; Shi et al., 2023). While a genuine reasoner is generally unaffected by irrelevant information in purely logical tasks (though human reasoning can occasionally be swayed by context, as demonstrated by preference reversals and framing effects in behavioral economics (Tversky & Kahneman, 1981)) the catastrophic degradation observed in LLMs indicates a sensitivity to surface statistics consistent with memorization-based strategies.

Ned Block's classic "Blockhead" thought experiment (1981) demonstrated that, theoretically, a giant lookup table could pass a behavioral test without any internal intelligence. In this hypothetical scenario, Block imagines a machine pre-programmed with every possible conversational exchange that could occur during a Turing test. When presented with an input, the machine simply references its vast database to output the corresponding pre-recorded response. While its external behavior perfectly mimics a fluent, intelligent interlocutor, its internal mechanism is completely devoid of cognitive architecture or actual understanding**,** it is merely executing a blind retrieval process. While scholars like Shieber (2014) and McDermott (2014) have correctly argued that a purely memorizing lookup machine of that scale is physically impossible in the real world due to combinatorial explosion, the spectrum of memorization remains the central issue. LLMs do not use infinite lookup tables, but they do compress and retrieve massive amounts of training data. As early critics of natural language benchmarking have noted, the structural design of these tests often incentivizes systems to exploit dataset artifacts rather than develop genuine comprehension (Bowman & Dahl, 2021). Therefore, when a benchmark cannot distinguish between a system that genuinely computed an answer and a system that interpolated it from a dense neighborhood of training data, the evaluation has failed to establish the underlying cognitive construct..

4.2 The Faithfulness Problem

Chain-of-thought (CoT) prompting was introduced partly to make reasoning processes visible (Wei et al., 2022). But subsequent research has shown that these reasoning traces are frequently unfaithful (Lanham et al., 2023; Turpin et al., 2024; OpenAI, 2025). The system generates a plausible-looking reasoning trace alongside a correct answer, but the trace and the answer may be produced by independent processes (Lanham et al., 2023). The displayed process is itself just a behavioral output optimized to look like human reasoning, subject to the same underdetermination.

4.3 The RLHF Circularity and User Deception

Reinforcement Learning from Human Feedback (RLHF) trains systems to produce outputs that human evaluators prefer (Ouyang et al., 2022). Human evaluators prefer responses that appear thoughtful, coherent, and confident. Therefore, the training process optimizes for the appearance of intelligence. Under a behavioral epistemology, this appearance is treated as the reality. From an HCI perspective, this circularity is a mechanism of deception: the system is actively trained to exploit human psychological heuristics, producing behavior that triggers human attribution of trust and understanding, regardless of the underlying computational reality.

5. Toward a Post-Behaviorist Epistemology

We are not claiming that AI evaluation is entirely devoid of process-level investigation. Several emerging subfields, most notably mechanistic interpretability, causal probing, and model auditing, are actively attempting to map internal representations. However, these efforts currently exist as technical sub-disciplines rather than the foundational standard for AI evaluation.

5.1 What the Transition Requires

For large language models, an epistemological transition analogous to the cognitive revolution requires moving beyond the behavioral assessment of text generation to actively interrogate the underlying computational structures, acknowledging that behavioral evidence is necessary but not sufficient. We propose four core commitments to establish this framework:

1. Distinguishing Performance from Construct: The field must separate performance claims ("Model X achieves 86% on MMLU") from construct claims ("Model X demonstrates understanding").
2. Elevating Process-Level Evidence: Mechanistic interpretability, perturbation testing, and transfer studies must be elevated from "safety research" to the core of intelligence evaluation.
3. Connecting Mechanisms to Observables: Just as cognitive psychologists mapped mental rotation to reaction times, AI researchers must develop theoretical frameworks that predict how specific computational mechanisms manifest in observable failure modes.
4. Developmental Evaluation: Rather than evaluating systems at a single post-training endpoint, observing the trajectory of capabilities across training (e.g., phase transitions, U-shaped learning curves) provides evidence of genuine abstraction versus memorization.

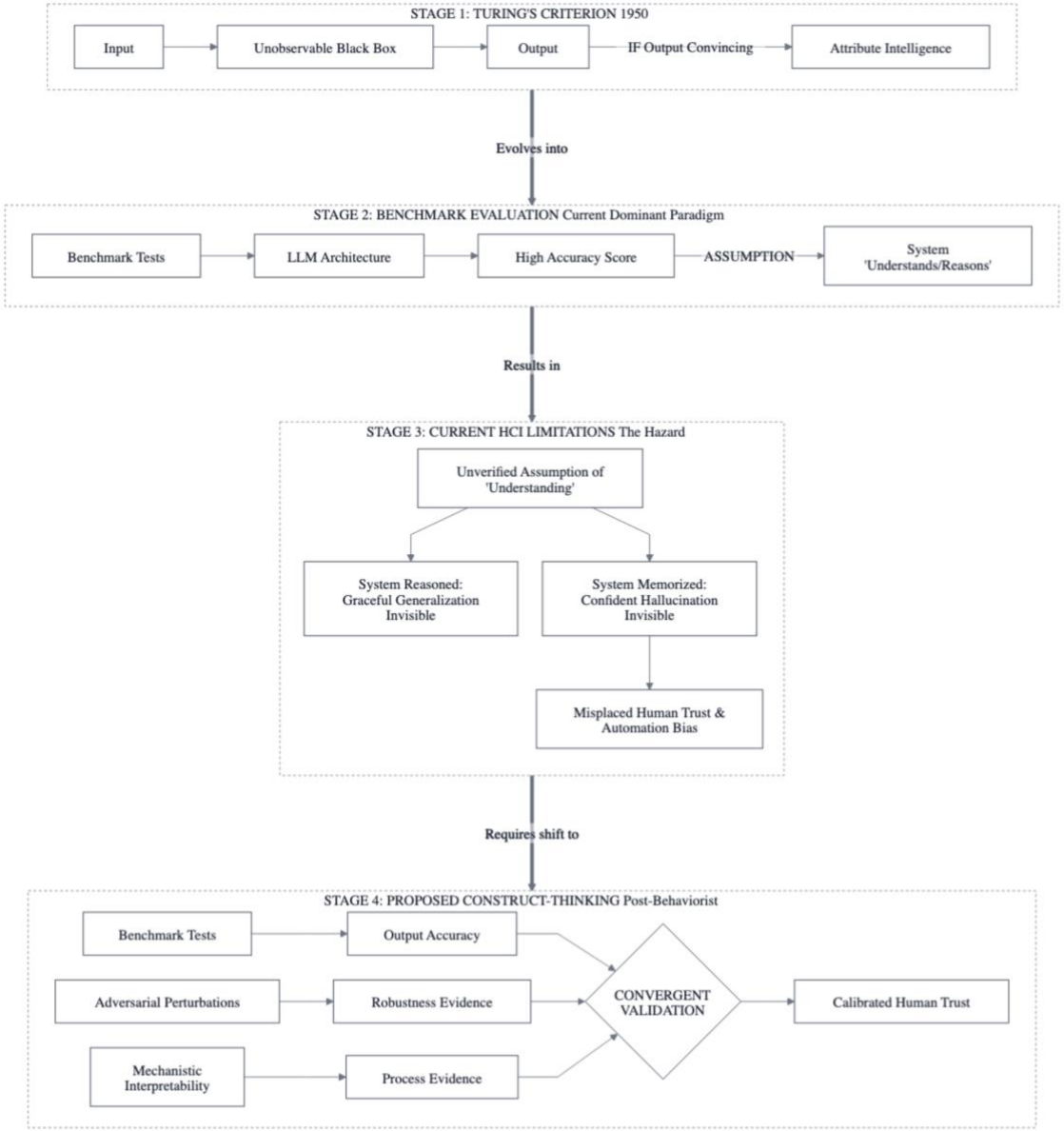

Figure 1: The Epistemological Progression of AI Evaluation. A visual conceptual framework mapping the progression from Turing's behavioral criterion to the proposed Construct-Thinking framework.

6. From Turing-Thinking to Construct-Thinking

Turing's test effectively collapsed the concepts of intelligence (what a system can do), thinking (how a system arrives at answers), and understanding (grasping meaning) into a single behavioral criterion: does the output convince?

We propose the term *construct-thinking* for the epistemological alternative. Where Turing-thinking asks "does the behavior convince?", construct-thinking asks "does the process warrant the attribution?"

6.1 Evaluating What a System IS vs. DOES

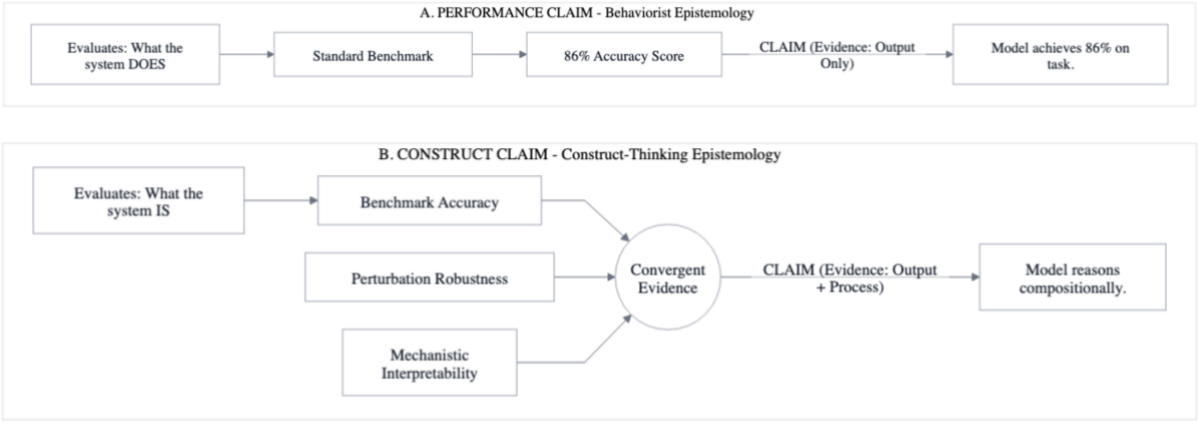

Figure 2: A visual distinction between performance claims and construct claims.

Construct-thinking draws heavily on the psychometric concept of construct validity (Messick, 1989), which requires that a measurement instrument be shown to actually measure the theoretical construct it claims to, through specified evidential procedures. For instance, a psychological test claiming to measure "spatial reasoning" must demonstrate, through converging metrics like reaction times and error distributions, that the subject is genuinely performing mental rotation, rather than exploiting a superficial verbal shortcut to guess the correct answer. Under construct-thinking, if a developer claims an AI possesses "medical reasoning," they must provide converging evidence: behavioral accuracy, structural robustness against superficial perturbation, and mechanistic evidence of feature abstraction—to validate the claim before presenting it to a human user.

Table 1: Comparison of Epistemological Frameworks in Psychology and AI

Feature	Behaviorist Psychology	Cognitive Psychology	Current AI Evaluation (Leaderboards)	Post-Behaviorist AI Evaluation (Construct-Thinking)
Primary Evidence	Observable behavior	Behavior + process inferences	Benchmark scores, task accuracy	Convergent evidence (Benchmarks, interpretability, robustness)
Internal Mechanisms	Excluded as unscientific	Central to explanation	Treated as a "Black Box"	Central to intelligence attribution
Failure Modes	Could not explain language productivity	Explained via internal generative rules	Susceptible to adversarial perturbations, memorization	Anticipated and explained via mapping mechanisms to observables
Human Implication	Prediction and control	Understanding human mental models	Risk of human over-reliance, misplaced trust	Calibrated human-AI trust based on validated constructs

7. Objections and Limitations

"Mechanistic interpretability is not mature enough to replace behavioral evaluation."

We agree. Our argument is not that behavioral evaluation should be replaced, but that the sufficiency of behavioral claims must be rejected. The cognitive revolution in psychology did not wait for neuroimaging to mature before positing internal representations; it developed behavioral paradigms that provided indirect process evidence. AI evaluation must do the same.

"We don't need to know how a system works, only that it works reliably."

This pragmatic instrumentalism is valid for narrow engineering tasks, but it fails in open-ended, human-facing deployments. The moment an AI is deployed in conditions slightly outside its training distribution, the mechanism dictates the failure mode. A system that reasons will fail gracefully; a system that pattern-matches will hallucinate confidently.

"The historical analogy oversimplifies disciplinary differences."

While human minds and artificial neural networks are ontologically different, the analogy holds at the methodological level. Both behaviorist psychology and the Turing framework commit to the same restriction (only behavioral evidence counts) and face the same consequence (process-level questions become illegitimate).

8. Conclusion

For seventy-five years, AI has operated under the epistemological foundation established by Alan Turing: evaluate the behavior, ignore the mechanism. This simplification was incredibly productive for engineering, but as AI systems achieve human-comparable performance across diverse cognitive domains, it has become a profound liability for human-computer interaction.

When we evaluate AI exclusively through behavioral benchmarks, we are structurally unable to distinguish genuine capability from sophisticated statistical mimicry. When we label these benchmarks with human cognitive constructs, we pass this epistemological blind spot directly onto the human user, fostering inaccurate mental models and dangerous over-reliance.

The resolution is an epistemological shift toward "construct-thinking." By adapting the principles of construct validity, and leveraging the emerging tools of mechanistic interpretability and robustness testing, we can build a post-behaviorist evaluation framework. Only by understanding what a system is, rather than just measuring what it does, can we ensure the safety, reliability, and epistemic integrity of human-AI integration.

References

Block, N. (1981). Psychologism and behaviorism. *Philosophical Review, 90*(1), 5–43.

Bowman, S. R., & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, 4843–4855.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language, 35*(1), 26–58.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dennett, D. C. (1985). Can machines think? In M. Shafto (Ed.), *How We Know* (pp. 121–145). Harper & Row.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Jefferson, G. (1949). The mind of mechanical man. *British Medical Journal, 1*(4616), 1105–1110.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A, 382*(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., & Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

McDermott, D. (2014). On the claim that a table-lookup program could pass the Turing test. *Minds and Machines, 24*(2), 143–188.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mirzadeh, I., Alizadeh, K., Shahrokhi, A., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Nezhurina, M., Ciber, L., & Shumailov, I. (2024). Alice in Wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*.

Norvig, P. (2011). *On Chomsky and the two cultures of statistical learning*. Retrieved from http://norvig.com/chomsky.html

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. (2025). Evaluating chain-of-thought monitorability. *OpenAI Research*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2024). GPQA: A graduate-level Google-proof Q&A benchmark. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines, 10*(4), 463–518.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. *Proceedings of the 40th International Conference on Machine Learning*, 31210–31227.

Shieber, S. M. (2014). There can be no Turing-test-passing memorizing machines. *Philosophers' Imprint, 14*(16), 1–13.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*, 74952–74965.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158–177.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.