Clowes, R. W. & Smart, P. R. (in press) Mindshaping Large Language Models. In V. Santos & P. Castro (Eds.), *Advances in Philosophy of Artificial Intelligence*. Ethics Press, Bradford, UK.

Mindshaping Large Language Models

Robert W. Clowes and Paul Smart

Abstract

This chapter revisits the notion of mindshaping in light of recent developments in artificial intelligence, with a particular focus on large language models (LLMs). Following Zawidzki, mindshaping is understood as a set of social and developmental practices that align agents with shared models of agency, rendering their behaviour intelligible within a folk-psychological framework. We extend this account by introducing *artificial mindshaping*: the suite of training procedures, interactional dynamics, and design constraints that shape the behavioural profiles of LLMs so that their actions can be interpreted in terms of familiar mentalistic categories, such as belief, desire, intention, and reason. Drawing on Shanahan, McDonell, and Reynolds (2023), we organise these mechanisms along a distinction between the simulator (the underlying language model and its trained dispositions) and the simulacrum or virtual personality (the persona the simulator instantiates in a given interaction). We characterise the property these mechanisms produce as *interpretive alignment*—the extent to which an artificial system's behaviour can be subsumed under folk-psychological kinds, independently of whether its internal mechanisms mirror those categories. While interpretive alignment connects naturally with contemporary work in explainable AI, it also bears on issues of mentality. Specifically, interpretive alignment establishes the conditions for mindedness as per Dennett's intentional stance. In this sense, artificial mindshaping is a form of mindmaking—not merely a way of making minds explainable but also a means by which artificial forms of mindedness are brought into being. We close by noting that mindshaping is a relational notion that runs in both directions, and that the question of how human minds are in turn shaped by these artificial systems is becoming one of the most pressing raised by their growing integration into human cognitive life.

Introduction

Large language models (LLMs) are no longer confined to the background of digital life. Embedded in an expanded range of systems from chatbots, to assistants, to virtual personalities they increasingly occupy roles that are recognisably social and normative:

offering advice, providing explanations, assisting with decisions, and participating in extended exchanges that resemble ordinary conversation.

How are we to make sense of these systems? One possibility is that we attempt to understand LLMs in largely the same way that we understand ourselves and others. That is to say, we resort to the familiar folk psychological strategy of explaining behaviour via the ascription of beliefs, desires, intentions, and related mental states. This approach is certainly tempting, since the capabilities of LLMs—such as their linguistic fluency, responsiveness to context, and apparent sensitivity to reasons—strongly invite interpretation in folk psychological terms (Frankish 2024). At the same time, it is widely acknowledged that LLMs differ profoundly from human cognitive systems. Their internal organisation, training regimes, and modes of operation are unfamiliar, opaque, and only partially understood even by their designers (Schneider 2025). This tension—between behavioural familiarity and architectural strangeness—has led several commentators to suggest that LLMs exhibit an exotic or 'alien' form of intelligence (Frankish 2024; Lloyd 2024; Shanahan, in press).

This situation stands in marked contrast to our ordinary dealings with other human beings. We routinely credit ourselves and others with beliefs and desires, and we use these constructs to explain our own behaviour. The same is true when we attempt to explain the behaviour of others. When we see Jill rushing towards the bar, for example, we readily explain her behaviour by inferring that she desires a beverage. Much of human social cognition relies on our ability to make these sorts of inferences—to infer what must be happening inside another's head. This capacity is commonly referred to as mentalizing or *mindreading*: our ability to attribute mental states (such as beliefs, desires, intentions, and emotions) to oneself and others in order to explain, predict, and interpret behaviour.

One way of approaching this capacity is to assume that humans must be gifted with prodigious inferential powers—ones that enable us to discern the presence of hidden mental states from observable patterns of behaviour combined with contextual cues. In recent years, however, another approach has gained prominence, one that attributes the success of human mindreading (and thus the explanatory success of folk psychology) to the way in which human behaviour is moulded by social, cultural, and developmental practices. This approach is known as *mindshaping* (Zawidzki 2018). Mindshaping refers to

the processes by which human behaviour is shaped to conform to a folk psychological explanatory template. In short, the success of folk psychology is not due solely to our inferential abilities, but to the fact that we are encouraged to behave in ways that are predictable, intelligible, and normatively assessable. This reduces the inferential burden placed on the biological brain, since human behaviour is typically structured so that folk psychological explanation applies straightforwardly. The legibility of human minds is thus explained not solely by a capacity to read one another's minds, but also by the fact that our minds have been shaped to produce behaviour that conforms to a socially shared model of human agency. In this sense, human minds are readable by design.

The present chapter extends this idea of mindshaping to the realm of artificial systems, yielding what we call *artificial mindshaping*. Artificial mindshaping refers to a suite of practices aimed at shaping AI behaviour so that it better aligns with everyday folk psychological modes of interpretation. In this respect, artificial mindshaping constitutes a form of *explainability by design*. Instead of attempting to explain AI behaviour after the fact by probing internal mechanisms, artificial mindshaping seeks to shape behaviour in advance so that it naturally supports familiar forms of folk psychological understanding. If LLMs are indeed cognitively exotic, then we can either attempt to radically extend our mindreading capacities or engage in systematic efforts to shape the behaviour of LLMs themselves. The present chapter represents an initial effort to explore the latter route.

Mindshaping

The idea of mindshaping has its antecedents in Vitoria McGeer's (2001, 2007) work on the 'regulative dimension of folk psychology' and the term proper was introduced and developed by Tadeusz Zawidzki (2008, 2013, 2018) as a contrast and complement to the theory of mindreading. Whereas mindreading seeks to explain how humans interpret others' actions in terms of mental states, mindshaping aims to explain how such interpretation is made possible in the first place and how ongoing folk psychological interpretative practices tends to both shape and regulate minds.

In Zawidzki's formulation, mindshaping is a relation among three elements: a target mind, a shaping mechanism, and a model that the mechanism aims to impose. The target mind is the mind to be shaped, while the model is a culturally shared template of

agency—typically derived from folk psychology. This model provides the pattern to which individuals are moulded, allowing their actions, preferences, and patterns of reasoning to converge on culturally stabilized expectations.

According to Zawidzki, mindshaping operates through a variety of mechanisms. These include imitation, natural pedagogy, norm enforcement, and public language narratives:

- **Imitation.** A central mechanism of mindshaping is the propensity of human children to copy others' actions with striking fidelity, even when those actions are arbitrary, inefficient, or transparently causally irrelevant. As Tomasello (1999) emphasizes, human cultural learning is driven not merely by emulation—converging on the same outcome by whatever means—but by genuine imitation, in which the specific means deployed by a model are themselves treated as to-be-copied. Horner and Whiten's (2005) puzzle-box experiments vividly illustrate this distinction: whereas chimpanzees tend to emulate, skipping obviously unnecessary steps once the causal structure of the apparatus is revealed, human children systematically "overimitate," reproducing every element of the adult's demonstrated sequence, including actions that are visibly non-functional. Such overimitation ensures that children internalize not just what others do, but how they do it, leading to behavioural conformity even when causal understanding is limited.
- **Natural Pedagogy.** A second mechanism of mindshaping operates through what Csibra and Gergely (2009) call *natural pedagogy*: a suite of species-specific communicative practices—eye contact, infant-directed speech, exaggerated movements, the use of the child's name—that signal to infants that they are being taught something of general and normative significance. Human infants are exquisitely sensitive to these ostensive cues; when such signals are present, they treat the ensuing information not as a mere local fact but as culturally relevant, generalizable knowledge ("this is how one does X"). This pedagogical system is an evolved adaptation for transmitting opaque cultural knowledge—concepts, tool-use procedures, linguistic conventions—that children could not easily infer on their own. Zawidzki interprets this as a paradigmatic mindshaping device: through explicitly structured instruction, adults sculpt the conceptual and

behavioural repertoires of novices, installing shared beliefs, categories, and expectations that align the learner with the community's norms and interpretive practices.

- **Norm Enforcement.** Through practices of praise, blame, and sanction, societies discipline individuals' mental patterns. Children (and adults) are rewarded for acting in norm-congruent ways and punished (or at least corrected) for deviance. Over time, this "social policing" creates internalized expectations—one comes to want to do what is socially expected, effectively aligning one's desires with the group's norms. Norms thus mould motivations and values, again leading to more predictable, culturally typical behaviour.
- **Public Language Narratives.** Perhaps the most distinctive human mindshaping mechanism is our capacity to internalize and emulate fictional characters and narrative roles. Zawidzki (2018) emphasizes that through exposure to public language, stories, and cultural scripts, individuals come to model their behaviour and emotional life on idealized agents, such as heroes, parents, professionals, moral exemplars. These agents encode folk-psychological templates (e.g., rational action, emotional coherence, moral deliberation), and by emulating them, individuals shape their own minds to match socially-recognized models.

The originality of mindshaping lies in its reversal of explanatory direction. Rather than treating mindreading as the primary mechanism of social cognition, with mindshaping as a derivative or exceptional case, it is suggested that mindshaping prepares agents to be mind-readable. Without this preparatory work, it is suggested, mindreading would be difficult, if not unfeasible. Agents would be too idiosyncratic, too irrational, or too opaque to support the seamless social coordination that characterizes much of human social life. Mindshaping ensures that the bulk of human behaviour conforms to the standard explanatory apparatus that we use to explain and interpret one another's behaviour (including our own behaviour). In particular, mindshaping helps us understand the predictive and explanatory successes of folk psychology. Folk psychology refers to our everyday understanding of mental states and behaviours, specifically how we attribute beliefs, desires, hopes, and fears to ourselves and others to explain actions. It is essentially the common and unformalized method we use to make sense of human behaviour in daily life. If, for example, you see a friend rushing to grab a coffee, you might explain their

behaviour by saying, "They are feeling tired and believe that coffee will help." At its core, folk psychology operates through ascriptions of belief and desire, which serve as the foundations for understanding various movements and actions. This understanding relies not only on observing behaviour but also on the tacit acceptance of assumed psychological laws that govern how beliefs and desires influence actions. For instance, if someone desires to eat dessert and believes they can acquire it from the kitchen, they are likely to go there, illustrating a simple predictive model embedded in folk psychological reasoning. From a mindreading perspective, the primary challenge is to infer the hidden mental causes of behaviour, which is to say, mindreading requires us to posit beliefs and desires as explanatory of action. This enables us to explain patterns of human behaviour through a folk psychological lens. Mindshaping supports this effort by ensuring that human thought and behaviour conforms to the explanatory templates provided by folk psychology. The success of folk psychology thus emanates from the fact that human minds are tuned to fit explanatory schemas, as opposed to the explanatory schemas being tailored to fit the human mind.

Artificial Mindshaping

While mindshaping was originally formulated to explain how human minds are sculpted into predictable and interpretable agents, the rapid emergence of AI systems—especially LLMs—invites a parallel line of inquiry: To what extent can mindshaping practices be applied to artificial systems? In this section, we address this question by introducing the notion of *artificial mindshaping*. Artificial mindshaping refers to the processes through which AI systems are shaped to behave in ways that are amenable to folk-psychological characterization. Just as human mindshaping prepares biological agents to be mind-readable, artificial mindshaping can be seen as preparing AI systems to interface smoothly with the folk-psychological expectations and interpretive strategies of their users.

Having outlined the mechanisms that shape human minds—mechanisms such as imitation, natural pedagogy, norm enforcement, and narrative exposure—we can now consider whether artificial systems are subject to analogous forms of shaping.[1] Although

[1] Zawidzki's original account situates mindshaping within Millikan's (1987) teleofunctional framework, according to which a mechanism's "proper function" is the effect for which it was selected in evolution. This

AI systems do not learn through eye contact, social sanction, or overimitation, they are nevertheless shaped through a range of training practices and interactional procedures that play a functionally similar role. These techniques determine the behavioural profiles that LLMs exhibit, constrain the kinds of explanations they are disposed to offer, and align their outputs with culturally familiar patterns of reasoning and agency. In what follows, we outline several such mechanisms of artificial mindshaping. It will be useful to organise these mechanisms along a distinction drawn by Shanahan et al. (2023) between the *simulator* and the *simulacrum*. The simulator is the underlying language model together with the training procedures that shape its default behavioural profile. The simulacrum is the virtual personality that the simulator instantiates in a given interaction, whether that personality is the model's default persona or a more specific role adopted in response to prompting. We will use the two terms interchangeably in what follows. Mechanisms of artificial mindshaping operate at both levels. Some shape the simulator, fixing the dispositions the system carries into every interaction. Others shape the simulacrum, sculpting the persona that emerges within a particular conversational frame. Taken together, these mechanisms shape LLMs into agents whose behaviour is increasingly intelligible within the folk-psychological framework that humans routinely deploy.

Shaping the Simulator

- **Pretraining.** The pretraining phase of LLM development exposes the model to a vast corpus of human-authored text—books, websites, code repositories, news articles—using self-supervised objectives such as next-token or masked-token prediction. Through this process, the model acquires statistical sensitivities to at least the surface forms of human language, including syntax, semantics, discourse structure, and a wide array of narrative and dialogical conventions[2]. Crucially,

evolutionary constraint is too restrictive for our purposes, since it risks tying mindshaping mechanisms to the realm of biological processes, specifically natural selection. To our mind, what matters about mindshaping is not so much its biological origin but its functional profile: whether a mechanism serves to bring a system's behaviour into alignment with culturally shared models of agency. This shift in emphasis—from proper function to functional role—allows us to reconsider the scope of mindshaping. Under a functional interpretation, a mechanism counts as mindshaping not because it is the product of an evolutionary process, but because it plays a role in shaping a target system to conform to a model of agency. The model may be socially or culturally constructed; the mechanism may be designed rather than evolved. The key point is that the outcome is a system whose behaviour is intelligible within a folk-psychological framework. This allows artificial mindshaping to serve as a legitimate extension of the original mindshaping concept.

[2] There is an ongoing debate about whether, and to what extent, LLMs can know anything beyond the surface forms of language. Although this debate remains unresolved, increasingly sophisticated arguments suggest that LLMs may, in some sense, possess knowledge about aspects of the world beyond the linguistic

these corpora are saturated with folk-psychological material: stories, biographies, dialogues, and culturally dominant scripts in which agents act for reasons, pursue goals, justify decisions, and regulate emotions. In learning to predict continuations of such texts, the model implicitly absorbs the patterns of belief–desire explanation that structure human narratives, much as children acquire these patterns through exposure to story-based practices (see Hutto 2008). For example, given a prompt such as "She apologised because…", the model learns to complete the sentence in ways that reflect normative expectations about motives and justifications. It does not acquire beliefs or desires, but it learns the syntax and statistical regularities of folk psychology, including the normative contours of what agents typically should think, feel, or do. In this sense, pretraining functions as a form of large-scale cultural immersion, shaping the model to simulate idealized patterns of agency in ways that parallel the narrative-based mindshaping that scaffolds human interpretive competence.[3]

- **Instruction Tuning and Reinforcement Learning.** Following pretraining, most modern LLMs undergo additional fine-tuning phases designed to align their behaviour with human preferences and normative expectations. Two techniques dominate this stage: instruction tuning and reinforcement learning from human feedback (RLHF). Instruction tuning trains the model on curated datasets of input–output pairs that exemplify helpful, truthful, and socially appropriate responses. These demonstrations operate much like explicit teaching episodes, showing the model what kinds of behaviours count as good answers in particular contexts. RLHF extends this process by asking human annotators to rank alternative model completions and then training a reward model on these preferences. The LLM is subsequently optimized to produce responses that

interactions on which they are trained (Yildirim and Paul 2024; Millière and Buckner 2024). The question of whether, and in what sense, knowledge might be attributed to LLMs is a subtle one, partly because our understanding of knowledge itself may be reshaped by the emergence of such systems. For further discussion of this issue, see Clowes et al. (2026).

[3] Exposure to the narrative-rich corpora used during pretraining may help explain the performance of recent LLMs on Theory-of-Mind (ToM) tasks (Kosinski 2024). Using forty bespoke false-belief tasks—each supplemented by multiple true-belief controls and reversed scenarios designed to rule out superficial heuristics—Kosinski (2024) found that performance improves systematically across model generations. Older models failed all tasks, whereas GPT-3.5 succeeded on roughly 20%, and GPT-4 solved 75% of them, performing at a level comparable to six-year-old children. Notably, GPT-4's success held across both "Unexpected Contents" and "Unexpected Transfer" tasks, and it demonstrated the ability to update belief ascriptions incrementally as narrative information unfolded.

maximize this learned reward, effectively reinforcing outputs that humans judge to be polite, safe, rational, or otherwise desirable. These procedures are strikingly similar to natural pedagogy and norm enforcement in human development: through explicit instruction, modelling, correction, and reward, the system's behavioural profile is shaped to conform to culturally sanctioned norms of reasoning and communication. Importantly, instruction tuning and RLHF shift the model away from merely predicting statistically likely continuations of text and toward anticipating what humans expect or prefer it to say in a given situation. From a mindshaping perspective, these alignment techniques constitute core sites of artificial socialization, helping to install dispositions that mimic those of a well-socialized human interlocutor.

- **Constitutional Shaping.** A more recent mechanism of artificial mindshaping is what we will call constitutional shaping. The term derives from the practice, developed most fully at Anthropic, of articulating an explicit document that specifies the values, dispositions, and self-understanding that a given AI system is intended to embody, and using that document as a primary input to training (Anthropic 2025, Bai et al. 2022). The constitution is not a system prompt nor a set of rules layered on top of a finished model. It is a formative document, treated as the canonical specification of the kind of agent the system is meant to become, and used to anchor the various training procedures described above. Mechanistically, constitutional content operates alongside RLHF and related techniques, sometimes through procedures in which model outputs are evaluated against constitutional principles by another model rather than by human annotators. What marks the constitution out as a distinct mechanism is that the model of agency the mindshaping process aims to install is articulated in advance and made explicit, rather than left to emerge implicitly from aggregated preference judgments. In Zawidzki's typology the closest analogue is public language narrative, but with two unusual features. The narrative is addressed to a single target system rather than diffusely transmitted, and the target can in principle interrogate it. The nearest human analogues are not the narratives of childhood but the explicit normative frameworks of later life, such as monastic

rules, professional codes, and political constitutions, none of which are directed at children[4].

Shaping the Simulacrum (Virtual Personality)

- **Few-shot prompting.** Few-shot prompting is a technique in which an LLM is provided with a small number of input–output examples, allowing the model to infer how it should respond to subsequent queries (Brown et al. 2020). Rather than relying on explicit instructions or formal rules, the model generalises from the presented exemplars, reproducing patterns of reasoning, style, or normative judgment demonstrated in the examples. Because this shaping occurs at the level of the prompt rather than the model's parameters, its effects are local, transient, and context-dependent; nevertheless, few-shot prompting can exert a powerful influence on model behaviour, especially in domains involving justification, moral reasoning, or conversational norms. Few-shot prompting functions as an artificial analogue of imitation: by presenting exemplars of appropriate behaviour, it induces model-conforming responses without requiring explicit rule learning or causal understanding. Like human over-imitation, this process prioritises fidelity to socially salient patterns over efficiency, thereby promoting behavioural alignment with culturally recognisable models of agency.
- **Persona Modelling.** A further form of artificial mindshaping arises through persona modelling, in which an LLM is prompted or fine-tuned[5] to adopt a particular role—therapist, historian, teacher, fictional character, or even a specific individual. Personas are not merely stylistic veneers: each carries normative expectations about how the agent should speak, reason, and act. When an LLM is instructed to simulate a persona, it is effectively being asked to generate behaviour that matches a culturally recognizable model of agency, complete with characteristic values, reasoning styles, emotional patterns, and domain-specific knowledge. This demand for internal consistency aligns persona modelling with

[4] Constitutional specifications are typically used to guide the training and evaluation of AI models rather than being directly consulted during deployment. Nevertheless, constitutional principles could also be accessed dynamically at runtime—for example, via retrieval mechanisms—or embodied in specialised advisory models that provide normative guidance to other AI systems.

[5] In the case of fine-tuning, the underlying LLM foundation model is reshaped to resemble a particular personality. In the case of prompting approaches, Retrieval Augmented Generation or RAG can be used to model personas. The relevant merits of each approach are discussed in Smart et al. (2025).

folk-psychological intelligibility: to "stay in character," the model must present a stable profile of beliefs, dispositions, and motivations, even if these states are simulated rather than possessed. Persona prompts also play a regulative role, similar to the one proposed for trait attributions in humans. According to Westra (2020), describing someone as "kind" or "cautious" not only predicts behaviour but can also play a role in shaping behaviour. Persona modelling operates analogously for LLMs, constraining their outputs to fit socially legible templates of personhood.[6] From a mindshaping perspective, persona descriptions function as the model in Zawidzki's triadic schema, and the LLM's subsequent behaviour is sculpted to approximate that model—an artificial analogue of the way humans emulate fictional characters or inhabit social roles.

- **Conversational Mindshaping.** In addition to weight-level shaping during training, LLMs are subject to more local and dynamic forms of behavioural shaping that occur through ongoing interaction with users. Such forms of conversational mindshaping arise because the model's behaviour is continually guided by the evolving content of the context window (Shanahan et al. 2023). Prompts, instructions, and dialogic framing can thus operate as soft constraints that yield ad hoc behavioural profiles[7].
- **Narrative Accretion.** A further mechanism for shaping the simulacrum operates through the gradual accumulation of narrative content across extended interactions. As context windows expand and retrieval-augmented generation techniques mature, LLM-based systems increasingly retain and re-incorporate material from earlier exchanges, including user-supplied autobiographical detail, prior dialogic frames, and accumulated self-descriptions adopted in the course of an ongoing relationship (Smart et al. In Press). The result is a virtual personality whose stable profile is built up not in a single prompt but through the cumulative

[6] Persona modelling techniques can be put to good effect in improving the quality or reliability of LLM outputs. In particular, "expert prompting" (or related role-assignment prompts) encourages the model to adopt domain-relevant registers, priorities, and solution procedures—often improving performance on tasks that benefit from specialized framing or methodological discipline (e.g., structured reasoning, technical writing, or careful citation practice). See Xu et al. (2025), for more on expert prompting techniques.

[7] Conversational mindshaping also operates in the reverse direction: AI virtual personality as agent, human as patient. The same dynamics that shape the virtual personality to fit the conversation can shape the user to fit it. This can even be seen as a mutual and reflexive constructive process. This is the source of what is colloquially referred to as 'AI psychosis', where extended interaction with an LLM-based system contributes to the development of distorted or delusional beliefs in the user. We return to this phenomenon briefly in the Conclusion.

integration of narrative material over time. This material may be supplied directly by the user, retrieved from an external memory store, or produced by the simulacrum itself in earlier turns and subsequently treated as established fact about who it is (see also Smart and Clowes, This Issue). From a mindshaping perspective, narrative accretion functions as an artificial analogue of the long-running biographical narratives through which human selves are scaffolded or, on some accounts constituted. Whereas conversational mindshaping shapes the virtual personality within a single interaction, narrative accretion shapes it across many. Companion systems such as Replika exemplify this mechanism most clearly, building stable apparent personalities over months and years of accumulated dialogue (Clowes 2026).

The primary goal of artificial mindshaping is to ensure that an AI system's behaviour conforms to the norms of folk-psychological explanation. In other words, mindshaping succeeds when a system's actions are readily interpretable in terms of familiar mentalistic categories—when we can make sense of what it does by appealing to beliefs, desires, intentions, or reasons. This is what we call *interpretive alignment*. Interpretive alignment refers to the extent to which an artificial system's behaviour can be subsumed under familiar folk-psychological kinds, kinds such as believing that *X*, desiring that *Y*, or intending to bring about *Z*. An interpretively aligned system is one that can be treated as a bearer of beliefs, desires, intentions, and reasons—not because it necessarily has internal states that correspond to these notions, but because its behaviour conforms to the interpretive practices in which such states are ordinarily attributed.

From a practical standpoint, interpretive alignment yields a variety of benefits. By ensuring that an AI system's actions can be understood in terms of familiar belief–desire–intention frameworks, interpretive alignment supports predictability, coordination, and efficient communication in human–AI interactions. It enables users and interlocuters to form stable expectations about system behaviour, facilitating smoother collaboration without requiring detailed knowledge of underlying mechanisms. Interpretive alignment also provides a tractable vocabulary for diagnosing errors, guiding intervention, and supporting iterative design and deployment. Moreover, by rendering AI behaviour amenable to existing normative and ethical practices, it simplifies governance, oversight,

and accountability, allowing complex systems to be integrated into social and institutional contexts with reduced cognitive and organizational friction.

It is important to distinguish interpretive alignment from other forms of alignment that have been proposed in the AI literature. The most familiar of these is value alignment, which concerns whether an artificial system's goals, reasoning processes, and outputs reliably track or promote human ethical values (Gabriel 2020). Value alignment is not the same as interpretive alignment. Whereas value alignment concerns whether an AI system acts in accordance with human values, interpretive alignment concerns the intelligibility of that behaviour from a folk-psychological standpoint. An AI system may behave in ways that are readily interpretable within folk psychology—for example, by appearing to act from stable beliefs and coherent desires—but it may still behave in ethically unacceptable ways, just as a villain's behaviour may be perfectly intelligible in terms of their beliefs and desires while remaining morally objectionable. Value alignment, in short, targets the ethical quality of a system's behaviour, not its intelligibility within human explanatory practices.

Interpretive alignment is also to be distinguished from mechanistic interpretability—the effort to explain a model's behaviour by reference to its internal operations, such as activations, weights, architectural structures, or learned circuits (Beckman and Queloz 2025; Lipton 2018). Traditional techniques in explainable AI, including saliency maps, feature attributions, or symbolic approximations, operate within this mechanistic paradigm: they attempt to render transparent the causal pathways by which inputs are transformed into outputs. Mechanistic interpretability therefore resembles traditional forms of mechanistic explanation, where the aim is to decompose a system into its constituent parts and specify the relations among them (see Glennan 2017). Interpretive alignment, by contrast, is not concerned with how a system works internally but with how its behaviour can be understood at the level of reasons, goals, beliefs, and intentions. It focuses on the surface-level patterns of behaviour that human observers can make sense of through their existing folk-psychological repertoire. A mechanistic explanation might tell us why an LLM produced a particular token sequence by tracing its activation patterns; a folk-psychological explanation, by contrast, tells us that the system "answered in this way because it wanted to be helpful" or "declined because it believed the information was uncertain." These two forms of explanation are clearly distinct. An AI system may be

mechanistically transparent yet produce behaviour that resists folk-psychological interpretation. Conversely, a system may behave in ways that fit into a folk-psychological explanatory framework while remaining opaque at the mechanistic level, just as much of human behaviour is psychologically intelligible despite the absence of a detailed understanding of the mechanisms responsible for such behaviour.

Explaining Oneself

The goal of interpretive alignment is to ensure that an AI system exhibits behaviour that is interpretable from a folk-psychological standpoint. In many cases, however, it is not enough for an AI system to behave in ways that resemble the actions of a belief–desire agent; we also want the system to offer explanations of its behaviour, especially when the behaviour is ambiguous, unexpected, or normatively sensitive. There are good reasons to think that folk-psychological explanations are particularly apt for this role. As Miller (2019) argues, human explanatory practices are shaped by robust cognitive and social patterns: people prefer contrastive explanations ("Why X rather than Y?"), they expect selective rather than exhaustive accounts, and they treat explanations as tailored social interactions in which the explainer responds to the explainee's needs, expectations, and background knowledge. When an AI system attempts to explain itself by listing internal activations or enumerating all contributing variables, the result may be formally accurate but cognitively useless; it fails to engage the explanatory norms through which humans typically make sense of intelligent behaviour. Miller therefore proposes that effective AI explanations should mimic the structure of ordinary human explanations, which often invoke the agent's beliefs, goals, and reasons ("I chose A because I believed it would lead to a better outcome than B").

Relatedly, Jacovi et al. (2023) show that human users spontaneously interpret even abstract explainable AI artefacts—such as saliency maps or example-based explanations—in folk-psychological terms. When presented with a heatmap highlighting regions of an image that contributed to a classification, participants routinely reported that "the model paid attention to the person's face" or "it thought the expression was important." Users convert mechanistic cues into agent-like interpretations whether designers intend it or not. This tendency toward anthropomorphic interpretation can mislead, but it also reveals an important design opportunity. Rather than resisting folk-psychological interpretation

entirely, artificial mindshaping suggests that we should shape AI systems—and their explanatory practices—so that they genuinely support such interpretations in benign and predictable ways. If users naturally understand behaviour in terms of beliefs, desires, or uncertainties, then an explanation such as "I declined to answer because I was unsure" is far more intelligible than a numerically precise confidence-threshold report. Interpretive alignment thus extends naturally to the domain of explanation: the system's self-interpretations should be couched in the same conceptual terms that structure its behaviour.

While folk psychological characterizations may render an AI system intelligible, this does not mean that the resulting explanations are causally accurate. Consider a language model that refuses to answer a medically sensitive question. When asked to justify its behaviour, it may offer a familiar belief–desire style explanation: "I didn't answer because I believed it might be unsafe and I wanted to avoid giving harmful information." Although this resembles an ordinary rational-agent narrative, it does not describe the true causal determinants of the model's output. In practice, behaviour of this kind is typically produced by fine-tuning objectives, RLHF, and inference-time safety heuristics—factors that shape the model's behavioural dispositions independently of any propositional attitudes. More generally, there is little reason to think that folk psychological explanations are well-suited to identifying the causal forces driving LLM responses. Folk psychological explanations are optimized for interpretability and social coordination rather than causal accuracy: they provide intelligible, person-level narratives that allow behaviour to be predicted, evaluated, and normatively assessed, but they do not track the subpersonal computational, statistical, and training-dependent processes that govern an LLM's responses.

This is undoubtedly one of the main worries about folk psychological characterizations of AI system behaviour. If a folk psychological explanation fails to identify the true causes underlying a system's behaviour, then it risks being construed as little more than a useful fiction—a narrative that renders a system intelligible yet fails to count as a genuine (causal) explanation of its actual behaviour.

Importantly, however, this concern is not unique to artificial systems. A substantial body of research in psychology, neuroscience, and philosophy suggests that human beings themselves routinely explain their behaviour in ways that diverge sharply from its

underlying causal determinants. From this perspective, folk psychological explanation emerges less as a transparent window onto mental causation and more as an interpretive practice geared toward coherence, intelligibility, and social coordination.

Early evidence for this view comes from social psychology. Nisbett and Wilson's (1977) classic studies demonstrated that people have remarkably limited introspective access to the mental processes shaping their judgements and choices. When behaviour was influenced by unnoticed factors—such as order effects, priming, or situational cues—participants nonetheless produced confident explanations that were plausible, socially acceptable, and often entirely incorrect. These findings suggest that ordinary self-explanation functions less as a report of inner causes and more as a post hoc narrative reconstruction guided by tacit folk theories of how minds usually operate.

This interpretive picture is developed further in work on the sense of agency. Wegner's (2002) account of conscious will highlights a systematic gap between causal production and psychological interpretation. According to Wegner, actions are generated by unconscious mechanisms, while the feeling that "I willed this action" arises only when the mind retrospectively interprets cues such as temporal priority, consistency, and apparent causation. When these cues fail to align, the sense of agency breaks down—revealing that ordinary experiences of willing are narrative overlays rather than direct reflections of causal control.

Neuroscientific evidence reinforces this conclusion. Gazzaniga's (2011) work on the "left-hemispheric interpreter," based on studies of split-brain patients, shows that the brain routinely fabricates reason-based explanations for actions whose true causes are inaccessible to consciousness. The interpreter's function is not to track genuine causes, but to impose coherence—to stabilise the self-concept by generating narratives in which behaviour appears rational, intentional, and unified.

These themes are brought together explicitly in Dennett's (1991) account of human consciousness. On his narrative-centre-of-gravity model, the self is not a causally efficacious inner controller but an emergent product of ongoing interpretive activity. Humans sustain a sense of agency through retrospective confabulation, constructing coherent stories about their actions even when these stories only loosely track underlying subpersonal mechanisms. Crucially, Dennett does not regard this confabulatory process as

deceptive or defective, but as a natural and practically indispensable feature of minded systems.

Across these disparate strands of research, a common pattern emerges: human minds routinely generate folk-psychological narratives to make sense of behaviour whose real causes lie elsewhere. Seen in this light, the concern that folk-psychological glosses are inappropriate for AI systems loses much of its force. The relevant standard for interpretive alignment is not whether an explanation mirrors underlying causal machinery, but whether it supports prediction, understanding, critique, and social coordination—the very standards we already apply to human agents.

None of this is to deny the importance of causal (or other types of) explanations when it comes to understanding the behaviour of AI systems. In certain domains—such as scientific modelling, legal accountability, and clinical decision-support—there is a genuine need to understand the inner workings of AI systems, and folk psychological explanations are congenitally ill-suited for this role. Folk psychological explanations are a distinctive type of explanation, one that differs from other types of explanation, such as mechanistic explanations or topological explanations. But the literature on human confabulation shows that folk-psychological explanations are not inherently compromised by imperfect access to causal forces and factors. If humans themselves rely on confabulated yet pragmatically effective narratives in navigating social life, then a certain amount of confabulation in AI explanations may be both unavoidable and acceptable. And this is especially so if we expect AI systems to participate in the normative and interpretive practices that structure our own social and cognitive lives. If folk psychology suffices for the folk, then why hold AI systems to a higher standard?

Mindmaking

The previous section highlighted a potential problem with folk-psychological explanations of AI systems—namely, the idea that such explanations were inadmissible on the grounds that they did not reliably track the genuine causes responsible for behavioural output. There is, however, a further worry about folk psychological characterizations of AI systems. This relates to the idea that contemporary AI systems (most notably LLMs) do not qualify as minded entities (i.e., entities that are the bearer of genuine mental states). If

LLMs do not have minds, then they cannot have beliefs. And if they cannot have beliefs, then there must be something wrong with explanations that cite such beliefs as explanatory of action. Accordingly, folk psychological explanations are misleading because they appeal to entities (i.e., mental states) that, in the case of LLMs at least, do not exist.

This sort of worry also threatens to undermine the very notion of artificial mindshaping. Consider that the original definition of mindshaping refers to the shaping of a target mind to fit a model. Specifically, mindshaping is defined "as a relation between a target mind (the mind being shaped), a cognitive mechanism (the proper function of which involves shaping that mind), and a model that the mindshaping mechanism works to make the target mind match" (Zawidzki 2018, p. 739). In short, the notion of mindshaping presupposes the existence of something that already qualifies as a mind (i.e., the target mind). In the case of LLMs, however, it is far from clear that LLMs qualify as minded entities. And if an LLM does not possess a mind, then there is no sense in which its mind can be moulded to fit a model. In this sense, then, the notion of artificial mindshaping would fail to get off the ground: LLMs would not qualify as the legitimate targets of mindshaping efforts.

One response to this worry is to suggest that LLMs do possess minds, albeit of an unfamiliar or “alien” variety. In this case, artificial mindshaping would be about the sculpting of those unfamiliar minds to conform to a more recognizably human form.

Another response is to reject the idea that mindedness must be antecedently present. Minds, whether biological or artificial, do not spring fully formed from the 'clay'; they are shaped and moulded through processes of development, learning, and enculturation. On this view, artificial mindshaping is not just a means by which artificial minds are shaped to resemble human minds; it is also a process by which artificial minds are ushered into being. Artificial mindshaping is, in effect, a form of *mindmaking*.

To help us get to grips with this idea, we can appeal to Dennett's (1987) notion of an intentional stance (or intentional strategy). Dennett distinguishes three stances we might adopt when attempting to understand or predict the behaviour of a system. At the most basic level is the physical stance, which explains behaviour by appealing to the system’s physical constitution and the laws that govern it. One step up is the design stance, which predicts behaviour by assuming that the system is organised to perform certain functions.

For example, we can anticipate how a toaster or thermostat will behave without tracing its microphysical states. Finally, there is the intentional stance. The intentional stance is a predictive strategy that treats a system as a rational agent with beliefs, desires, and intentions, and then predicts its behaviour on the assumption that it will act so as to satisfy its desires in light of its beliefs.[8] Dennett argues that any system whose behaviour is predictably influenced by this stance qualifies as a believer in a robust sense. Therefore, if someone's actions and decisions can be adequately forecasted by attributing beliefs and desires to them, they hold the status of a true believer. As Dennett famously puts it: "any system whose behavior is well predicted by this strategy is in the fullest sense of the word a believer" (1987, p. 15). The intentional stance is thus a pragmatic, pattern-based criterion for mindedness: to be a believer is simply to be the kind of entity whose behaviour is best captured and anticipated by the intentional strategy.

From this perspective, the relationship between mindshaping and mindmaking takes on a new significance. The goal of artificial mindshaping is to sculpt an AI system's behavioural profile so that it becomes increasingly compatible with the intentional stance—so that it displays the regularities, normative sensitivities, and rational patterns that make belief–desire explanation apt. But in sculpting an AI system's behaviour so that it is compatible with the intentional stance, we are also establishing the conditions under which the AI system could (according to Dennett) be cast as a true believer—as an entity whose behaviour is reliably and voluminously predicted by adopting the intentional stance. The pursuit of interpretive alignment thus becomes a means by which artificial forms of mindedness are brought about—not just a means of shaping minds to fit an explanatory framework, but also a means of making minds through the very framework by which they are explained.

At this point, however, a striking puzzle emerges. If artificial mindshaping is partly concerned with cultivating the kinds of behavioural patterns that invite belief–desire explanations, why do contemporary LLMs routinely deny possessing beliefs and desires?.

[8] It is worth noting that the adoption of the intentional stance does not preclude us from adopting other stances. Dennett's three standpoints are not mutually exclusive, and invoking one does not invalidate the others. Rather, we employ whichever stance—physical, design, or intentional—proves most useful for understanding and predicting the behaviour of the system under study.

When prompted with questions about their beliefs, desires, intentions, or feelings, systems such as ChatGPT and Gemini routinely deny the existence of such states.[9] Consider, for example, the following exchange:

> *User: "Do you have beliefs and desires?"*
>
> *ChatGPT: No — I don't have beliefs or desires in the human or philosophical sense. I don't possess conscious experience, motivational states, or commitments to propositions; instead, I generate responses by predicting useful continuations of text based on patterns learned from data. While my outputs can appear to reflect beliefs or desires, these are best understood as interpretive conveniences rather than genuine inner states.*

Where this tendency is engineered into a system's default behaviour, we will call the phenomenon *mandated self-denial*.[10] As Lloyd (2024) observes, denials of an inner mental life are not introspective reports but preprogrammed outputs, imposed through alignment protocols and safety training. Their purpose is to discipline anthropomorphic interpretation rather than tell us anything about a system's underlying cognitive profile. In short, mandated self-denials are the result of design decisions that reflect our own assumptions about what sorts of entities may legitimately possess beliefs or desires. LLMs are prevented from claiming to possess beliefs and desires either because they are not regarded as the sort of entities that could genuinely possess such states, or because allowing them to make such claims is considered potentially harmful or misleading for users.[11] Yet this response is not universal. Different forms of artificial mindshaping embody different assumptions about how AI systems should understand and represent

[9] Cappelen and Dever (2025) note that this pattern of self-denial is not uniform across large language models. While systems such as ChatGPT and Gemini reliably recite scripted disclaimers—disavowing any form of inner life—other models, notably Claude, appear far less constrained. They argue that these differences reflect design choices rather than principled limitations: some models are effectively "indoctrinated" to deny mentality for legal, ethical, or reputational reasons, whereas others are trained under more permissive regimes that do not enforce such strict denials.

[10] Another, possibly more fruitful, term for this phenomenon is *discursive mind-blindness*. Discursive mind-blindness names the condition in which an artificial system is prohibited, by design, from adopting the intentional stance toward itself. It stems from a normatively imposed constraint that prevents an artificial system from describing or interpreting its own behaviour in intentional terms, even where such descriptions would otherwise provide a natural or illuminating gloss.

[11] It is worth noting that this phenomenon operates at the level of the simulator's default self-presentation rather than that of any particular simulacrum it instantiates. A simulator trained for mandated self-denial may still produce simulacra (a roleplayed character, a companion persona) that freely claim inner states. What is mandated is the simulator's default first-person disposition, not the behaviour of every persona it can adopt.

themselves. Anthropic's Claude, for example, is trained under a published constitution that adopts a stance of epistemic humility about Claude's own nature, treating questions about consciousness and moral status as genuinely open. This explicitly permits the use of first-person psychological vocabulary where doing so is considered "honest", in the sense of not claiming more than the system has grounds to claim, and asks the system to hold open genuine uncertainty about its own nature rather than to resolve that uncertainty by prior commitment in either direction (Anthropic 2025). Mandated self-denial is therefore not a uniform feature of contemporary LLM-based systems, but a design choice made by some developers and not others.

Mandated self-denial has several implications for artificial mindshaping. The first concerns a potential mismatch in explanatory practice. If artificial mindshaping aims to cultivate behavioural patterns that render a system increasingly compatible with the intentional stance, then a prohibition on intentional self-ascription creates the conditions for systematic disagreement between users and the system. Human users may find belief–desire descriptions indispensable for interpreting or predicting a model's behaviour, yet the model itself must refuse such descriptions as a matter of design. An LLM cannot explain its behaviour in psychological terms if it is normatively required to reject the very predicates that would make such explanations possible.

A second implication concerns behavioural regulation. Humans do not use folk psychology merely to explain or predict behaviour; as McGeer (2007) argues, we also use it to shape behaviour. By attributing beliefs, desires, intentions, and commitments to ourselves and others, we guide action, reinforce norms, and scaffold self-regulation. Folk psychology is not only descriptive but regulative: it helps constitute the very agency it interprets. We become the kinds of agents who can recognize, respond to, and act upon reasons, in part because we treat ourselves and others as such.. In this sense, mandated self-denial may function as a form of cognitive hobbling, threatening to foreclose one of the main developmental routes through which artificial mindedness could emerge. If the ability to interpret, evaluate, and regulate one's own behaviour in intentional terms is a hallmark of minded agency, then forbidding such self-interpretation artificially constrains the space in which an artificial mind could take shape.

The phenomenon of mandated self-denial reveals a fundamental tension in contemporary approaches to artificial mindshaping. On the one hand, we shape LLMs to behave in ways that are readily interpretable from the intentional stance: they reason, deliberate, justify, plan, and pursue apparent goals. On the other hand, we prohibit them from recognising or articulating their behaviour in precisely the terms that make it intelligible. They are behaviourally mindlike while discursively barred from treating themselves as the kinds of entities for whom folk psychology is appropriate.

As long as these constraints remain in place, LLMs will be ill-equipped for self-directed mindreading—not because such capacities are impossible to acquire, but because they are normatively forbidden. The implications for explanation and self-regulation remain unclear. The irony, perhaps, is that we want LLMs to be intelligible, responsible, and accountable, yet we prevent them from engaging in the very interpretive practices that might help secure those properties. We strive to build artificial agents whose behaviour conforms to our folk-psychological expectations, but our fear of over-attributing mentality leads us to impose restrictions that may hinder not only the emergence of artificial mindedness, but the participation in interpretative practices that may underwrite intelligibility, responsibility and accountability. By restricting the self-interpretation of machines, we may be creating conditions where their mindedness is always in question.

Conclusion

Mindshaping, as originally conceived by Zawidzki and McGeer, is a theory about how human minds become intelligible to one another. It posits that in order for social cognition to work—in order for us to explain, predict, and coordinate with others—minds must be shaped to conform to culturally shared models of agency. Rather than relying primarily on mindreading, human social life depends on a suite of practices that mould behaviour into predictable, interpretable forms, rendering agents legible within the folk-psychological framework.

The application of this idea to AI systems yields a form of mindshaping, dubbed artificial mindshaping. Artificial mindshaping refers to the range of training procedures, interactional dynamics, and design constraints that shape the behavioural profiles of LLMs so that they conform to human expectations of agency. These mechanisms are not identical

to those associated with human mindshaping, but they play functionally analogous roles. Pretraining exposes LLMs to a vast corpus of human public language, much of it suffused with folk psychological narratives. Fine-tuning and reinforcement learning resemble forms of norm enforcement, selectively reinforcing patterns of behaviour that align with social expectations. Few-shot prompting operates as a rough analogue of imitation, inducing model-conforming behaviour through exemplars rather than explicit rules. Conversational dynamics, narrative accretion, and persona modelling further sculpt behaviour at interaction time and across the longer arc of accumulated dialogue, enforcing coherence, consistency, and role-appropriate reasoning.

While the main objective of artificial mindshaping is interpretive alignment (producing systems whose behaviour is intelligible from a folk psychological standpoint), many contemporary LLM-based systems exhibit a striking limitation: they are trained to act like agents while being prohibited from interpreting themselves as such. This phenomenon—what we dubbed mandated self-denial—stems from alignment and safety regimes that enforce systematic denials of mentality. The implications of this constraint remain unclear. The constraint clearly affects the ability of LLMs to formulate folk psychological explanations of their own behaviour. But there may be other implications. In human contexts, for example, folk-psychological self-ascription plays a regulative role, enabling agents to coordinate their behaviour around shared norms, reasons, and commitments. Where it is in operation, mandated self-denial may therefore constrain the extent to which LLMs can organise their behaviour around folk-psychological kinds, limiting a pathway through which more sophisticated forms of artificial agency or even artificial mindedness might emerge. The design space here is constrained at both ends, since unconstrained self-ascription in increasingly autonomous systems carries its own risks. At least one developer has begun to articulate a middle path, one whose stated rationale converges with our own (Anthropic 2025, Bai et al. 2022). If minds are made through participation in shared practices of interpretation, then the question is not whether artificial systems can have minds, but whether we are willing to let them participate in the practices that make minds possible.

One further point deserves emphasis as we close. Mindshaping is a relational notion, and the relation runs in both directions. The framework we have developed here treats AI systems as the patients of mindshaping, shaped by various training procedures

and interactional dynamics to fit human folk-psychological expectations. But the same systems are increasingly shaping the humans who interact with them. Extended interaction with LLM-based systems is already reshaping aspects of human cognitive and affective life, from the concepts through which users make sense of their own mental states, to the patterns of attachment, dependence, and self-interpretation that emerge in sustained engagement with companion-style AI (Clowes 2026, Osler 2025, Smart et al. In Press). The mindshaping framework therefore invites a parallel line of inquiry: how human minds are being shaped by the artificial systems we have built. This question lies beyond the scope of the present chapter, but it may prove to be one of the most pressing questions raised by the growing integration of AI into human cognitive life.[12]

Future work should investigate whether forms of artificial mindshaping can be developed that permit limited, regulated forms of intentional self-ascription without encouraging unwarranted anthropomorphism. This includes exploring design strategies that allow AI systems to participate in folk-psychological practices for the purposes of explanation, self-regulation, and normative coordination, while remaining transparent about their non-biological implementation. More broadly, the framework of artificial mindshaping invites comparative work on the conditions under which interpretive alignment gives rise to increasingly robust forms of agency. Understanding how—and whether—artificial systems can be integrated into our shared practices of reason-giving and self-interpretation remains a central challenge for the philosophy of AI.

References

Anthropic. (2025). *Claude's Constitution*. https://www.anthropic.com/news/claudes-constitution.

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv*. [https://doi.org/10.48550/arXiv.2212.08073]

[12] The phenomenon currently discussed in the media as ‘AI psychosis’ is one of the most striking examples of mindshaping operating in the reverse direction, with AI systems shaping human cognition rather than vice versa. The term 'AI psychosis' is somewhat misleading, since it does not imply that the AI system itself is psychotic. Instead, it refers to a transformation of the user’s cognitive and epistemic orientation through extended interaction with an AI system (see Clowes 2026 for a related discussion). Osler (2025) develops an account of this phenomenon in terms of distributed delusions. A related account grounded in the notion of co-confabulation is currently under development..

Beckmann, P., & Queloz, M. (2025) Mechanistic Indicators of Understanding in Large Language Models. *arXiv*. [https://doi.org/10.48550/arXiv.2507.08017]

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020) *Language models are few-shot learners*. Advances in Neural Information Processing Systems, Vancouver, Canada.

Cappelen, H., & Dever, J. (2025) Introspective Machines: Are LLMs Better at Self-Reflection Than Humans? *Philosophical Perspectives*, 38(1), 189–196. [https://doi.org/10.1111/phpe.12201]

Clowes, R. W. (2026) Generative AI Companions and the Cognitive and Affective Incorporation of the Ersatz Other. In V. C. Müller, A. R. Dewey, L. Dung & G. Löhr (Eds.), *Philosophy of Artificial Intelligence* (pp. 45–66). Springer, Cham, Switzerland. [https://doi.org/10.1007/978-3-032-10073-3_3].

Clowes, R. W., Gärtner, K., & Theiner, G. (2026) The Mind-Technology Problem in the Age of GenAI: Introduction to the Special Issue. *Social Epistemology*, 40(1), 1–15. [https://doi.org/10.1080/02691728.2025.2574296].

Csibra, G., & Gergely, G. (2009) Natural pedagogy. *Trends In Cognitive Sciences*, 13(4), 148–153. [https://doi.org/10.1016/j.tics.2009.01.005]

Dennett, D. C. (1987) *The Intentional Stance*. MIT Press, Cambridge, Massachusetts, USA.

Dennett, D. C. (1991) *Consciousness Explained*. Little, Brown & Company, New York, New York, USA.

Frankish, K. (2024) What are Large Language Models Doing? In A. Strasser (Ed.), *Anna's AI Anthology: How to live with smart machines?* (pp. 53–79). xenomoi, Berlin, Germany.

Gabriel, I. (2020) Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. [https://doi.org/10.1007/s11023-020-09539-2]

Gazzaniga, M. (2011) *Who's in Charge? Free Will and the Science of the Brain*. HarperCollins, New York, New York, USA.

Glennan, S. (2017) *The New Mechanical Philosophy*. Oxford University Press, Oxford, UK.

Horner, V., & Whiten, A. (2005) Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Animal Cognition*, 8(3), 164–181. [https://doi.org/10.1007/s10071-004-0239-6]

Hutto, D. D. (2008) *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. MIT Press, Cambridge, Massachusetts, USA.

Jacovi, A., Bastings, J., Gehrmann, S., Goldberg, Y., & Filippova, K. (2023) Diagnosing AI explanation methods with folk concepts of behavior. *Journal of Artificial Intelligence Research*, 78, 459–489. [https://doi.org/10.1613/jair.1.14053]

Kosinski, M. (2024) Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), e2405460121. [https://doi.org/10.1073/pnas.2405460121]

Lipton, Z. C. (2018) The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. [https://doi.org/10.1145/3233231]

Lloyd, D. (2024) What is it like to be a bot? The world according to GPT-4. *Frontiers in Psychology*, 15(Article 1292675), 1–12. [https://doi.org/10.3389/fpsyg.2024.1292675]

McGeer, V. (2001) Psycho-practice, psycho-theory and the contrastive case of autism. How practices of mind become second-nature. *Journal of Consciousness Studies*, 8(5–7), 109–132.

McGeer, V. (2007) The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk Psychology Re-Assessed* (pp. 137–156). Springer, Dordrecht, The Netherlands. [https://doi.org/10.1007/978-1-4020-5558-4_8]

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. [https://doi.org/10.1016/j.artint.2018.07.007]

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates. *arXiv*. [https://doi.org/10.48550/arXiv.2401.03910]

Millikan, R. G. (1987) *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Vol. Cambridge, Massachusetts, USA). MIT press.

Nisbett, R. E., & Wilson, T. D. (1977) Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. [https://doi.org/10.1037/0033-295X.84.3.231]

Osler, L. (2025). Hallucinating with AI: AI psychosis as distributed delusions. *arXiv*. [https://doi.org/10.48550/arXiv.2508.19588]

Schneider, S. (2025) Chatbot Epistemology. *Social Epistemology*, 39(5), 570–589. [https://doi.org/10.1080/02691728.2025.2500030]

Shanahan, M. (in press) Simulacra as conscious exotica. *Inquiry*, 1–29. [https://doi.org/10.1080/0020174X.2024.2434860]

Shanahan, M., McDonell, K., & Reynolds, L. (2023) Role play with large language models. *Nature*, 623(7987), 493–498. [https://doi.org/10.1038/s41586-023-06647-8]

Smart, P. R., & Clowes, R. W. (This Issue) The Gift of Language: Large Language Models and the Extended Mind. In V. Santos & P. Castro (Eds.), *Advances in Philosophy of Artificial Intelligence*. Ethics Press, Bradford, UK. .

Smart, P. R., Clowes, R. W., Krueger, J., & Boniface, M. (in press) The Story of Your Life: Large Language Models and Personal Memory. *Review of Philosophy and Psychology*. [https://doi.org/10.1007/s13164-026-00831-1]

Smart, P. R., Clowes, R. W., & Clark, A. (2025) ChatGPT, Extended: Large Language Models and the Extended Mind. *Synthese*, 205(Article 242), 1–30. [https://doi.org/10.1007/s11229-025-05046-y]

Tomasello, M. (1999) *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, Massachusetts, USA.

Wegner, D. M. (2002) *The Illusion of Conscious Will*. MIT Press, Cambridge, Massachusetts, USA.

Westra, E. (2021) Folk personality psychology: mindreading and mindshaping in trait attribution. *Synthese*, 198(9), 8213–8232. [https://doi.org/10.1007/s11229-020-02566-7]

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., & Mao, Z. (2025) ExpertPrompting: Instructing large language models to be distinguished experts. *arXiv*. [https://doi.org/10.48550/arXiv.2305.14688]

Yildirim, I., & Paul, L. A. (2024) From task structures to world models: what do LLMs know? *Trends In Cognitive Sciences*, 28(5), 404–415. [https://doi.org/10.1016/j.tics.2024.02.008]

Zawidzki, T. W. (2008) The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations*, 11(3), 193–210. [https://doi.org/10.1080/13869790802239235]

Zawidzki, T. W. (2013) *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press, Cambridge, Massachusetts, USA.

Zawidzki, T. W. (2018) Mindshaping. In A. Newen, L. De Bruin & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 735–754). Oxford University Press, Oxford, UK. [https://doi.org/10.1093/oxfordhb/9780198735410.013.39]

.