AI & SOCIETY (2026) 41:5177–5189
https://doi.org/10.1007/s00146-026-02875-4

OPEN FORUM

Personalised LLMs and the risks of the digital twin metaphor

Marco Annoni[1] · **Davide Battisti**[1] · **Beatrice Marchegiani**[2]

Received: 22 September 2025 / Accepted: 14 January 2026 / Published online: 29 January 2026

Abstract
Can an AI truly be your digital twin? Technology companies, startups, and even academic researchers increasingly claim so. From grief-bots that promise to let you talk with deceased loved ones to clinical tools designed to predict patients' treatment preferences, personalized Large Language Models are being marketed as faithful replications of individual identity, personality, and values. The digital twin label—borrowed from industrial engineering, where it describes computational models precisely mirroring physical systems—lends these claims an aura of scientific credibility. But is this metaphor appropriate, or is it dangerously misleading? This paper argues that applying the digital twin metaphor to personalized LLMs constitutes a systematic mischaracterization with serious ethical consequences. To this end, we first outline three plausible interpretations of the PDT metaphor—behaviorist, representational, and phenomenal. We then show that current AI systems do not satisfy the conditions of any of these interpretations, rendering the PDT metaphor ultimately inappropriate. The consequences of this metaphorical overreach are far from abstract. When vulnerable individuals—grieving families, patients facing incapacity, people seeking psychological support—interact with systems marketed as preserving human essence, they risk forming attachments to sophisticated illusions. The metaphor fosters misplaced trust, distorts public understanding of AI capabilities, and shapes policy debates on false premises. As these technologies proliferate into healthcare, legal proceedings, and intimate relationships, metaphorical precision becomes an ethical imperative. We conclude by proposing alternative frameworks that honestly represent what these systems can and cannot do.

Keywords Digital twins · Large Language Models · Personalized AI · Scientific metaphors · AI ethics · Anthropomorphism · Griefbots

1 Introduction

The recent rise of generative AI and large language models (LLMs) has driven a wave of efforts to create personalized computational systems replicating specific individuals. This technological phenomenon encompasses diverse applications ranging from commercial "griefbots" that simulate deceased individuals to proposed clinical decision support systems to predict patient treatment preferences (Hollanek and Nowaczyk-Basińska 2024; Earp et al. 2024; Danaher and Nyholm 2024, 2025). Despite their heterogeneous implementations and varied purposes, these systems are increasingly unified under a common metaphorical framework: they are marketed and conceptualized as "digital twins".

The genealogy of the digital twin concept traces back to industrial applications, where virtual models maintain correspondence with physical systems through continuous integration of sensor data. These models track quantifiable parameters—temperature gradients, pressure differentials, structural stress patterns—in domains spanning manufacturing, aerospace engineering, and biomedicine. Recently, however, this framework has been appropriated for a fundamentally different purpose: describing LLMs configured to mimic the cognitive, linguistic, and behavioural patterns of specific individuals, termed "Personalized Digital Twins" (PDTs) by Jimenez et al. (2020). However, this appropriation occurred without a strong and rigorous conceptual analysis

✉ Marco Annoni
marco.annoni@cnr.it

Davide Battisti
davidebattisti@cnr.it

Beatrice Marchegiani
beatrice.marchegiani@kellogg.ox.ac.uk

[1] National Research Council, Rome, Italy

[2] University of Oxford, Oxford, United Kingdom

of the metaphor, which is necessary to evaluate its actual appropriateness and potential ethical implications.

This paper advances the thesis that the application of the digital twin metaphor to personalized LLM-based models constitutes not merely a semantic imprecision but an epistemologically and ethically problematic mischaracterization.

More specifically, the paper is structured as follows. In Sect. 2, we discuss the function and relevance of metaphors in scientific discourse, then tracing the history of the digital twin metaphor's evolution. In Sect. 3, we review the current usage of the PDT metaphor in the debate, highlighting its intuitive rather than analytical use. In Sect. 4, we propose three interpretations of what is required of an LLM system for the PDT metaphor to make sense: behaviourist, representational, and phenomenal interpretations. In Sect. 5, we use a technical analysis to evaluate whether the criteria described above are met by the current technical capabilities of these systems, arguing that the PDT metaphor systematically misrepresents personalized LLM-based systems. In Sect. 6, we argue that such misrepresentation creates the conditions for substantial harm in high-stakes decision-making contexts. After looking at possible alternative conceptual frameworks to the PDT metaphor in Sect. 7, we conclude by arguing that our analysis demonstrates how understanding the structural inadequacies of this metaphor illuminates the necessity for more research into developing suitable conceptual frameworks to accurately characterize these emergent computational applications.

2 The epistemological function and limitations of scientific metaphors

Metaphorical reasoning is pervasive in human cognition (Thibodeau and Boroditsky 2011). As Lakoff and Johnson (1980, p.3) noted in their seminal work, "Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature". Metaphors operate as cognitive scaffolding, facilitating comprehension of novel phenomena through systematic mapping from familiar source domains to unfamiliar target domains.

This metaphorical structuring is particularly pronounced in scientific discourse, where abstract concepts often require concrete analogical grounding (Taylor and Dewsbury 2018). Molecular biologists conceptualize genetic processes through textual metaphors of "reading," "transcribing," and "translating" DNA sequences; quantum physicists describe subatomic phenomena as both "particles" and "waves" despite these entities' fundamental dissimilarity to macroscopic analogs; evolutionary biologists employ arboreal metaphors—"trees of life"—to represent phylogenetic relationships. The designation "artificial intelligence" itself exemplifies this tendency, attempting to render comprehensible a novel class of computational phenomena through the conjunction of two established conceptual categories: the synthetic ("artificial") and the cognitive ("intelligence")—a process that Floridi and Nobre have recently defined as "conceptual borrowing" (2024).

The distinction between metaphor and simile carries significant cognitive implications. While similes preserve explicit comparative structure through linguistic markers ("like", "as"), metaphors effectuate conceptual fusion between tenor (the target domain) and vehicle (the source domain). This linguistic compression—asserting the heart *is* a pump rather than functions *like* a pump—generates cognitive efficiency but simultaneously introduces systematic ambiguity. The intended mappings between source and target domains often remain underspecified, permitting unwarranted property transfers from vehicle to tenor. This ambiguity generates both heuristic utility and epistemic hazard (Pigliucci and Boudry 2011). The pump metaphor productively guides cardiovascular medicine by highlighting mechanical properties of cardiac function; conversely, the computer metaphor for brain function, while generative in certain contexts, has demonstrably constrained neuroscientific theorizing by imposing inappropriate computational frameworks onto biological neural processes. As Ahmedien (2023, p.314) articulates, metaphors can function as cognitive constraints, "obstructing perspectives, limiting understanding, and confining choices".

The specific hazard of anthropomorphic metaphors in AI discourse warrants particular attention. As Salles et al. (2020) demonstrate, anthropomorphic framing of AI systems systematically misleads users regarding system capabilities, potentially leading to forms of "over trust" in automated systems. Anthropomorphic metaphors in AI may create a "double bind": they simultaneously facilitate initial understanding while obscuring the fundamental mechanistic differences between human cognition and computational processes.

The application of digital twin terminology to personalized LLM systems represents a particularly consequential instantiation of metaphorical reasoning. Unlike established scientific metaphors that have undergone decades of critical refinement and empirical validation, this conceptual framework has achieved rapid adoption absent systematic scrutiny, despite its profound implications for fundamental concepts including personal identity, consciousness, and the ontological status of human relationships in digital contexts (Voinea et al. 2025). As we argue below, the velocity of this metaphor's migration—from industrial control systems to putative replications of human identity within a single decade—has manifestly exceeded our capacity for critical evaluation. To appreciate the extent of this conceptual strain, it is necessary to first examine the origins and original function of the digital twin concept in its native engineering context.

2.1 The evolution from industrial digital twins to personalized LLM-based systems

The digital twin paradigm emerged from computational simulation discourse in the 1990s (Gelernter 1991, Braun 2021) and achieved concrete instantiation in aerospace engineering applications. NASA's 2010 formalization defined digital twins as "ultra-realistic probabilistic simulations" that mirror the complete lifecycle of air vehicles (Grieves and Vickers 2017). This conceptualization established digital twins as dynamic computational models that evolve with their physical counterparts through continuous data assimilation, enabling predictive analytics and system optimization.

Digital twin applications have since proliferated across manufacturing, urban planning, healthcare, and other domains, yet retain consistent fundamental characteristics: they model physical systems with measurable parameters, generate empirically testable predictions, and maintain temporal synchrony through continuous data integration. Fidelity—the quantifiable correspondence between model behaviour and physical system dynamics—constitutes the fundamental evaluative criterion. Physical digital twins must satisfy two necessary conditions: a virtual representation of a physical system and continuous data streams reflecting real-time state changes. Under this definition, an automotive dashboard display updating from vehicular sensors instantiates a legitimate digital twin; a static three-dimensional model, regardless of geometric precision, does not.

Concurrently, a parallel transformation was occurring in artificial intelligence research. The 2012 breakthrough in deep neural network architectures, culminating in the 2022 release of ChatGPT, demonstrated unprecedented capabilities in human-like text generation. These transformer-based architectures produce linguistically coherent, contextually appropriate text often indistinguishable from that written by humans. Seeking analogical frameworks to conceptualize these novel capabilities, researchers and entrepreneurs began appropriating the digital twin framework, asserting that LLMs could mirror not only mechanical systems but purportedly human-specific phenomena: natural language competence, reasoning patterns, personality traits, individual psychology, and identity.

This appropriation represents a categorical discontinuity with the original concept. The relationship between an industrial digital twin and its physical referent is unambiguous: no engineer conflates a turbine's computational model with the physical turbine itself. With PDTs, however, the metaphor invites precisely such conflation—users may perceive themselves as interacting with the actual person or an ontologically equivalent entity. The system is framed not as a representation but as a replication of personal identity.

3 Contemporary applications: from commercial griefbots to academic proposals for digital immortality

The trajectory from physical digital twins to personalized LLM systems reveals a progressive erosion of the empirical constraints that originally defined the twin concept. Understanding this historical evolution is essential for two reasons. First, it demonstrates how metaphorical drift—the gradual expansion of a metaphor beyond its original domain—can fundamentally alter a concept's epistemic status. Second, it illuminates how the rhetorical authority accumulated by digital twins in industrial contexts has been strategically appropriated to legitimize fundamentally different technologies. Contemporary PDT applications exploit this semantic slippage, leveraging the credibility of industrial digital twins while abandoning their defining characteristics of measurable fidelity and empirical validation.

Current implementations of personalized LLM-based digital twins span an extensive spectrum from commercial bereavement technologies to academic proposals. Commercial applications dominate current deployment. Eternos.life markets systems that purportedly ensure an individual's "essence lives on in perpetuity". HereAfter AI, StoryFile, and You, Only Virtual integrate conversational AI with multimedia archives to facilitate posthumous interaction—technologies critics characterize as "digital necromancy" (Morse 2023). Tavus.io advances more ambitious claims, processing email corpora and social media histories to generate what they term "high-fidelity personality replications" for applications in customer service, therapeutic intervention, and social companionship. These "griefbots" or "thanabots" are marketed as technologies capable of achieving digital resurrection through AI systems trained on individuals' digital footprints. Significantly, they all rely on the twin metaphor (Hollanek and Nowaczyk-Basińska 2024).

The promotional discourse surrounding these applications reveals fundamental conceptual conflations. When Eternos claims to preserve "essence" or Tavus promises "emotionally intelligent AI that looks and seems real", these assertions transcend mere marketing hyperbole—they constitute implicit ontological claims regarding the nature of identity and consciousness that exceed their systems' technical capabilities. Within these contexts, the twin metaphor functions not as incidental branding but as a foundational framework shaping both user expectations and system architecture. While marketing has long relied on non-factual claims, the endorsement of the digital twin metaphor in some academic literature—combined with the complexity of the technology and the generally low level of digital literacy among users—makes this metaphor especially likely to mislead.

Academic proposals exhibit comparable ambition. Giubilini et al. (2024) propose iSAGE, characterized as a "digital ethical twin" functioning as a "real-time consultative interface that helps individuals make morally sound decisions based on their own evolving set of values and beliefs". Earp et al. (2024) conceptualize a "personalized patient preference predictor"—a digital "psychological twin" designed to infer treatment preferences from digital footprints when patients lack decisional capacity. Xie et al. (2025) introduce PsyDT, a framework that constructs digital twins of psychological counsellors with personalized counselling styles, capturing individual therapists' unique linguistic patterns and therapeutic techniques through LLM-based modelling. Meanwhile, Fawkes and Burden (2025) have proposed developing "Digital Human Twins" for defence applications, including training, mission planning, operational decision-making, health monitoring, equipment design and acquisition, and human–machine teaming. Most ambitiously, Iglesias et al. (2024) propose "digital doppelgangers" that purportedly extend some dimensions of individual existence and personhood by maintaining interpersonal relationships beyond biological death. Even critiques of this proposal continue to employ the term 'digital twin' (Battisti 2025).

These applications demonstrate considerable technical sophistication and address legitimate human needs: grief processing, healthcare accessibility, and memory preservation. Nevertheless, their reliance on the twin metaphor implies capabilities that current technology cannot substantiate. The fundamental issue concerns representational adequacy: existing LLM architectures cannot achieve the fidelity standards that "digital twin" nomenclature implies. This representational overreach transforms potentially beneficial tools into instruments of systematic deception, particularly when vulnerable individuals in high-stakes contexts such as healthcare engage with systems that promise genuine psychological correspondence while delivering only algorithmic pattern matching.

4 Analysing the digital twin metaphor: a systematic deconstruction

The preceding analysis has documented how the digital twin metaphor has proliferated across bioethical and philosophical discourse, often deployed based on intuitive appeal rather than systematic conceptual analysis. Consequently, the necessary and sufficient conditions for appropriate application of this metaphor remain critically underspecified. This section undertakes a systematic identification of the conceptual features that a system must possess for the metaphor to maintain philosophical coherence. While not exhaustive, this analysis isolates key characteristics implicit in contemporary discourse that could constitute an analytically rigorous account of what constitutes a legitimate digital twin.

Before doing so, we need to clarify a relevant point. The digital twin metaphor fundamentally presupposes a relation of "sameness" between the computational system and the individual it purports to replicate. However, philosophical analysis reveals that "sameness" admits of multiple interpretations, each with distinct ontological implications. In this context, the classical distinction between numerical and qualitative identity has important implications for understanding PDTs. For example, two children may have the same bicycle in one sense, and the same mother in another. Numerical identity constitutes the relation an entity bears exclusively to itself—a relation that admits no degrees and cannot be shared. Qualitative identity, conversely, denotes exact similarity in properties—a relation that admits of degrees and can obtain between numerically distinct entities. Parfit's (1984) canonical example illuminates this distinction: two white billiard balls may achieve qualitative identity through shared properties while remaining numerically distinct. When one ball is subsequently painted red, qualitative identity dissolves while the painted ball maintains numerical identity with its earlier state.

This distinction proves crucial also for analysing the digital twin metaphor. We hold that the metaphor necessarily implies qualitative rather than numerical identity between person P and their purported digital twin. The semantic structure of "twin" presupposes two numerically distinct entities exhibiting high degrees of qualitative similarity. This clarification exposes a fundamental conceptual confusion in recent literature. Iglesias et al. (2024) ask whether a digital twin might constitute "an extension of the self or of personal identity"—a formulation that appears to conflate numerical and qualitative identity. If proponents wish to investigate numerical identity between person P and their digital representation, they must abandon the twin metaphor entirely in favour of alternative frameworks such as "digital self" or "digital continuation" that do not presuppose numerical distinctness.

Furthermore, the relation of qualitative identity need not be perfect for the metaphor to maintain substantive coherence. Monozygotic twins—our biological paradigm—exhibit substantial phenotypic and psychological variation despite genetic identity. Accordingly, a charitable interpretation of the metaphor should invoke not exact qualitative identity but rather what we term "strong similarity"—a concept that, while admittedly vague, provides sufficient analytical precision for ethical discourse where many central concepts similarly resist precise definition.

Having clarified that the metaphor implies strong qualitative similarity rather than numerical identity, we now specify the dimensions along which such similarity must be assessed. To do so, we identify three interpretations of the

metaphor that have some plausibility, namely: the *behaviourist interpretation*, the *representational interpretation*, and the *phenomenal interpretation*. In this way, we do not contend that there is a single plausible construal of the PDT metaphor, but rather at least three, each of which appeals to different combinations of distinct features that may be considered necessary and sufficient to define a system as a PDT.

4.1 The behaviourist interpretation

The first and perhaps most basic intuitive interpretation of the metaphor is the "behaviourist interpretation", according to which all that matters for the metaphor to be viable is the generation, by the PDT, of a plausible output resembling that of person P. On this interpretation, strong similarity in outputs (the capacity to generate responses closely approximating those the modelled individual would produce) is both necessary and sufficient. This condition requires that person P, or informed third parties, can recognise in the system's outputs both stylistic and substantive features plausibly attributable to P.

Importantly, meeting strong similarity in outputs does not require satisfying a maximally demanding counterfactual test that would demand exact replication of P's responses in specific contexts. Human responses exhibit substantial variability contingent on factors ranging from recent conversations to physiological states. Requiring digital twins to achieve such precise counterfactual accuracy would render the metaphor vacuous. Moreover, verifying such accuracy would be methodologically impossible in most cases.

Therefore, strong similarity in outputs requires only that outputs be plausibly recognisable as belonging to P on the basis of demonstrated stylistic and content-based correspondence with P's documented expressions. Ideally, strong similarity would involve a form of self-recognition, such that when person P examines the outputs of their digital twin, they judge them as sufficiently similar, something they plausibly could have said and would be willing to endorse in that context. However, this criterion can be unreliable, since it cannot be applied when the person is no longer alive (as in the case of Griefbots) and may vary across individuals with different thresholds for what they consider plausibly 'theirs', which is why we also propose a method that does not depend on the person's own assessment. What we have in mind here is a variation on the Turing test, originally designed to assess whether a machine's conversational ability is indistinguishable from a human's. This could be adapted to assess a digital twin: human evaluators familiar with person P would attempt to determine whether a given message originated from the real P or from their LLM "twin". If evaluators cannot reliably distinguish between the two, the system could be said to achieve strong similarity for that context. Something very similar has already been operationalized for a PDT of philosopher Daniel Dennett. In that study, experts on Dennett's work and readers of his philosophy blog were asked to distinguish between texts produced by Dennett himself and texts produced by his PDT. The researchers posed ten philosophical questions to the real Dennett and then asked the PDT the same questions, collecting four unedited responses from the model for each prompt (Schwitzgebel et al. 2024). As we will discuss in greater detail in Sect. 5, the results of this study showed that evaluators were largely able to distinguish the real Dennett from his digital twin, indicating the difficulty of achieving strong similarity.

4.2 The representational interpretation

A stronger interpretation of what it means for a system to qualify as a PDT requires more than mere output similarity. While producing responses consistent with what a person might say is a necessary condition, it is far from sufficient. To merit the metaphor of a "twin", a PDT must also share access to the informational substrate that underlies that person's identity and must exhibit *unified agency*, the capacity to act coherently as a single agent across diverse contexts. This entails maintaining a reasonable degree of consistency of beliefs, values, and decision-making principles across professional, personal, and social domains. We refer to this as the representational interpretation of PDT.

To illustrate, imagine a large language model trained exclusively on Derek Parfit's philosophical writings. Such a model might generate new philosophical arguments in Parfit's style, yet it would fail to capture his passion for photography, his personal relationships, or the broader life experiences that shaped his thought. More importantly, the deficiency here is not merely instrumental. Its lack of access to such information would not only limit its expressive accuracy but would also exclude it from being a *representation* of Parfit at all. To be Parfit in any meaningful sense is to have access to the informational world that Parfit inhabited, including the facts, memories, and experiences that structured his cognition and perspective. A system deprived of that informational grounding could imitate his writings but not instantiate his personhood. Even if it performed isolated tasks that resembled aspects of Parfit's intellectual output, it would lack the continuity and coherence necessary to be recognised as a single, unified agent across domains. For a certain unified agency to be exhibited, it is important—beyond facts, memories, and experiences—that also certain patterns of thinking and reacting be observable across domains, ones that, generally speaking, belong to Parfit alone, marking his character and personality.[1]

[1] In this context, we do not refer to identity theories such as psychological continuity or narrative identity. Psychological continuity is a theory that evaluates a person's persistence over time through the continuity of their mental states and memories, beliefs, desires, and

Under the representational interpretation, a genuine digital twin must therefore exhibit a sufficient degree of *cross-domain coherence*. It should be capable of navigating professional, personal, and social contexts according to consistent underlying patterns, rather than fragmenting into task-specific functions. This view contrasts with the distinction proposed by Voinea and colleagues (2025), who differentiate between *task-specific* digital twins, designed for specialized functions such as academic writing or medical decision support, and *relational* ones, intended to emulate interpersonal roles. Their framework assumes that unified agency across domains is unattainable. We argue that such an assumption undermines the metaphor's representational power by reducing the person to a collection of context-bound outputs rather than a coherent, continuous self.

4.3 Phenomenal interpretation

The most demanding interpretation requires not only strong similarity in output and unified agency, but also in the cognitive and emotional processes that generate outputs or, at least, emerge in the outcome production. We refer to this interpretation as the *phenomenal interpretation.* This implies that digital twins must instantiate some form of consciousness, experiencing interactions with users from a first-person perspective similar to the modelled individual's subjective experience.

A clarification is needed. Following Ned Block (1995), we use "consciousness" to mean phenomenal consciousness, not access consciousness. Access consciousness concerns information processing, while phenomenal consciousness concerns lived, qualitatively subjective experience. Second, several theories of consciousness have been proposed, which attribute to phenomenal consciousness a different role in producing output. Some authors believe that phenomenal consciousness may have a causal role in generating outputs, since it represents the capacity of a system of integrating information (Tononi 2012). Others believe, instead, that consciousness does not play any causal role in generating output even in human beings since it would be epiphenomenal (Robinson 2010). Then a system might have a similar functional causal structure without having similar conscious properties. We do not need to take a stance on this complex debate; what is important to claim here is that there is at least one plausible interpretation of the metaphor of PDT that rests on the conviction that the first-person experience is something relevant to maintain that a system can be considered a PDT of a person P.

Returning to the previous example, what matters here is not only that Digital Parfit answers as Parfit would answer, or that it possesses a unified system capable of expressing coherence across several tasks and domains, but that the PDT *experiences* a way similar to how Parfit himself experienced performing his daily tasks. In this sense, the term "digital twin" is understood in a comprehensive way, encompassing not only the elements previously discussed but also an internal phenomenal dimension, one that presupposes some form of intentional consciousness on the part of the digital twin. That is, the PDT would experience interactions with the user from a first-person perspective, in a manner that resembles the way the person it is modelled on would experience them.

Interpretation of the PDT metaphor (from least to most demanding)	Corresponding dimension of qualitative similarity
Behaviourist	Strong similarity in output the PDT generates sufficiently similar outputs to what person P would have done in certain contexts
Representational	Strong similarity in unified agency (in addition to meeting the requirements of behaviourist interpretation) the PDT should be a single system that has access to the comprehensive informational substrate underlying personal identity of person P, and reflects the person's ability to act across multiple contexts
Phenomenal	Strong similarity in underlying processes (in addition to meeting the requirements of behaviourist and representational interpretations) the PDT should instantiate some form of phenomenal consciousness, experiencing interactions from a first-person perspective analogous to P''s subjective experience

Footnote 1 (continued)

personality traits (Parfit 1984). While those aspects play an important role in enabling users to recognise a system as the PDT of person P, we do not invoke them here as criteria of personal persistence or psychological continuity, but rather as part of the informational background that supports recognisability of the PDT by users as a unified agent. Along similar lines, we avoid appealing to narrative identity, since it involves self-conceptions (DeGrazia 2005), that is, a first-person standpoint of the system, which is not required under the representational interpretation of the digital twin metaphor. What matters, instead, is that the PDT exhibits patterns of behaviour, memories, and related features that can adequately represent person P.

4.4 Three diagnostic frameworks to assess the digital twin metaphor appropriateness

As shown, the three interpretations can be understood as exhibiting increasing demandingness: the behaviourist interpretation requires only output similarity; the representational interpretation adds requirements for comprehensive training data and unified agency; and the phenomenal interpretation further demands similarity in underlying cognitive and emotional processes.

These interpretations function not as strict definitions but as analytical tools, offering structured frameworks for evaluating whether, and to what extent, the digital twin metaphor applies to particular systems. Importantly, our subsequent analysis does not employ these interpretations as a definitive checklist but as a diagnostic framework. A system's failure to meet even the minimal behaviourist interpretation would definitively disqualify it from legitimate use of the twin metaphor; partial satisfaction might warrant qualified or restricted use; and full satisfaction across all dimensions would justify unrestricted application. This approach preserves analytical rigor while accommodating the inherent vagueness of metaphorical language.

5 How the metaphor fails: comparing the digital twin metaphor with current personalised LLMs

Having established a diagnostic framework for evaluating the digital twin metaphor, we now undertake a systematic assessment of whether current LLM-based implementations satisfy these criteria. This section offers a deliberately high-level and conceptual assessment rather than an evaluation of any specific commercial system. We take this approach because LLM-based architectures evolve quickly, and focusing on a single platform would risk tying our arguments to contingent design choices. Our aim is, therefore, to articulate criteria that apply broadly across current and near-future systems, independently of particular implementations**.** Our evaluation employs the previously defined interpretative taxonomy not as a binary checklist but as a graduated analytical tool, examining the degree to which current systems approximate each dimension of similarity. The analysis reveals that even under the most charitable interpretation, current implementations fail to justify the twin designation.

LLMs work by learning complex statistical patterns in massive amounts of text to predict and generate human-like language. At a basic level, an LLM learns to predict what piece of text—called a *token*, which may be a word or part of a word—should come next based on the text that came before it. During training, the model reads sequences of text and adjusts its internal parameters so that its predicted next tokens get closer to the real ones. It does this repeatedly until it becomes good at anticipating what should come next. At inference time, the model takes an input prompt, runs it through its transformer layers, and then generates the next tokens one by one by sampling from the probabilities it has learned. Once trained, an LLM can be turned into a conversational agent by providing a system instruction or prompt that helps enforce turn-taking behaviour with the system and user, and defines its behaviour, tone, and goals (for example, instructing it to act as a helpful assistant, tutor, or advisor). Creating a personalized digital twin with an LLM typically begins with a general-purpose model, which can then be adapted using personal data. One way to do this is Fine-Tuning (FT), which means training the model further on a person's own text, such as messages, emails, essays, or transcripts. This process lets the model learn that individual's tone, style, and preferences (Zhang et al. 2024; Liu et al. 2025). FT could be combined with Retrieval-Augmented Generation (RAG), a method where the model looks up relevant information from an external collection of a person's documents or data whenever it generates a response (Wang et al. 2024). RAG ensures that the model can ground its responses in the data available about the person (Klesel 2025), while FT helps the model learn their tone, style, and preferences.

With this description of how digital twins can be implemented using current LLMs' technologies, we can now examine whether they live up to the promise implied by the digital twin metaphor. Specifically, let's evaluate their resemblance to the individual they imitate across the three interpretations of the metaphor identified in the last section.

The first (behaviourist) interpretation of the digital twin metaphor concerns similarity in output, that means having the model reliably produce responses that match what the person it emulates would say. Referring to the Turing-test standard for strong similarity proposed above, the limited existing empirical evidence suggests that this bar cannot currently be met even in the case of a narrowly focused PDT trained for imitating a person in specific tasks. Recalling the results of the experiment on Daniel Dennett's PDT mentioned earlier, where a model was fine-tuned on Dennett's philosophical writings, both experts on his work and philosophy-blog readers were reliably able to distinguish the model's responses from Dennett's own (Schwitzgebel et al. 2024). This confirms that the PDT did not pass a domain-specific Turing test. We believe that there are three main limitations to achieving strong similarity of output, two technical and one conceptual. First, lack of relevant and complete training data. Personal datasets used in FT are frequently sparse or fragmented (more on this later), leaving the model without enough context to generate accurate responses in some circumstances. In such cases, the model may either fail to respond meaningfully or fall back on the

generic knowledge encoded in the base model, neither of which reflects what the person would actually say. For example, if a user asks a digital twin of Parfit, "Where is Parfit's favourite place to take photographs?" and that information is not in (or inferable from) the personalized training data, the model would either admit that it does not know or guess a plausible answer based on what is present in the original base model. Therefore, missing data will inevitably lead to divergence between the model's output and the real individuals for certain interactions. Second, even when the data is available in the personalized training set, LLMs are prone to hallucination (confidently producing false or misleading outputs that do not align with what is present in the training data) (Huang et al. 2025).The probabilistic nature of next-token prediction makes hallucinations a structural feature of LLMs, not just a bug that will be fixed (Xu et al. 2024). Third, there is a deeper conceptual issue: even if a model could produce accurate responses, which version of a person is it supposed to reflect? Human beings change over time (their beliefs, preferences, and behaviours evolve) whereas an LLM is static once trained. This creates an unavoidable mismatch between the dynamic nature of real individuals and the frozen snapshot represented by their LLM-based digital twin.

The second representational interpretation of the digital twin metaphor concerns sufficiency in information access and unified agency. The idea is that a digital twin should have access to a body of data that resembles the knowledge, memories, and experiences of the individual it represents, and should exhibit a coherent set of beliefs, values, and competencies across the various tasks and contexts it engages in. Current implementations fall far short of this standard for two reasons.

The first limitation concerns the amount of data that a model needs to have access to satisfy the representational interpretation. While today's individuals leave behind far more digital traces than any previous generation, the data available to the model (either through fine-tuning or through an external database) still represents only a tiny fraction of a person's actual lived experience. As Vallor (2024, pp. 23–24) notes, human experience unfolds as a continuum, while data is discrete, fragmented, and selectively captured. So even if we could record every word spoken or written, we would still be missing vast amounts of meaningful data. Fine-tuning chat-based LLMs relies almost entirely on text, which captures only part of human communication. The tone of voice, facial expressions, gestures, and pauses that shape meaning in a conversation are lost in transcripts. Even more importantly, internal states (such as thoughts, emotions, intentions) are currently unrecorded in training data, yet they are central to how humans make decisions and express themselves. Even if we were able to record more information and create a more comprehensive dataset, the fundamental problem would remain unresolved. This is because the very act of capturing data can introduce distortion: people may alter their behaviour when they know they are being recorded, leading to a version of the self that is curated rather than authentic. Similarity in information access also implies a similarity in the causal relationship between data and output. It is not enough for a digital twin to produce statistically plausible or "correct" answers; the metaphor also suggests that the model produces the right outputs for the right reasons. For instance, a PDT of Parfit might predict that Parfit enjoyed taking photographs simply because many people in his demographic do. While this may be statistically accurate, it misses the specific, contextual causes that give the preference meaning. Perhaps he developed a love for photography after spending many trips photographing Venice. In this case, to achieve similarity in information substrate, it's not enough for the memory of trips to Venice to be accessible to Parfit's PDT; we also want that specific information to play a causal role in shaping the model's outputs, much like how such a memory would influence a real person's response.

The second limitation concerns the lack of a single system able to act consistently across all contexts, as it is required for unified agency. Current LLMs are limited by their narrow interaction modalities (primarily text-based) and lack the embodied, multimodal awareness that supports unified human agency. But even when we restrict our expectations to the text modality alone, serious limitations persist. LLMs often struggle to maintain consistency across different interactions across domains. Some suggest training models to simulate specific relationships, but doing so within a single unified model risks context leakage, where sensitive or role-specific information is inappropriately reused or generalized. As a result, the best we can currently achieve is to deploy separate instances or specialized configurations of a model for different tasks and then build an orchestrating structure around them to give the appearance of a single coherent agent.

The last interpretation of the digital twin metaphor concerns similarity in the processes through which the digital twin generates its outputs. While we have already discussed the need for a causal relationship between training data and output, this goes a step further: it implies a kind of psychological similarity between the individual and their digital counterpart. Unlike surface-level similarities in training data or output, this criterion requires that the digital twin replicate, at least in some form, the cognitive and emotional mechanisms a human would engage with when producing a response. Popular culture, particularly science fiction, has helped entrench this expectation, often portraying digital twins as fully conscious replicas capable of reasoning, feeling, and intentional communication. However, this image diverges sharply from technological reality. LLMs' output

holds no communicative intentions (Magnus 2025). They are almost certainly not phenomenally conscious, and it is debated to what extent they hold mental states. LLMs' data-driven prediction fundamentally differs from human cognition's theory-based, causal reasoning (Felin and Holweg 2024). We can therefore safely assume that current personalized LLMs are not capable of simulating inner mental states of the individual they are trying to emulate. In this sense, their behaviour is more akin to role-playing or acting than thinking: they can mimic the appearance of cognition, much like an actor playing a character on stage, but without actually being the character (Shanahan et al. 2023).

This systematic evaluation demonstrates that current LLM-based systems fail to satisfy even minimal criteria for legitimate application of the digital twin metaphor. They achieve at best weak statistical correspondence in limited domains while failing entirely on dimensions of data sufficiency, unified agency, and phenomenal similarity. When viewed through the three interpretations of the metaphor outlined in Sect. 4, the conclusion is clearer: even under the behaviourist interpretation of the metaphor, output similarity remains too inconsistent and fragmented to reliably satisfy this standard; under the representational interpretation, lack of unified agency disqualifies stronger claims to twinning; and the phenomenal interpretation is not met at all, given the absence of replication of the subject's cognitive processes. Thus, across all three frameworks, current implementations fail to justify the digital twin metaphor.

6 The ethical implications of metaphorical misrepresentation

Having established the inadequacy of the digital twin metaphor for current LLM systems, we now examine why this misrepresentation matters ethically. The analysis reveals that inappropriate metaphorical framing generates both individual harms in high-stakes contexts and collective distortions in public understanding and policy formation.

Metaphors shape not merely description but comprehension, particularly for non-expert users navigating complex technologies. When metaphors systematically misrepresent system capabilities, they create conditions for what Miranda Fricker (2007) terms "hermeneutical injustice"—the absence of adequate interpretive resources for understanding one's experiences. In the context of AI systems, misleading metaphors deprive users of the conceptual tools necessary to accurately comprehend their interactions with technology. This harm distributes unevenly across populations. Research demonstrates significant overlap between digitally marginalized communities and those with limited technical literacy (van Deursen and van Dijk 2019). These populations—already vulnerable to digital exclusion—face heightened susceptibility to metaphorical misrepresentation. When the "digital twin" metaphor suggests capabilities that systems cannot deliver, it could perpetrate a form of epistemic injustice that compounds existing inequalities.

The anthropomorphic nature of the twin metaphor amplifies these risks. Anthropomorphic framing encourages systematic attribution of human-like properties—intentions, emotions, consciousness—to systems that possess none of these characteristics. While experts may recognize such attributions as metaphorical convenience, lay users often interpret them literally, forming false beliefs about system capabilities that shape their interactions and decisions. False beliefs about AI's capabilities (particularly those stemming from anthropomorphic representations) can lead to several harms. These include diminished user autonomy, unwarranted trust in the system, and an increased likelihood of over-disclosing personal or sensitive information (Akbulut et al. 2024; Gabriel et al. 2024; Marchegiani 2025).

Beyond individual comprehension, metaphors function as frameworks structuring collective understanding and action. The Italian regulation of cultivated meat provides an instructive parallel (Fino et al. 2024). Opponents successfully reframed "cultivated meat" as "synthetic meat", triggering associations with artificiality and danger that culminated in prohibitive legislation. This case demonstrates how metaphorical framing can determine regulatory outcomes independent of empirical evidence. The digital twin metaphor risks analogous distortions in AI governance. By suggesting that LLMs can genuinely replicate individual identity, the metaphor shapes policy discussions around questions of digital personhood, posthumous rights, and consent for computational simulation. These discussions proceed from false premises about technological capabilities, potentially establishing regulatory frameworks that either inadequately protect individuals or unnecessarily restrict beneficial applications.

Experimental evidence from Thibodeau and Boroditsky (2009) demonstrates that metaphorical framing systematically influences policy preferences. Participants presented with crime described as a "virus" favoured social reform interventions; those encountering crime as a "predator" preferred punitive enforcement. Similarly, framing LLMs as "digital twins" rather than "statistical simulators" may shape public and policymaker responses to these technologies along certain directions instead of others.

The concrete manifestation of metaphorical harm emerges most acutely in high-stakes applications where vulnerable individuals encounter these systems during moments of profound need. In healthcare contexts, the proposal for "personalized patient preference predictors" (Earp et al. 2024) exemplifies how metaphorical framing shapes system design and deployment. Framing these as "psychological twins" encourages developers to prioritize superficial

resemblance—replicating patient mannerisms, speech patterns, even visual appearance—over functional utility. Surrogate decision-makers confronting systems that mimic their loved ones face impossible psychological burdens, potentially deferring to algorithmic recommendations even when possessing relevant information about actual patient preferences. The metaphor transforms a statistical tool into something different, potentially enhancing vulnerability precisely when clarity is most essential (but see Earp et al. 2025).

Legal proceedings face parallel challenges when proposals emerge for digital twins providing posthumous testimony. The twin metaphor suggests these systems can "speak for" deceased individuals rather than merely analyzing data patterns. This conflation between statistical inference and authentic testimony could undermine adversarial legal processes that depend on accountable human witnesses subject to cross-examination.

Commercial "griefbots" represent perhaps the most ethically troubling application. Marketing these systems as preserving the "essence" of deceased individuals exploits profound human vulnerability. Users may form dependencies that prevent healthy grief processing, experiencing what has been called "second loss" when services discontinue or algorithms update (Lindemann 2022). The twin metaphor transforms the necessary psychological work of accepting loss into technologically mediated denial.

The most profound harm may be what we term "phenomenal deception"—the suggestion that these systems possess or can simulate subjective experience. When users interact with a "digital twin," the metaphor implies engagement with something possessing interiority, consciousness, intentionality. This represents not merely overstatement but conceptual error, attributing phenomenal properties to systems operating through fundamentally different mechanisms. This deception proves particularly pernicious because it exploits fundamental human cognitive tendencies toward anthropomorphism and social projection. Humans naturally attribute mental states to entities exhibiting behavioural complexity—a tendency that served evolutionary purposes but proves maladaptive when engaging with sophisticated pattern-matching systems. The twin metaphor amplifies rather than corrects this tendency, encouraging users to perceive consciousness where none exists.

7 Toward adequate conceptual frameworks

The failure of the digital twin metaphor necessitates alternative conceptual frameworks that accurately represent the capabilities and limitations of personalized LLMs while maintaining practical utility for users, developers, and policymakers.

A notable recent attempt to address the inadequacy of the digital twin metaphor comes from Voenia et al. (2025), who critique it as failing to capture "the distinctive features of these models: […] their simulation of aspects of a specific individual's psychology". They propose instead the nomenclature "AI Simulated Mind" (AI SIM), which they argue more accurately represents how these systems are "designed to simulate that specific person, not a closely related but distinct individual". While this represents an important recognition of the metaphor's limitations, their analysis remains insufficiently specified, particularly regarding which psychological "aspects" undergo simulation. Moreover, their critique proceeds from opposite premises to ours: whereas they contend the digital twin metaphor understates psychological similarity, we have demonstrated that it systematically overstates the degree of psychological correspondence achievable through current LLM architectures. The AI SIM framework risks suggesting a level of psychological fidelity that exceeds technical capabilities.

Shanahan et al. (2023) propose understanding LLMs through "role-playing" rather than identity replication. This framework accurately captures how these systems generate performances based on learned patterns without implying genuine identity or consciousness. A "role-playing system" trained on individual data performs for that individual much as an actor performs a character, through surface-level mimicry rather than deep identification. This metaphor offers several advantages: it maintains clear boundaries between performance and identity, acknowledges the constructed nature of outputs, and avoids suggesting phenomenal properties. Users understanding that they are interacting with a sophisticated role-playing system maintain appropriate emotional distance and critical evaluation.

A more technical but precise framework is to describe these systems as "linguistic pattern engines"—mechanisms that identify and reproduce statistical regularities in language use. This characterization foregrounds the mechanical nature of their operation while avoiding misleading biological or psychological analogies. This framework proves particularly valuable in professional contexts where precision matters more than accessibility. Healthcare providers, legal professionals, and policymakers benefit from understanding these as pattern-matching tools rather than identity-preserving technologies.

The aim of this paper is not to support a precise alternative to the digital-twin metaphor, but rather to highlight structural weaknesses of the latter. The alternatives we have mentioned are promising, yet they require further reflection before they can be validated from a conceptual standpoint. Moreover, in this respect, it would be valuable to examine, from an empirical perspective, the users' attitudes elicited by different labels. Such evidence could, in turn, inform the

conceptual analysis and help develop a more accurate label for personalized LLMs.

8 Conclusion

The "digital twin" metaphor, when applied to personalized LLMs, constitutes not merely semantic imprecision but systematic misrepresentation with profound ethical implications. Our analysis demonstrates categorical failure across all dimensions that would justify the twin designation: current systems produce statistically plausible but not authentic outputs (Battisti 2025), lack access to the comprehensive experiential data constituting human identity, cannot maintain unified agency across contexts, and operate through pattern-matching rather than cognitive processes. This metaphorical misrepresentation generates concrete harms from both collective and individual stand points.

The path forward requires conceptual frameworks that maintain fidelity to technological reality. Understanding LLMs as role-playing systems, statistical simulacra, or linguistic pattern engines preserves clarity about their actual capabilities—sophisticated pattern matching—while avoiding the suggestion of preserved identity or consciousness. These frameworks may lack the seductive appeal of "digital twin", but they serve the interests of users, developers, and society by maintaining clear boundaries between performance and personhood, between mimicry and identity, between statistical inference and human experience.

As these systems spread across different domains of human experience—from healthcare to law, and even to the most intimate personal relationships—the stakes of metaphorical precision increase. Although these tools have demonstrated their usefulness, there are reasonable doubts about our ability to understand their nature. The genuine benefits these technologies can offer emerge only when we abandon the pretence that they can—or should—preserve or replicate human identity.

The analysis presented here represents more than an academic exercise in conceptual precision. It constitutes an urgent ethical imperative as these systems shape increasingly consequential decisions about human life and death, memory and legacy, identity and consciousness. We must resist the allure of metaphors that promise digital transcendence of mortality through computational simulation. These are not twins preserving human essence but sophisticated pattern-matching systems generating statistical approximations of human discourse. Recognizing this distinction—maintaining vigilance against metaphorical seduction—represents a crucial task for preserving conceptual integrity in the coming age of artificial intelligence.

The broader lesson extends beyond any single metaphor: the language we choose to describe emerging technologies shapes not merely understanding but reality, determining how these systems are designed, deployed, and governed. In choosing our metaphors, we choose our futures. The responsibility to choose wisely—to prioritize accuracy over appeal, clarity over comfort—falls to researchers, developers, policymakers, and society collectively. The stakes—for autonomy, and the very meaning of human identity—demand nothing less.

Acknowledgements The authors acknowledge the use of ChatGPT (version 5, OpenAI) and Opus 4.1 and 4.5 (Claude AI, Anthropic) as AI-assisted language editing platforms to improve the final text in its clarity, coherence, and consistency. This editing process was undertaken without introducing substantive changes to the scholarly content, arguments, or conclusions. All authors retain full responsibility for the accuracy, interpretation, and scholarly integrity of the final manuscript. The authors would like to thank Dr Gary D. O'Brien for his helpful comments and feedback on an earlier draft of this paper.

Author contribution MA, DB, and BM contributed equally to all aspects of this work, including conceptualization, methodology development, original draft preparation, and critical revision of the manuscript. All authors have reviewed and approved the final version of the manuscript for submission.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ahmedien DAM (2023) Analysing bio-art's epistemic landscape: from metaphoric to post-metaphoric structure. BioSocieties 18:308–334. https://doi.org/10.1057/s41292-022-00270-y

Akbulut C, Weidinger L, Manzini A, Gabriel I, Rieser V (2024) All too human? Mapping and mitigating the risk from anthropomorphic AI. Proc AAAI/ACM Conf AI Ethics Soc 7(1):13–26. https://doi.org/10.1609/aies.v7i1.31613

Battisti D (2025) The authenticity requirement: why using digital twins for achieving person-span extension goods can be self-defeating. Am J Bioeth 25(2):120–123. https://doi.org/10.1080/15265161.2024.2441761

Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18(2):227–247
Boroditsky L, McClelland J, Thibodeau P (2009) When a bad metaphor may not be a victimless crime: the role of metaphor in social policy. In Proceedings of the Annual Meeting of the Cognitive Science Society. 31(31)
Braun M (2021) Represent me: please! Towards an ethics of digital twins in medicine. J Med Ethics 47:394–400. https://doi.org/10.1136/medethics-2020-106134
Danaher J, Nyholm S (2024) Digital duplicates and the scarcity problem: might AI make us less scarce and therefore less valuable? Philos Technol 37:106. https://doi.org/10.1007/s13347-024-00795-z
Danaher J, Nyholm S (2025) The ethics of personalised digital duplicates: a minimally viable permissibility principle. AI Ethics 5:1703–1718. https://doi.org/10.1007/s43681-024-00513-7
Earp BD, Porsdam Mann S, Allen J, Salloch S, Suren V, Jongsma K, Braun M, Wilkinson D, Sinnott-Armstrong W, Rid A, Wendler D, Savulescu J (2024) A personalized patient preference predictor for substituted judgments in healthcare: technically feasible and ethically desirable. Am J Bioeth 16:1–14. https://doi.org/10.1080/15265161.2023.2296402
Earp BD, Mann SP, van Veenendaal T, Allen J, Salloch S, Jongsma K, Savulescu J (2025) The enduring promise of personalising patient preference prediction: responding to critics of the 'P4' proposal. preprint
Fawkes A, Burden D (2025) Digital human twins and the military metaverse: opportunities and challenges. AI Soc. https://doi.org/10.1007/s00146-025-02508-2
Felin T, Holweg M (2024) Theory is all you need: AI, human cognition, and causal reasoning. Strat Sci 9(4):346–371. https://doi.org/10.1287/stsc.2024.0189
Fino MA, Anzà B, Bairati L, Bertini I, Biolatti B, Biressi S, Cannizzo FT, Cavallarin L, Conti L, Deriu M, Gargioli C, Loera B, Lo Sapio L, Marchisio D, Pallante L, Stano S, Torri L, Bertero A, Massai D (2024) Cultivated meat beyond bans: ten remarks from the Italian case toward a reasoned decision-making process. One Earth 7(12):2108–2111. https://doi.org/10.1016/j.oneear.2024.11.002
Floridi L, Nobre AC (2024) Anthropomorphising machines and computerising minds: the crosswiring of languages between artificial intelligence and brain & cognitive sciences. Mind Mach 34:5. https://doi.org/10.1007/s11023-024-09670-4
Fricker M (2007) Epistemic injustice: power and the ethics of knowing, https://doi.org/10.1093/acprof:oso/9780198237907.001.0001, accessed 9 Sept. 2025.
Gabriel I, Manzini A, Keeling G, Hendricks LA, Rieser V, Iqbal H, Manyika J (2024) The ethics of advanced ai assistants. arXiv preprint arXiv:2404.16244.
Gelernter D (1991) Mirror worlds: or: the day software puts the universe in a shoebox...How it will happen and what it will mean https://doi.org/10.1093/oso/9780195068122.003.0008
Giubilini A, Porsdam Mann S, Voinea C et al (2024) Know thyself, improve thyself: personalized LLMs for self-knowledge and moral enhancement. Sci Eng Ethics 30:54. https://doi.org/10.1007/s11948-024-00518-9
Grieves M, Vickers J (2017) Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen J, Flumerfelt S, Alves A (eds) Transdisciplinary perspectives on complex systems. Springer, Cham. https://doi.org/10.1007/978-3-319-38756-7_4
Hollanek T, Nowaczyk-Basińska K (2024) Griefbots, deadbots, postmortem avatars: on responsible applications of generative AI in the digital afterlife industry. Philos Technol 37:63. https://doi.org/10.1007/s13347-024-00744-w
Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T (2025) A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. https://doi.org/10.1145/3703155
Iglesias S, Earp BD, Voinea C, Mann SP, Zahiu A, Jecker NS, Savulescu J (2024) Digital doppelgängers and lifespan extension: what matters? Am J Bioeth 25(2):95–110. https://doi.org/10.1080/15265161.2024.2416133
Jimenez JI, Jahankhani H, Kendzierskyj S (2020) Digital twin technologies and smart cities. Health care in the cyberspace: medical cyber-physical system and digital twin challenges. Springer, Switzerland, pp 79–92. https://doi.org/10.1007/978-3-030-18732-3
Klesel M, Wittmann HF (2025) Retrieval-Augmented generation (RAG). Bus Inf Syst Eng. https://doi.org/10.1007/s12599-025-00945-3
Lakoff G, Johnson M (1980) Metaphors we live by. University of Chicago Press, Chicago. https://doi.org/10.2307/430414
Lindemann NF (2022) The ethics of 'deathbots.' Sci Eng Ethics 28(6):60
Liu J, Qiu Z, Li Z, Dai Q, Zhu J, Hu M, King I (2025) A survey of personalized large language models: progress and future directions. arXiv preprint arXiv:2502.11528
Magnus PD (2025) On trusting chatbots. Episteme. https://doi.org/10.1017/epi.2024.29
Marchegiani B (2025) Anthropomorphism, false beliefs, and conversational AIs: how chatbots undermine users' autonomy. J Appl Philos. https://doi.org/10.1111/japp.70008
Morse T (2023) Digital necromancy: users' perceptions of digital afterlife and posthumous communication technologies. Inf Commun Soc 27(2):240–256. https://doi.org/10.1080/1369118X.2023.2205467
Parfit D (1984) Reasons and persons. Oxford University Press, London, England
Pigliucci M, Boudry M (2011) Why machine-information metaphors are bad for science and science education. Sci Educ 20:453–471. https://doi.org/10.1007/s11191-010-9267-6
Robinson WS (2010) Epiphenomenalism. Wiley Interdisciplinary Reviews: Cognitive Science 1(4):539–547
Salles A, Evers K, Farisco M (2020) Anthropomorphism in AI. AJOB Neurosci 11(2):88–95. https://doi.org/10.1080/21507740.2020.1740350
Schwitzgebel E, Schwitzgebel D, Strasser A (2024) Creating a large language model of a philosopher. Mind Lang 39(2):237–259. https://doi.org/10.1111/mila.12466
Shanahan M, McDonell K, Reynolds L (2023) Role play with large language models. Nature 623:493–498. https://doi.org/10.1038/s41586-023-06647-8
Taylor C, Dewsbury BM (2018) On the problem and promise of metaphor use in science and science communication. J Microbiol Biol Educ 19:10.1128/jmbe.v19i1.1538. https://doi.org/10.1128/jmbe.v19i1.1538
Thibodeau PH, Boroditsky L (2011) Metaphors we think with: the role of metaphor in reasoning. PLoS ONE 6(2):e16782. https://doi.org/10.1371/journal.pone.0016782
Tononi G (2012) Integrated information theory of consciousness: an updated account. Archives italiennes de biologie 150(4):293–329
Vallor S (2024) The AI mirror: how to reclaim our humanity in an age of machine thinking. Oxford University Press, Oxford
van Deursen AJ, van Dijk JA (2019) The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 21(2):354–375. https://doi.org/10.1177/1461444818797082
Voinea C, Porsdam Mann S, Earp BD (2025) Digital twins or AI SIMs? What to call generative AI systems designed to emulate specific

individuals, in healthcare settings and beyond. J Med Ethics, in press

Wang Z, Li Z, Jiang Z, Tu D, Shi W (2024) Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4891–4906). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.281

Xie H, Chen Y, Xing X, Lin J, Xu X (2025) PsyDT: using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1081–1115, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.55

Xu Z, Jain S, Kankanhalli M (2024) Hallucination is inevitable: an innate limitation of large language models. arXiv preprint arXiv:2401.11817.

Zhang Z, Rossi RA, Kveton B, Shao Y, Yang D, Zamani H, Wang Y (2024) Personalization of large language models: a survey. arXiv preprint arXiv:2411.00027.