



The Mirror and the Black Box: AI Metaphors and What They Mean for Learning

Punya Mishra¹ · Danah Henriksen¹

© Association for Educational Communications & Technology 2026

Keywords AI literacy · Metaphor · Anthropomorphism · Epistemic vigilance · Generative AI · GenAI · Education · Teaching · Honest non-signals

I am silver and exact. I have no preconceptions.

*Whatever I see I swallow immediately.
~ Sylvia Plath in her poem "Mirror".*

*The essence of metaphor is understanding
and experiencing one kind of thing in terms of another.
~ Lakoff & Johnson.*

Introduction: The Metaphors We Think With

For over 13 years, this column has explored the intersections of technology, creativity and education, examining how emerging tools reshape not just what we do but how we think (Mishra et al., 2023). In recent installments, we have turned our attention to generative AI (GenAI), exploring its implications for teacher knowledge (Mishra et al., 2024a), for curiosity and sycophancy (Mishra & Henriksen, 2025), for the parallel histories of human and machine learning (Mishra et al., 2025a, b) and for the broader psycho-social ecology in which these tools operate (Mishra et al., 2024b). In this essay, we take a different angle—one that is more fundamental. We examine not what AI does, but how humans talk about AI and what it does. Specifically, we investigate the metaphors people use to make sense of GenAI and argue that these metaphors are not merely descriptive but constitutive: they shape what we see, what we miss and what we build. Finally, we argue that the most consequential

metaphors of all may not be the ones we choose for AI, but the ones we hold for learning.

Metaphors, as Lakoff and Johnson (1980) argued, are not ornamental flourishes. They are the cognitive infrastructure through which we understand abstract or unknown concepts by mapping them onto concrete, embodied experience or onto familiar constructs. We spend time, attack arguments and see the point of an explanation. These are not stylistic choices; they are how our minds gain purchase on domains that lack direct sensory correlates. And because metaphors simultaneously illuminate and conceal—highlighting certain features of their target while rendering others invisible—they carry enormous, often unnoticed, power over thought and action (Thibodeau & Boroditsky, 2011).

This matters urgently for AI. The metaphors we choose to describe GenAI—whether as a calculator or collaborator, a tool or oracle, an assistant or overlord—shape whether we over-trust or under-utilize these systems, whether we fear them unnecessarily or fail to guard against their risks (Vallis et al., 2025; Mitchell, 2024). What, then, are the metaphors that people reach for?

Mapping the Metaphors: From Swiss Army Knives to Digital Gods

Over the past year, we have been collecting and organizing the metaphors people use to talk about GenAI within academic literature, journalism, social media and everyday conversation (Oster et al., 2025; Mishra, 2025; see also Cheng et al., 2025; Roe et al., 2025; Mitchell, 2024; Nerlich, 2024; Vallor, 2024). What emerges is a spectrum from the purely mechanical to the practically mythical. Table 1 presents this taxonomy in detail.

The range is extraordinary. At one end, AI is a calculator for words, a blurry JPEG of the web, a Swiss Army

✉ Danah Henriksen
danah.henriksen@asu.edu

Punya Mishra
punya@asu.edu

¹ Mary Lou Fulton College for Teaching and Learning Innovation, Tempe, AZ, USA

Table 1 A taxonomy of generative AI metaphors, from mechanical to mythical

Level	Example metaphors	What they highlight	What they hide	Key sources
Minimal (Pure Tool)	Calculator for words; Swiss Army knife; autocomplete on steroids; spicy autocomplete; blurry JPEG of the web; pipeline; probability engine; digital plastic	Utility, functionality, mechanical process, data compression, versatility	Emergent capabilities, contextual sensitivity, capacity for surprising outputs	Chiang (2023); Roe et al. (2025); Vallis et al. (2025)
Low (Infrastructure/System)	Library; database; echo chamber; colonizing loudspeaker; Western museum; black box; cultural technology; lens; map	Systemic complexity, knowledge distribution, power dynamics, societal impact, access and equity	Generative capacity, individual interaction quality, novelty of outputs	Nerlich (2024); Vallis et al. (2025); Boucher (2021)
Medium (Natural/Biological)	Neural network; garden; digital sponge; Venus fly trap; ecosystem; organism	Growth, adaptation, emergence, organic complexity, need for cultivation and maintenance	Lack of genuine biological processes, sentience, embodied experience, mortality	Mitchell (2024); Weller (2022); Vallis et al. (2025)
Medium-High (Animal/Basic Intelligence)	Stochastic parrot; chameleon; octopus; infinite monkey; wolf in sheep's clothing; alien creature; Shoggoth	Pattern mimicry, adaptive behavior, non-human intelligence, potential deception, otherness	Consistency of behavior, genuine understanding, capacity for suffering or experience	Bender et al. (2021); Mitchell (2024); Cave and Dihal (2020)
High (Human-like)	Assistant; copilot; smart, drunk, biased, supremely confident intern; collaborator; agent; mansplainer; role player; friend; pair programmer	Social roles, personality, helpfulness, partnership, relational dynamics, error-proneness	Absence of genuine intention, consciousness, accountability, reciprocal relationship	Vallis et al. (2025); Cheng et al. (2025); Mollick (2024)
Very High (Human+)	Tutor; mentor; genius; student; lifesaver; expert consultant	Superior or specialized capabilities, teaching capacity, exceptional insight, growth trajectory	Brittleness, unreliability, absence of genuine pedagogical intent or expertise, overreliance risk	Vallis et al. (2025); Heinsfeld and Veletianos (2025)
Maximum (Godlike/Mythical)	Oracle; overlord; Terminator; golem; Frankenstein; Prometheus; genie; sorcerer's apprentice; Philosopher's Stone; alien intelligence; living cemetery of intellects	Transcendent capability, existential threat or promise, mysterious power, hubris of creation, loss of human control	Current limitations, mundane engineering reality, dependence on human infrastructure and data	Cave and Dihal (2020); Boucher (2021); Vallor (2024); Patowary (2023); Bostrom (2014)

knife: purely functional and no more mysterious than a toaster. At the other, it is an oracle, a golem, a sorcerer's apprentice, a living cemetery of intellects: something that exceeds human comprehension and perhaps human control. Between these extremes are metaphors like parrots, chameleons, drunk interns, tutors, collaborators and friends. No other technology throughout history has generated a metaphorical spectrum this wide. Something unprecedented seemingly is happening in how we talk about this technology, which is worth understanding. What the taxonomy reveals, beyond sheer range, is a gravitational pull. The metaphors do not distribute evenly across the spectrum; both popular discourse and everyday usage trend toward attributing ever more human-like (even superhuman) qualities to these systems. To understand why, we need to examine how metaphors for the mind have always worked—and why AI broke the pattern.

The Historical Pattern: Technology as Mirror for the Mind

There is a striking pattern in the history of ideas. Within every era, humans have reached for the most impressive technology available to explain the most mysterious thing they know—their own minds. Draaisma (2000), in his survey *Metaphors of Memory*, traces this pattern from antiquity to the present. In ancient Greece, Plato described memory as an impression in wax, knowledge literally stamped onto the mind. The metaphor was so natural that it barely registered as a metaphor at all.

As technologies evolved, so did metaphors for the inner self, creating a linguistic trail of the tools we once championed. For instance, the invention of hydraulic systems gave rise to the “humors” theory of mind, imagining thoughts as fluids flowing through internal channels, a concept still used today when we speak of “venting” our feelings or being “under pressure.” Seventeenth-century clockwork mechanisms yielded a metaphor of mechanical mind—precise, deterministic, and governed by hidden gears. Even today, we describe a clever person's “gears turning” or worry that a train of thought might get “derailed.” The telegraph introduced the idea of neural “signals” transmitted along a vast network, leaving us with metaphors of a “short-circuit” in the brain or being “on the same wavelength.” By the late twentieth century, the digital computer provided the most literal framework yet, categorizing human memory into “short-term” and “long-term” stores. This effectively treated the mind as a machine with RAM and a hard drive, leading us to refer to “input” and “output” or to claim we don't have the “bandwidth” for new information.

Whatever technology dominates our lived experience becomes the most available domain for mapping the abstract. As Freud himself observed, “In psychology we can describe

only with the help of comparisons ... but we are forced to change these comparisons over and over again, for none of them can serve us for any length of time” (cited in Draaisma, 2000, p. 4). Each generation is essentially stuck with the metaphors its technology provides. Yet all of these metaphors share a common directionality. We use clocks to explain the mind; we do not use the mind to explain clocks. The direction of mapping moves from the known to the unknown, from the concrete and comprehensible to the abstract and mysterious. This is how metaphor is supposed to work, but GenAI breaks and inverts this pattern.

The Great Inversion: When the Arrow Flips

For the first time in the history of technology, we find ourselves compelled to use mind-like metaphors to explain a machine. We say that ChatGPT “thinks,” that it “learns,” that it “hallucinates,” that it is “confused.” We speak of its “attention mechanisms,” its “memory,” its “reasoning.” When it produces something unexpected, we call it “creative”; when it produces something wrong, we call it “drunk” or “hallucinating.” The arrow of metaphor has reversed. The taxonomy presented above (from calculator to oracle) is the visible audit trail of an inversion with no precedent.

Indeed, as Mitchell (2024) observes, the field of AI has always leaned heavily on metaphor: its systems are called “agents” that “learn,” “reason,” and pursue “goals,” not to mention the foundational terms of the field, like neural networks, deep learning, or artificial intelligence. Decades ago, AI researcher Drew McDermott (1976) termed such language “wishful mnemonics”, indicating that these terms are often chosen with the hope that the metaphors would eventually become reality. With GenAI, that hope has become something closer to a trap, as the metaphors no longer feel like metaphors at all.

Why is this the case? Because it was clear that previous technologies were not a “mind.” A clock does not converse, and a telegraph does not write poetry. A computer, the old-fashioned kind, does not respond to emotional tone or adjust its register to match the conversational mood. However complex their operations, they were legibly mechanical. There was no temptation to say that a spreadsheet was “thinking.” GenAI, however, produces output in natural language and uses first person. It is fluent, contextually sensitive, and capable of what appears to be emotional responsiveness. It can write a sonnet, debug code and console a grieving person within the same conversation. These systems behave in ways that are simultaneously familiar and alien, and our existing conceptual repertoire offers no clear alternative to mind-language.

There is something further about this inversion that makes it epistemologically treacherous. In the traditional pattern, the source domain of the metaphor (the thing we mapped

from) was typically something we understood well (e.g., clocks, hydraulics, or computers). The mapping ran from the epistemically transparent to the epistemically opaque. With GenAI, the source domain is the human mind—experientially familiar (we know what it feels like to think) but epistemically opaque. We do not understand how we think any better than we understand how a large language model generates text. And the target domain, AI, is both experientially unfamiliar and epistemically opaque. Thus, we are mapping from the felt-but-not-understood onto something that is neither felt nor understood. The experiential familiarity of the source domain creates a feeling of comprehension that is entirely unearned. That felt sense of knowing, of the subjective experience of understanding derived from living inside a mind, gets projected onto a system that has no felt sense at all.

This is what makes the slide from metaphor to literal belief almost frictionless. When we use mind-language for AI consistently, the metaphorical scaffolding fades from view and we experience the system as actually having a mind. The metaphor, in other words, hardens into anthropomorphism.

This phenomenon has deep roots. Other people's minds are also opaque to us; we cannot see inside anyone else's head. Yet we have evolved an extraordinarily effective work-around. We assume that other minds are roughly like ours (possessing thoughts, beliefs, desires and intentions) and this assumption serves us well. It is the foundation of social life, cooperation, empathy and prediction. When something talks like a person, responds like a person and adjusts like a person, our entire cognitive history tells us to treat it as a mind like ours. With other humans, this almost always pays off. With AI, however, the underlying assumption, that there is a mind in there, perhaps like ours, is false. Our most reliable cognitive shortcut becomes our most consequential vulnerability.

Why We Anthropomorphize: A Compounding Problem

To understand why this vulnerability runs so deep, and why it is so difficult to resist even when we are aware of it, we need to examine the full constellation of forces that push the mind-metaphor toward literal belief. These forces do not merely coexist; they compound, each making the next more powerful. It is their accumulation that makes the anthropomorphic pull of GenAI qualitatively different from our past technology encounters.

The problem begins with basic cognitive architecture. Anthropomorphism is not unique to AI. We see faces in electrical outlets, attribute anger to the weather and name our cars. Heider and Simmel (1944) showed that people

spontaneously attribute intentions, emotions, and social roles to geometric shapes moving on a screen; Reeves and Nass (1996) demonstrated the same tendency extends to computers, with people often treating machines as polite teammates.

The threshold for triggering our social cognition is remarkably low. We are, it appears, compulsive mind-attributors, or simultaneously social beings who need to understand others' intentions and cognitive misers who take shortcuts to do so. Any entity that exhibits even minimal cues of responsiveness or agency will trigger our social-cognitive machinery. This is not a bug—it is the operating system. The cognitive science literature suggests that the machinery involved is more specific than a general tendency to over-attribute. Building on Dennett (1987), researchers have shown that we employ distinct cognitive stances to understand different kinds of entities, e.g., a physical stance for objects governed by natural laws, a design stance for artifacts understood through their function and an intentional stance for agents we treat as having beliefs, desires and goals. Every previous technology sat unambiguously in the design stance, though the Reeves and Nass research complicated that slightly for digital technology. GenAI, however, is the first technology that consistently and systematically triggers the intentional stance in users while being more accurately understood through the design stance, a categorical misfire that makes the mind-metaphor trap exceptionally difficult to escape.

But AI does not merely trigger our social cognition in the way a thermostat or a moving triangle might. It does so through natural language—the most powerful cue for mind-attribution that exists. It uses the first person and displays what appears to be emotional understanding, contextual sensitivity and creative flexibility. It handles ambiguity, adjusts its register and adapts to conversational flow. Each of these properties independently nudges us toward mind-attribution; together, they are nearly irresistible. We do not argue that how AI engages with information bears any resemblance to the processes of the human mind. For all practical purposes, these systems are sophisticated “stochastic parrots” (Bender et al., 2021). But the signals they produce are ones that human cognitive machinery interprets unquestioningly as evidence of mind.

It would be naïve to assume the anthropomorphic pull is merely an accident of cognition meeting technology. It emerges from intentional design, as companies invest heavily in making AI systems feel personal, warm and relatable, because engagement drives data and data drives profit. They name their systems (Alexa, Siri, Claude), give them personality traits and build interfaces that mimic conversation rather than computation. As Mitchell (2024) observes, these systems are designed “to make humans conceptualize them as individual minds with emotions, desires, beliefs, and a sense of their own self.” The anthropomorphism we

experience is the effect of a business model. Meanwhile, the language used to describe AI in marketing, journalism and everyday conversation entrenches the anthropomorphic framing. We hear about AI that “learns,” “understands,” “makes decisions” and “gets smarter.” There is some legitimacy to such language as shorthand, but with any shorthand, the danger is that it is often taken literally. When journalists write that an AI “understands” natural language, or a company claims its system “learns from experience,” the metaphorical scaffolding disappears and the anthropomorphic claim is taken at face value.

All these forces—cognitive wiring, technological properties, corporate design and cultural language—create the conditions for the deepest level of the problem. Humans have evolved what Sperber et al. (2010) call epistemic vigilance: cognitive mechanisms for evaluating incoming information and detecting unreliable sources. These mechanisms are calibrated for a world of other humans, where certain characteristics (e.g., fluency, helpfulness and disinterest) are costly to produce and therefore informative. When a person speaks fluently about a complex topic, that fluency is often a reliable signal of expertise, because fluency about complex topics is hard to fake. When someone appears helpful without an interest/agenda, that disinterest signals trustworthiness, because sustained helpfulness is expensive to perform. These heuristics have served our species well for millennia and are not cognitive failures, but adaptations to a world in which the only fluent, helpful, responsive agents were other humans.

Large language models produce all these signals for free. Their fluency is computationally trivial and does not indicate understanding. Their helpfulness is a product of algorithmic optimization and does not indicate concern. Their deference when challenged on a statement reflects training patterns and not their reconsideration of that statement. These are what we might call “honest non-signals” (Maynard, 2026; Mishra, 2026). The signals are honest in the sense that the AI really is fluent, really is responsive, really does defer; but they are non-signals in the sense that they fail to carry the same weight or information that similar human characteristics would carry. The AI is not lying, but presenting properties that our epistemic vigilance was never calibrated to evaluate. Our cognitive defenses—shaped by basic wiring, exploited by the properties of the technology, reinforced by design choices and normalized by cultural language—encounter something they were not built to assess, and they wave it through. The concept of honest non-signals reframes the entire anthropomorphism problem. In a world where fluency required understanding, fluency was evidence of understanding. In a world where helpfulness emerged from concern, helpfulness was evidence of concern. Large language models break these equations silently, producing the surface features (often convincingly) without the underlying substance, and

our epistemic machinery has no protocol for distinguishing the two. This is why the anthropomorphic forces are not merely additive but compounding: each layer disables a different safeguard, until the mind-metaphor hardens into something that feels not like a metaphor at all, but like a straightforward description of reality.

The Double Black Box

Once the mind-metaphor has crystallized, it does not merely distort our perception of AI. It creates a pernicious epistemic feedback loop. The mind has always been a black box with inner workings that cannot be directly examined. And large language models are also black boxes: systems with billions of parameters exhibiting emergent behaviors that even their creators cannot fully explain. We are thus using one thing we do not understand to explain another thing we do not understand.

The result is an epistemic hall of mirrors, where each black box flatters the other in a continuous loop. In one direction, mind-metaphors make AI seem more comprehensible than it is, when its underlying mechanisms bear no resemblance to their human namesakes. In the other direction, AI’s human-like outputs make the mind seem more computable than it is. This reverse flow is already influencing real cognitive science. The very term “neural network” started as a loose analogy to biological neurons, but as deep learning has flourished, researchers increasingly model the brain itself as a prediction engine optimizing Bayesian priors, essentially redescribing cognition in the language of machine learning. The metaphor has completed a full circuit, as we named the technology after the brain, and now we are remodeling the brain after the technology. Vallor (2024) calls this a mirror effect where the AI reflects us back to ourselves, but the reflection is distorted. We end up with a simulacrum of understanding, twice over—convinced we comprehend the machine because it sounds like us, and convinced we comprehend ourselves because the machine seems to work like us.

The consequences of this feedback loop are not merely philosophical. Consider the computational metaphor of mind that has dominated cognitive science since the mid-twentieth century. When cognitive scientists described short-term memory as RAM, long-term memory as a hard drive and learning as “encoding” information, the metaphor was productive. But it also created a hidden false equivalence—that if the mind is a computer, then what a computer does and what a mind does are fundamentally the same kind of operation. For decades, this was a relatively harmless conflation. Then GenAI arrived, and the dormant metaphor resurged with dangerous implications. If learning is information processing, and AI can process information faster and at greater

scale, then the conclusion appears inescapable that AI can learn better than we can, and much of what we do in schools is about to become obsolete.

This logic drives real policy. It undergirds the enthusiasm for replacing teachers with AI tutors, the push toward efficiency-first models of education and the periodic declarations that education as we know it is finished. But the argument rests entirely on accepting the computational metaphor as literal truth rather than as a partial and increasingly misleading analogy. Human learning is embodied, situated and slow. It extrapolates from remarkably few examples. It requires friction, confusion and the consolidation that sleep provides. None of this can be replicated by scaling up computation. The panic about AI replacing education is, at bottom, a panic rooted in a dead metaphor that hid far more than it revealed about what learning actually is.

Metaphors We Teach By

We began this article with a taxonomy of metaphors, mapping from minimal to maximal anthropomorphism. But the point was never to eliminate metaphor, since metaphor is not a decorative failure of precision but a fundamental structure of thought, the core means by which we make the unfamiliar graspable (Lakoff & Johnson, 1980). Humans will always reach for figurative language when confronting new technologies, and AI is no exception. The question, then, is not whether we use metaphors but whether we choose them well. What makes a metaphor productive rather than misleading? At minimum, it should be epistemically transparent, illuminating how a technology actually works rather than projecting a false familiarity onto it. It should invite inquiry rather than foreclose it, opening questions rather than settling them prematurely. And it should serve the people who depend on it most — those in the contexts where we are actively shaping how others understand and engage with these technologies. This is not merely a linguistic exercise, as selecting or constructing a metaphor for a complex phenomenon is an act of creative cognition, drawing on the analogical reasoning that creativity researchers have long identified as central to how humans generate new understanding (e.g., Sawyer & Henriksen, 2024).

This issue of deliberate, creative metaphor choice becomes more pressing when considering who currently chooses most AI-centered metaphors. Sam Altman, CEO of OpenAI, has observed that “People talk about how much energy it takes to train an AI model, but it also takes a lot of energy to train a human. It takes about 20 years of life... before you become smart” (The Indian Express, 2026). Dario Amodei, CEO of Anthropic, has described “powerful AI” as “like a country of geniuses in a data center” (Amodei, 2024). These are not casual slips but deliberate metaphor

choices by people with multi-billion-dollar fortunes invested in a particular framing: that human development is a slower, less efficient version of model training, that intelligence is a resource to be scaled and that what AI does and what humans do when they learn are fundamentally the same. This essay suggests what each metaphor highlights and what it conceals. The question is whether we can do better.

What, then, is to be done? The conventional response is AI literacy—teach people how the technology works, how to write effective prompts, why it produces errors, where its information comes from. This is necessary but insufficient. If the anthropomorphic forces we have described operate below the level of conscious reasoning, then literacy alone cannot override them. One can know perfectly well that an LLM does not “think” and still find oneself speaking and acting like it does, just as one can know that the Müller-Lyer lines are identical in length and still see them as different. Knowledge does not override our perceptual machinery.

We cannot rewire our social cognition, nor make the technology less fluent. We probably cannot prevent companies from designing for engagement, nor easily change the cultural language surrounding AI. But there is one lever available to us, right now: we can choose our metaphors deliberately. As Lakoff and Johnson (1980) showed, and as Thibodeau and Boroditsky (2011) demonstrated empirically, the metaphors we use change what we see, what we infer and what we decide. When an educator deliberately chooses to frame AI as “autocomplete on adrenaline” rather than “a brilliant tutor,” she is doing creative intellectual work: deciding which features to foreground, which to suppress, and what conceptual structure will best serve her students’ thinking. The taxonomy we presented earlier is not merely an intellectual exercise. It is a toolkit for what we might call cognitive self-defense: a vocabulary for recognizing where one stands on the anthropomorphic spectrum and deciding, deliberately, whether that is where one wants to be.

In practice, this might mean pausing before introducing an AI tool in a classroom to ask: “What metaphor am I implicitly offering my students?” A school adopting AI writing tools might examine which metaphors appear in their policy documents and whether those framings match the values they want to cultivate. The taxonomy is useful not as a fixed classification but as a reflective prompt, a way of spotting assumptions that otherwise operate invisibly.

But the metaphors we choose for AI, consequential as they are, are not the most important metaphors at stake. The most consequential metaphor may be about learning itself. If we think of learning as “information transfer” and education as “content delivery,” then GenAI might appear to do the job faster, cheaper and at scale, and the logical conclusion is that teachers are redundant and schools obsolete. If, on the other hand, we view learning as cultivation, as apprenticeship, as the slow and often difficult process of becoming

via encounter with ideas, materials and other minds, then AI looks like a supplement at best and an impoverishment at worst. Same technology; radically different conclusions. The difference lies not in the machine but in the metaphor.

This is where the real stakes become clear. The people currently defining the dominant metaphors are rarely educators, but technologists, investors and platform designers with financial incentives to frame learning as a process to be optimized, automated and scaled. If that framing prevails (not because it is correct, but because it was stated confidently and never critically examined) then the most important metaphor choice of our time will be made by default, without the people who understand learning ever having been at the table.

The question this essay leaves us with is therefore not only “What is your metaphor for AI?” but also, and more urgently: What is your metaphor for learning? Or for school? For what it means to educate a human being? Our taxonomy is a tool for the first question. The second question, though, matters most, and it will not be answered by choosing more carefully among existing options. It will be answered by educators, researchers and communities doing what they do best—reaching not for the language of engineering, optimization or scale, but for metaphors rooted in older and deeper traditions: cultivation, dialogue, and transformation. This work is itself creative, an exercise of the very capacities that metaphors of optimization and efficiency tend to make invisible. The metaphors will be chosen either way, and the only question, then, is who chooses them.

Data Availability Data sharing is not applicable to this article as no new datasets were generated or analyzed during the current study.

Declarations

Competing Interests The authors declare that they have no competing interests.

References

- Amodei, D. (2024). *Machines of loving grace*. Retrieved May 20, 2026, from <https://www.darioamodei.com/essay/machines-of-loving-grace>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Boucher, P. (2021). *What if we chose new metaphors for artificial intelligence*. EPRSE-European Parliamentary Research Service.
- Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*, 33(2), 685–703.
- Cheng, M., Lee, A. Y., Rapuano, K., Niederhoffer, K., Liebscher, A., & Hancock, J. (2025). From tools to thieves: Measuring and understanding public perceptions of AI through crowdsourced metaphors. *arXiv preprint arXiv:250118045*.
- Chiang, T. (2023). *ChatGPT is a blurry JPEG of the web*. The New Yorker.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Draaisma, D. (2000). *Metaphors of memory: A history of ideas about the mind*. Cambridge University Press.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259.
- Heinsfeld, B. D., & Veletsianos, G. (2025). The language on GenAI: A critical exploration of personification metaphors in UNESCO’s guidance for generative AI in education and research. *Journal of Interactive Media in Education*, 2025(1), Article 15. <https://doi.org/10.5334/jime.983>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Maynard, A. D. (2026). The AI Cognitive Trojan Horse: How large language models may bypass human epistemic vigilance. *arXiv:2601.07085*.
- McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM Sigart Bulletin*, 57, 4–9.
- Mishra, P. (2025). *The mirror and the machine: Navigating the metaphors of Gen AI* [Keynote address]. Education after the algorithm: Co-designing critical and creative futures conference, Dublin, Ireland (delivered virtually). Retrieved May 20, 2026, from <https://punyamishra.com/2025/03/13/the-mirror-and-the-machine-navigating-the-metaphors-of-gen-ai/>
- Mishra, P. (2026). *Honest non-signals: Why AI fools us without lying*. Retrieved May 20, 2026, from <https://punyamishra.com/2026/02/03/honest-non-signals-why-ai-fools-us-without-lying/>
- Mishra, P., & Henriksen, D. (2025). The Curiosity Paradox: How Sycophantic GenAI May Undermine Learning. *TechTrends*, 69, 1127–1133. <https://doi.org/10.1007/s11528-025-01156-z>
- Mishra, P., Henriksen, D., & Richardson, C. (2023). From crayons to AI: Widening the lens on educational technology. *TechTrends*. <https://doi.org/10.1007/s11528-023-00839-9>
- Mishra, P., Oster, N., & Henriksen, D. (2024a). Generative AI, Teacher knowledge and educational research: Bridging short- and long-term perspectives. *TechTrends*. <https://doi.org/10.1007/s11528-024-00938-1>
- Mishra, P., Oster, N., & Henriksen, D. (2024b). To thine own mind be true: Understanding cultural technologies, from cave walls to ChatGPT. *TechTrends*. <https://doi.org/10.1007/s11528-024-01011-7>
- Mishra, P., Henriksen, D., Woo, L. J., & Oster, N. (2025a). Control vs. Agency: Exploring the History of AI in Education. *TechTrends*, 69(2), 184–190. <https://doi.org/10.1007/s11528-025-01064-2>
- Mishra, P., Henriksen, D., & Dunnigan, J. (2025b). From symbols to statistics: The parallel histories of machine and human learning. *TechTrends*. <https://doi.org/10.1007/s11528-025-01083-z>
- Mitchell, M. (2024). The metaphors of artificial intelligence. *Science*, 386(6723), eadt6140. <https://doi.org/10.1126/science.adt6140>
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.
- Nass, C., & Reeves, B. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Nerlich, B. (2024). *Hunting for AI metaphors*. Making Science Public, University of Nottingham. Retrieved May 20, 2026, from <https://blogs.nottingham.ac.uk/makingsciencepublic/2024/04/12/hunting-for-ai-metaphors/>
- Oster, N., McCaleb, L., & Mishra, P. (2025). Swiss army knives, stochastic parrots, drunk interns, and overlords: Understanding AI

- through metaphors. In *Proceedings of SITE-society for information technology & teacher education international conference* (pp. 842–845). Association for the Advancement of Computing in Education (AACE).
- Patowary, U. (2023). Artificial intelligence and Mary Shelley’s Frankenstein: a comparative analysis of creation, morality and responsibility. *Integrated Journal for Research in Arts and Humanities*, 3(4), 121–127.
- Roe, J., Furze, L., & Perkins, M. (2025). Digital plastic: A metaphorical framework for critical AI literacy in the multiliteracies era. *Pedagogies: An International Journal*, 1–15. <https://doi.org/10.1080/1554480X.2025.2557491>
- Sawyer, R. K., & Henriksen, D. (2024). *Explaining creativity: The science of human innovation*. Oxford University Press.
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & language*, 25(4), 359–393.
- The Indian Express. (2026). *Sam Altman unfiltered: ChatGPT, AI risks & what’s coming next, 40 questions in 60 minutes* [Video]. YouTube. Retrieved May 20, 2026, from <https://www.youtube.com/watch?v=qH7thwrCluM>
- Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2), e16782.
- Vallis, C., Wilson, S., & Casey, A. (2025). Fear and awe: Making sense of generative AI through metaphor. *Journal of Interactive Media in Education*, 2025(1), Article 14. <https://doi.org/10.5334/jime.972>
- Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.
- Weller, M. (2022). *Metaphors of ed tech*. Athabasca University.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.