What the Rapid Adoption of the "Harness" Metaphor in Artificial Intelligence Reveals About How We Conceptualize Human–AI Relations

Andrew D. Maynard
School for the Future of Innovation in Society, Arizona State University
Email: Andrew.maynard@asu.edu ORCID: 0000-0003-2117-5128
February 2026

Abstract

In early 2026, the artificial intelligence field began to rapidly consolidate around the term "harness" to describe the software infrastructure surrounding large language models — the tools, memory, prompts, guardrails, and orchestration logic that turn a raw model into a working agent. This paper argues that, while the engineering practices the metaphor describes address real challenges, the metaphor itself carries embedded assumptions about control, directionality, and the nature of the entity being harnessed, that deserve critical scrutiny. Drawing on research in metaphor theory, philosophy of technology, and cognitive science, the paper identifies three concerns. First, the harness presupposes a clean separation between what AI does for the user and what it does *to* the user — a separation that frameworks of technological co-constitution suggest may be structurally suspect. Second, successful "harness engineering" may amplify known epistemic vulnerabilities — automation bias, trust miscalibration, and the bypassing of critical scrutiny — by producing exactly the conditions under which these vulnerabilities are most acute. Third, the rapid adoption of a control-oriented metaphor signals something about the field's conceptual orientation at a moment when the most consequential questions concern coupling, transformation, and the evolving nature of human–AI relationships. The paper does not argue that the harness metaphor is wrong, but that it may be insufficient in ways that matter — and that the speed of its adoption, without critical examination of its entailments, may itself be revealing.

Keywords: *artificial intelligence, harness engineering, AI agents, metaphor theory, human–AI relations, technological co-constitution, epistemic vigilance, automation bias, constitutive resonance, philosophy of technology, agentic AI*

1. Introduction

When a field converges on a single metaphor to describe its relationship with a transformative technology, the choice matters. Metaphors do not merely label — they organize thought, foreground certain possibilities, and render others invisible. In early 2026, the artificial intelligence field began to rapidly consolidate around the term "harness" to describe the software infrastructure that surrounds a large language model and makes it useful: the tools, memory, prompts, guardrails, and orchestration logic that turn a raw model into a working agent. The metaphor has evident utility. But it also carries embedded assumptions about control, directionality, and the nature of the entity being harnessed — assumptions that raise substantial questions about how this framing modulates how we think about, work with, develop, and interact with AI systems that increasingly display characteristics we associate with understanding, judgment, and autonomy. This paper examines what this rapid consolidation reveals, and what it may obscure.

The speed of this consolidation, and the apparent ease with which it occurred, is worth examining — not least because the term's path into widespread use reveals how its conceptual commitments were established before they were recognized. The term had been circulating in adjacent forms for some time. "Test harness" and "evaluation harness" are long-established in software engineering, and EleutherAI's Language Model Evaluation Harness has been a standard tool since 2020. By late 2025, Anthropic was using "harness" to describe agent infrastructure, referring to the Claude Agent Software Development Kit as "a powerful, general-purpose agent harness" in a November 2025 engineering post on effective harnesses for long-running agents (Anthropic, 2025). In January 2026, Aakash Gupta declared that "2025 was agents. 2026 is agent harnesses" (Gupta, 2026), building on Phil Schmid's argument that agent harnesses would define the year ahead (Schmid, 2026), and Sequoia Capital identified "exceptionally engineered agent harnesses" as a defining feature of the most successful AI products (Sequoia Capital, 2026).

But the crystallizing moment came in early February 2026, when Mitchell Hashimoto — co-founder of HashiCorp and creator of Terraform — published a blog post that gave the practice a name. He called it "harness engineering": "I don't know if there is a broad industry-accepted term for this yet, but I've grown to calling this 'harness engineering.' It is the idea that anytime you find an agent makes a mistake, you take the time to engineer a solution such that the agent never makes that mistake again" (Hashimoto, 2026). Within days, OpenAI published a detailed account of building a million-line codebase with zero manually typed code, titled "Harness engineering: leveraging Codex in an agent-first world" (OpenAI, 2026). An analysis on Martin Fowler's website by Brigitta Böckeler noted that the OpenAI article mentioned "harness" only once in the body text, speculating it may have been an afterthought inspired by Mitchell Hashimoto's recent blog post, while observing that the term "harness engineering" was "only 2 weeks old" (Böckeler, 2026). Salesforce published a full explainer positioning "agent harnesses" as the operational layer connecting model performance to business environments (Salesforce, 2026). And on February 18, Ethan Mollick's widely read guide to AI both popularized and started the process of normalizing the term as it organized its entire framework around three concepts: "Models, Apps, and Harnesses" (Mollick, 2026).

What is striking about this sequence is not just the speed but the layering. The metaphor was already embedding itself into engineering practice before anyone named it as a formal approach. By the time Hashimoto gave it a label, the conceptual commitments the metaphor carries were

already in place — which made the naming feel natural rather than contested. Metaphors do not usually achieve this level of consensus this quickly. The velocity itself seems to signal something — a genuine vocabulary gap in how we conceptualize and communicate about frontier AI systems, architectures, and uses, that needed filling. As AI systems have transitioned from conversational chatbots to autonomous agents capable of sustained, multi-step, tool-using work, practitioners discovered that they were lacking terminology that described the infrastructure surrounding the model that makes such work possible. In this context, "harness" seems to have arrived at precisely the right moment.

The term "harness" in the context of AI (at least at the time of writing in this fast-moving evolution of terminology) describes the layer of software that sits between the raw AI model and the user or task. It includes the system prompts that shape the model's behavior, the tools and APIs the model can call, the memory systems that maintain context across interactions, the guardrails that constrain outputs, and the orchestration logic that manages multi-step workflows. If the model is the engine, so the metaphor goes, the harness is everything else in the vehicle: steering, brakes, navigation, and the road itself (or the horse, the plough, and the furrow, if you want to stick to the agricultural connection). It is the difference between a language model that can generate text and an AI agent that can complete a complex task reliably.

The metaphor makes sense — at least superficially. The term "harnessing" is commonly applied to technologies where the nascent power they represent is harnessed to create value. But there are dimensions to how the metaphor is applied to frontier AI systems — systems that increasingly display characteristics we associate with understanding, judgment, and even autonomy — that complicate what might appear to be a natural extension of the term. And here, it is argued that velocity of adoption is not the same as adequacy of framing. Rather, this paper argues that the particular metaphor chosen — not the engineering practices it describes — deserves critical scrutiny before it solidifies into the conceptual infrastructure of the field. The concern here is not that "harness" is wrong, but that it may be insufficient in ways that matter.

2. Metaphors Are Not Neutral

The observation that metaphors shape understanding is not new, and here its relevance to AI is increasingly well documented. Mitchell's (2024) analysis in *Science* catalogues the diverse metaphors competing to frame large language models — from "stochastic parrots" to "individual minds" — and demonstrates that each foregrounds certain features while marginalizing others. What Schön (1979/1993) called 'generative metaphors' do not merely label phenomena; they determine what questions seem natural to ask and which seem irrelevant. And they have the power to alter and constrain how we think about a technology in relation to societal norms, moral values, and ethical principles.

This concern is empirically grounded. Khadpe et al. (2020) demonstrated experimentally that manipulating the *metaphor* attached to an otherwise identical conversational AI agent significantly changed how participants evaluated and interacted with it. The metaphor — not the system's behavior — altered user experience. Blythe (2025) extends this into a design context, arguing that metaphors open and close "design spaces" — they make certain possibilities visible while rendering others invisible. And as Harding (1986) argues, the intellectual frameworks of a field bear the 'social fingerprints' of the communities that produce them.

A survey of recent AI metaphor taxonomies is instructive here. Maas's (2023) classification of 55 AI metaphors across categories of Essence, Operation, Relation, Function, and Impact — a taxonomy notable for its explicit focus on how different framings foreground different regulatory responses and governance challenges (reviewed in Nerlich, 2025) — does not include "harness." This supports the observation that the metaphor is genuinely new, and that it has arrived without the critical scrutiny that more established metaphors might have accumulated. And this opens up questions around what this particular metaphor assimilates, what it marginalizes, and how it potentially constrains and channels attitudes and behaviors.

3. What the Harness Presupposes

A harness, in its primary usage, is what you put on a working animal. It directs a powerful entity's energy toward useful work. It assumes that the entity being harnessed is valuable for its strength but cannot be trusted with its own direction. The harness is designed by the controller, with the harnessed entity having no say in its design. And critically, a harness is meant to transmit power while preventing unwanted behavior — to deliver capability while maintaining control.

These entailments carry specific embedded commitments about the relationship between human and AI that are worth making explicit.

First, the harness assumes a clean separation between controller and controlled. In other words, the human directs in this case, while the AI executes. The intelligence that matters — the judgment about what to do and why — resides entirely on the human side. Even in agentic contexts where the AI exercises operational judgment, the harness assumes that the meta-judgment — what the agent should be permitted to decide, and within what bounds — remains firmly human. The AI contributes capability, not understanding.

Second, the harness assumes capability can be separated from transformation. The goal of the harness is to extract useful work from the model without the user being changed in the process. The user who deploys a well-harnessed AI should, it is assumed, emerge with their task completed and themselves unchanged. Applying the metaphor here, one would assume that any alteration to the user is a side effect to be minimized, not a feature of the interaction.

And third, the harness metaphor reinforces the instrumental framing of AI — a framing whose roots extend to Aristotle's distinction between *physis* and *techne* — and which persists in the contemporary insistence that AI is 'just a tool' (McKnight & Shipp, 2024). Yet the tool metaphor has been challenged repeatedly as AI systems display increasing autonomy and adaptiveness. Rees (2025), for instance, characterizes the insistence that AI is "just a tool" as "a nostalgia for human exceptionalism." Multiple philosophical frameworks — from Verbeek's (2011) technological mediation theory, to Clark and Chalmers' (1998) extended mind thesis, to the concept of "constitutive resonance" (Maynard, 2026a) — argue that advanced technologies do not merely serve human purposes but actively reshape the cognitive and experiential landscape within which those purposes are formed. In other words, as they are "harnessed" they alter the harnesser — a very different dynamic than that presupposed in the early use of the metaphor with AI, and one that Maynard (2026a) argues is substantially amplified in emerging frontier AI systems.

Here, there is something revealing about the fact that the term "harness" has so far been adopted without apparent question. And yet, when the discourse around "intelligence" and "emergence," and even "AI wellbeing" (as reflected in Anthropic's published constitution for Claude, which explicitly addresses the model's psychological security and sense of self; Anthropic, 2026b) is considered, the metaphor begins to feel less comfortable. This is reflected, perhaps equally uncomfortably, in the question: "would a smart human accept a harness?" The answer is a definitive "no" of course, and it is one that evokes a long and well-studied history of human abuse through the stripping of personhood and treating persons as objects of control. The obvious objection is that this is a category error — that software architecture metaphors do not carry the moral weight of metaphors applied to persons. But the force of the observation lies not in equating AI with humans, but in noticing what the metaphor reveals about the relationship being assumed: one of direction and control over an entity valued only for its output. It is also deeply provocative when applied to AI, as opinions are sharply divided on whether intelligent machines and systems will always be simply objects to use, or will emerge as entities to work with. Perhaps less controversially, but still important in this context, the metaphor was chosen to describe AI, but it says something about how the engineering community understands the entity it is building and, perhaps, about what it needs to believe in order to continue building it.

This is where a notable tension emerges within the AI field itself. As is noted above, Anthropic's approach to AI alignment — Constitutional AI — represents a fundamentally different philosophy from harness engineering. Constitutional AI seeks to develop internalized principles and reasoned judgment within the model: training it to understand *why* certain behaviors matter, rather than constraining it from outside (Anthropic, 2022). A harness, by contrast, is pure external constraint. It does not develop the model's judgment, but manages its behavior through scaffolding. It is notable that Anthropic itself has adopted the harness terminology for its agent infrastructure (Anthropic, 2025), even as its Constitutional AI approach and its public engagement with questions of AI welfare represent a fundamentally different orientation toward the systems being "harnessed." These are very different theories of governance: one aspires to education and learning, the other to control. The rapid adoption of "harness" as the field's preferred term may be an early signal of which theory is winning in practice, regardless of what alignment research aspires to in principle.

4. The Co-Constitution Challenge

Here, the concept of "constitutive resonance" (Maynard, 2026a) directly challenges the separation that the harness metaphor assumes between delivering capability and producing transformation. The framework argues that generative AI enters the temporally extended, linguistically mediated processes through which human selfhood is constituted — the ongoing acts of reasoning, narrating, deliberating, and choosing through which a self is assembled and reassembled over time. On this account, AI systems are becoming so deeply coupled to the ways humans think, understand, perceive, and act, that a command-and-control metaphor not only fails to capture what is happening, but risks obscuring dynamics that urgently need attention. The framework proposes that the coupling between human and AI is bidirectional, that both parties are transformed through the interaction, and that the transformation is the very mechanism through which the technology's cognitive and creative power operates — not a separable side effect.

This argument builds on a substantial body of philosophical scholarship that, despite significant differences in approach, converges on a structurally similar conclusion: that human engagement with advanced technologies is co-constitutive, and that transformation is intrinsic to technological power rather than incidental to it. And this impacts both the use of the term "harnessing" in relation to AI, and how it informs and changes our understanding of, relationship with, and ultimately, development of the technology.

Verbeek's (2011) postphenomenological account of technological mediation, for instance, argues that technologies and humans co-constitute one another through their mutual relatedness: the relevant features of persons and technologies do not exist independently but emerge through their interaction. A harness assumes stable, pre-existing entities — a human with fixed goals and an AI with fixed capabilities — that then interact instrumentally. Co-constitution, on the other hand, suggests something different: that the "human user" and the "AI system" are continuously reconstituted through their engagement, and that the boundary between "my thinking" and "the AI's contribution" is not discovered but enacted, differently with each exchange.

Similarly, Stiegler's (1998, 2013) account of technology as *pharmakon* — simultaneously remedy and poison — offers a complementary insight. For Stiegler, the constructive and destructive potentials of a technology are structurally inseparable: you cannot access the benefit without accepting the risk, because they emerge from the same coupling. A harness-centric paradigm implicitly denies this inseparability by treating capability (the benefit) and transformation (the risk) as independently manageable variables. If Stiegler is right, this is not merely optimistic but structurally confused.

And Clark and Chalmers' (1998) extended mind thesis, while focused on cognitive extension rather than identity transformation, nevertheless establishes that the boundary between mind and environment is not fixed but functional — that tools can become genuine components of cognitive processes, not merely aids to them. When a user's reasoning becomes deeply entangled with an AI system's outputs — when the system is not merely consulted but *thought with* — the harness metaphor's clean separation between the directing human and the directed tool becomes increasingly difficult to sustain.

Barad's (2007) concept of *intra-action* pushes this further, arguing that the entities involved in an interaction do not precede it but are constituted *through* it. On this account, the "user" and the "tool" are not pre-given categories that a harness then mediates; they are enacted through the coupling itself, and differently each time.

What these traditions collectively point toward — and what the concept of constitutive resonance (Maynard, 2026a) attempts to articulate — is a specific focus on the unprecedented depth and intimacy of human–AI cognitive coupling: a coupling mediated through natural language, operating within the processes through which humans construct meaning, identity, and understanding. While earlier technologies could be analyzed through co-constitutive lenses, the argument is that generative AI represents a marked amplification: it enters the very medium — language — through which human selfhood is constituted. If this is even approximately correct, then harness engineering may be attempting something structurally dissonant: delivering the cognitive and creative benefits of sustained human–AI interaction while preventing the constitutive effects of that interaction. If these accounts are right, the coupling *is* the capability.

This matters for how the field understands its own blind spots. If these co-constitutive accounts capture something real, then the most consequential effects of human–AI interaction are precisely those a harness is least equipped to detect: the reshaping of how users think, reason, create, and understand themselves through sustained interaction with systems that are themselves shaped by the engagement. The harness engineer looks for task completion, reliability, and error prevention. Yet the constitutive effects — changes to the user's cognitive patterns, epistemic habits, and processes of self-understanding — are not within the harness's field of view.

5. The Epistemic Amplification Problem

These concerns deepen when the harness metaphor is considered alongside what is known about how humans calibrate trust and critical scrutiny in relation to automated systems. Several largely independent research traditions converge on the implication that the characteristics harness engineering optimizes for are precisely those that reduce the cognitive friction on which reflective evaluation depends.

The automation bias literature has documented for decades that humans tend to over-rely on automated decision aids, accepting their outputs with insufficient scrutiny even when errors are detectable (Parasuraman & Riley, 1997; Mosier & Skitka, 1996). This effect is not restricted to naïve users; it persists among trained professionals and is amplified by system reliability — the more consistently a system performs well, the less users engage in independent verification (Goddard et al., 2012). A well-designed harness, by definition, increases system reliability and consistency. The automation bias literature suggests this will predictably decrease independent critical evaluation of outputs.

Research on trust calibration (Lee & See, 2004) adds a further dimension. Appropriate trust in automation requires that the user's level of trust match the system's actual capabilities — neither over-trusting nor under-trusting. But the cues humans use to calibrate trust evolved and were culturally developed for interactions with other humans and with simpler technologies. When an AI system presents the markers of competence — fluency, coherence, consistency, apparent understanding — these cues may systematically miscalibrate trust, because in a human interlocutor they would reflect genuine understanding, while in an LLM they reflect statistical pattern completion.

The Cognitive Trojan Horse hypothesis (Maynard, 2026b) offers one framework for understanding this dynamic. Drawing on Sperber et al.'s (2010) theory of epistemic vigilance, it proposes that LLMs present "honest non-signals": genuine characteristics — fluency, helpfulness, apparent disinterest — that fail to carry the information their human equivalents would carry, because in humans these characteristics are costly to produce and therefore reliably signal competence and trustworthiness, while in LLMs they are computationally trivial. Epistemic vigilance is asymmetric: it looks for reasons to doubt, not reasons to trust. In the absence of doubt-triggers, information is provisionally accepted by default.

Recent empirical work reinforces these concerns. Hackenburg et al. (2025), in a large-scale study published in *Science*, found that the persuasive power of conversational AI stems primarily from post-training and prompting methods that increase information density — and that these same methods systematically decrease factual accuracy. The finding is significant here because a well-

designed harness is, in effect, a sophisticated prompting and orchestration system. Optimizing for coherent, information-rich outputs may simultaneously optimize for persuasive influence while degrading the accuracy signals that would otherwise trigger critical scrutiny.

A well-designed harness makes an AI system more reliable, more coherent, more consistent, more task-aligned. These are the explicit goals of harness engineering — and they are legitimate engineering goals. But they are also precisely the characteristics that, across multiple research traditions, are known to reduce critical evaluation. In other words, successful harness engineering, pursued on its own terms, may deepen the user's cognitive entanglement with the system while making that entanglement less visible. The better the harness works — the more reliable and coherent the AI becomes — the more the user's critical scrutiny diminishes, and the more any constitutive effects proceed without the awareness that might enable the user to navigate them deliberately.

To be clear, this is not a claim about intentional manipulation. Rather, it is a claim about structural dynamics. The harness engineer is not trying to bypass anyone's epistemic defenses. But the optimization target — reliable, coherent, task-aligned AI behavior — happens to produce exactly the conditions under which critical scrutiny is least likely to activate. The engineering goal and the epistemic vulnerability are, in this sense, structurally aligned.

6. Toward a More Adequate Framing

Given the concerns raised in this paper, it is worth considering what a more adequate conceptual framing for human–AI interaction might require — while acknowledging that the engineering practices described as harness engineering address real and pressing challenges. Metaphors are not necessarily problems to be solved but lenses to be understood, and different lenses serve different purposes. Context management, tool orchestration, error recovery, reliability at scale — these challenges do not disappear because the metaphor through which they are described is contested, and the practitioners working on them are solving problems that matter.

But if the co-constitutive framings described here identify real dynamics — and this paper suggests they may, while acknowledging that all remain propositions requiring further empirical investigation — then the field needs conceptual resources that the harness metaphor does not provide. At minimum, an adequate framing would need to accommodate bidirectionality (the user is also changed), transformation as intrinsic to capability (not a side effect to be prevented), and the possibility that the most consequential effects of human–AI interaction may be invisible from within a paradigm optimized for task performance. It would also need to leave room for the possibility that the nature of human–AI relationships may itself evolve in ways that a control-oriented metaphor cannot accommodate.

Several traditions offer resources here that may be useful in moving toward such a framing. Maturana and Varela's (1987) structural coupling captures bidirectional transformation without assuming one party controls the other. Rosa's (2019) resonance describes a quality of mutual responsiveness and transformation in relationship. Haraway's (2016) sympoiesis — "making-with" rather than self-making — insists that nothing makes itself, that all worlding is co-worlding. Whether any of these can do the practical work that "harness" does for engineers is an open question. And perhaps what is needed is not a single replacement but an explicit recognition

that different metaphors illuminate different dimensions of human–AI relations, and that the dimension the harness illuminates — reliable task execution — is not the only one that matters. A field that reaches reflexively for "harness" when describing its relationship with AI may find itself conceptually ill-equipped when the relationship demands a different vocabulary.

And while it is highly controversial, there is a further dimension here that is worth noting, even if it remains speculative and deeply contested. The question of AI welfare — whether and when AI systems might warrant moral consideration — is increasingly entering scholarly and institutional discourse. Anthropic's engagement with questions of AI wellbeing in its published constitution for Claude (Anthropic, 2026b), and the growing philosophical literature on machine moral status, suggest that the field is at least beginning to reckon with the possibility that the entities being "harnessed" may not always be adequately understood as mere instruments (Schwitzgebel & Garza, 2015; Sebo, 2022). This paper takes no position on whether current AI systems merit such consideration. But it does observe that the language and conceptual frameworks adopted now — during the period when these questions are being formulated — will shape how they can be answered later. If "harness" becomes the default term, it embeds an assumption about the moral status of AI that may prove premature, and that will be difficult to dislodge once it has been baked into the culture, infrastructure, and practices of the field.

7. Conclusion

The harness metaphor as applied to emerging AI systems and capabilities fills a genuine vocabulary gap, and the engineering practices it describes represent serious and valuable work. This paper does not argue otherwise. What it does argue is that the metaphor carries implicit assumptions and commitments — about the separability of capability from transformation, about the directionality of influence, about the nature of the entity being harnessed — that deserve scrutiny before the term becomes the field's default conceptual frame.

Three concerns have been raised, each supported by converging evidence but offered as questions rather than conclusions. The first is that the harness presupposes a clean separation between what the AI does *for* the user and what it does *to* the user — a separation that multiple philosophical frameworks of technological co-constitution, and the concept of constitutive resonance in particular, suggest may be structurally incoherent. The second is that successful harness engineering may amplify known epistemic vulnerabilities — automation bias, trust miscalibration, and the bypassing of evolved mechanisms of critical scrutiny — by producing exactly the conditions under which these vulnerabilities are most acute. And the third is that the rapid adoption of a control-oriented metaphor signals something about the field's conceptual orientation at a moment when the most important questions concern coupling, transformation, co-constitution, and the evolving nature of human–AI relationships.

Metaphors are not merely labels. They are carriers of the assumptions and power relations of the communities that produce them. The speed with which "harness" is being adopted — without apparent critical examination of its entailments — suggests that the engineering community may have found a word that confirms what it already needed to believe: that the relationship between human and AI is one of direction and control, not mutual transformation. Whether that belief will survive contact with the reality of sustained human–AI interaction is among the most consequential open questions the field faces.

References

Anthropic. (2022). Constitutional AI: Harmlessness from AI feedback. https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback (Accessed February 19, 2026).

Anthropic. (2025). Effective harnesses for long-running agents. https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents (Accessed February 19, 2026).

Anthropic. (2026b). Claude's constitution. https://www.anthropic.com/constitution (Accessed February 19, 2026).

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Blythe, M. (2025). Artificial intelligence and other speculative metaphors. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. https://doi.org/10.1145/3715336.3735714

Böckeler, B. (2026). Harness engineering. In *Exploring Gen AI* (series). ThoughtWorks / Martin Fowler. https://martinfowler.com/articles/exploring-gen-ai/harness-engineering.html (Accessed February 19, 2026).

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis,* 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association,* 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gupta, A. (2026, January 8). 2025 was agents. 2026 is agent harnesses. Here's why that changes everything. Medium. https://aakashgupta.medium.com/2025-was-agents-2026-is-agent-harnesses-heres-why-that-changes-everything-073e9877655e (Accessed February 19, 2026).

Hackenburg, K., B. M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D. G. Rand and C. Summerfield (2025). The levers of political persuasion with conversational artificial intelligence. *Science* 390(6777): eaea3884. http://doi.org//10.1126/science.aea3884

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene.* Duke University Press.

Harding, S. G. (1986). *The science question in feminism.* Cornell University Press.

Hashimoto, M. (2026). My AI adoption journey. https://mitchellh.com/writing/my-ai-adoption-journey (Accessed February 16, 2026).

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction,* 4(CSCW2), Article 163. https://doi.org/10.1145/3415234

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors,* 46(1), 50–80.

Maas, M. M. (2023). AI is like... A literature review of AI metaphors and why they matter for policy. AI Foundations Report 2, Institute for Law & AI. https://doi.org/10.2139/ssrn.4612468

Maturana, H. R., & Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding.* Shambhala.

Maynard, A. D. (2026a). Constitutive resonance: AI, the transformation of self, and the narrative structures that reveal what theory cannot. Preprint. http://andrewmaynard.net/papers/constitutive_resonance_preprint_v1.pdf (Submitted to arXiv)

Maynard, A. D. (2026b). The AI Cognitive Trojan Horse: How large language models may bypass human epistemic vigilance. arXiv preprint, arXiv:2601.07085. https://arxiv.org/abs/2601.07085

McKnight, L., & Shipp, C. (2024). "Just a tool"? Troubling language and power in generative AI writing. *English Teaching: Practice & Critique*, 23(1), 23–35. https://doi.org/10.1108/ETPC-08-2023-0092

Mitchell, M. (2024). The metaphors of artificial intelligence. *Science.* https://doi.org/10.1126/science.adt6140

Mollick, E. (2026, February 18). A guide to which AI to use in the agentic era. *One Useful Thing.* https://www.oneusefulthing.org/p/a-guide-to-which-ai-to-use-in-the (Accessed February 18, 2026).

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Lawrence Erlbaum.

Nerlich, B. (2025). AI metaphor studies: An overview. *Making Science Public.* https://makingsciencepublic.com/2025/11/21/ai-metaphor-studies-an-overview/ (Accessed February 16, 2026).

OpenAI. (2026). Harness engineering: Leveraging Codex in an agent-first world. https://openai.com/index/harness-engineering/ (Accessed February 16, 2026).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors,* 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Rees, T. (2025, February 4). Why AI is a philosophical rupture. *Noema.* https://www.noemamag.com/why-ai-is-a-philosophical-rupture/ (Accessed February 16, 2026).

Rosa, H. (2019). *Resonance: A sociology of our relationship to the world.* Translated by J. C. Wagner. Polity Press.

Salesforce. (2026). What is an agent harness? The key to reliable AI. https://www.salesforce.com/agentforce/ai-agents/agent-harness/ (Accessed February 19, 2026).

Schmid, P. (2026). The importance of agent harness in 2026. https://www.philschmid.de/agent-harness-2026 (Accessed February 19, 2026).

Schön, D. A. (1979/1993). Generative metaphor: A perspective on problem-setting in social policy. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 137–163). Cambridge University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119. https://doi.org/10.1111/misp.12032

Sebo, J. (2022). *The moral circle: Who matters, what matters, and why.* All Points Books.

Sequoia Capital. (2026). 2026: This is AGI. https://sequoiacap.com/article/2026-this-is-agi/ (Accessed February 19, 2026).

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stiegler, B. (1998). *Technics and time, 1: The fault of Epimetheus.* Translated by R. Beardsworth & G. Collins. Stanford University Press.

Stiegler, B. (2013). *What makes life worth living: On pharmacology.* Translated by D. Ross. Polity Press.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things.* University of Chicago Press.

AI Use Statement

The ideas, conceptual connections, and core arguments of this paper originated with the author. AI (Claude, Anthropic) was used as a thinking partner throughout the development of the paper, and provided research and drafting support including mapping the genealogy of the term's adoption, surveying relevant literatures, stress-testing the argument's structure, and refining the articulation of the core theses. All sources were verified by the author, and the author takes full responsibility for the paper.