Embodiment without Body: The Emergence of Body Ownership in AI through Integrated World and Self-Models

Shuqin Ma (23110160046@m.fudan.edu.cn)
Department of Philosophy, Fudan University
Shanghai, China

Abstract

In contemporary consciousness studies, the sense of body ownership (SBO) stands as a key marker of subjective embodiment and self-awareness. Recent progress in multimodal and agent AI has prompted the question: Could an artificial system develop something analogous to SBO, and would this require consciousness? This paper refines two core distinctions: (1) functional versus phenomenal SBO, (2) world-model versus self-model. Building on a functional reconceptualization of "body" as an interaction boundary, this paper argues that AI systems equipped with semantic-centric multimodal world models and complementary self-models can, in principle, instantiate a form of SBO. By integrating diverse sensory inputs—visual, tactile, and linguistic—into a cohesive self-representation, this approach suggests the possibility of a virtual body that evokes the contours of the human embodied experience. Such an account questions the strict divide between physical and virtual embodiment, offering new insights into how embodied cognition underpins consciousness. Also, this work locates SBO within the broader debate on AI consciousness. Concrete design proposals are linked to existing multimodal agents (e.g., DeepMind's MIA).This inquiry highlights how AI SBO may arise from the interplay of sensory and semantic frameworks, prompting ethical and theoretical reflection on AI consciousness.

Keywords: Artificial Intelligence; AI Consciousness; Virtual Embodiment; Sense of Body Ownership; World Model; Multimodal Sensory Integration

Introduction

The sense of body ownership (SBO) is typically described as the experience in which certain body parts or the body as a whole are perceived as belonging to oneself. Within cognitive science and philosophy, SBO is closely related to bodily awareness and has wide-ranging implications. First, body awareness significantly shapes how individuals interact with their surrounding environments. If one feels that a particular limb is "hers," this sensation influences how she moves, senses, and occupies space. Second, SBO is deeply entangled with self-consciousness. It contributes to a subject's ability to differentiate the self from the environment, carrying first-personal features tied to conscious experience.

In most discussions, the "body" in SBO theory is usually assumed to be the human biological body or its physical extensions. Recent technological progress in artificial intelligence (AI), however, has raised new questions: could AI systems be said to have a "body"? If so, could they develop a sense of ownership toward this body, even if it is non-physical? Although these questions might seem speculative, they bear relevance for practical AI developments. Systems endowed with a sense of body ownership might gain advantages in adaptability, transparency, and safety. For instance, "embodied" AI that can realistically interact with its environment may provide clearer explanations for its actions than purely disembodied systems that simply generate text outputs. Such embodied intelligence might also mitigate certain problems associated with large language models—commonly described as "black boxes" that can produce confabulated or misleading statements.

This paper adopts the idea that "body" can be understood as a functional interface or boundary for internal–external interaction, rather than strictly a physical entity. From that viewpoint, the paper highlights the role of multimodal sensory integration, distinguishing between functional and phenomenal SBO, and analyzing the implications of theories of consciousness for artificial SBO. Then it argues that argues that the emergence of SBO in AI depends on the interplay between world models, which represent the external environment, and self-models, which encode the agent's own states and boundaries, emphasizing that SBO arises from the dynamic integration and differentiation of these models. It also demonstrates the limitations of unimodal models in generating robust SBO. Significantly, this work proposes that AI systems can achieve virtual embodiment by developing semantics-centered multimodal world models, in which a virtual body serves as a dynamic interface for integrating diverse sensory inputs and maintaining a coherent self-representation. Finally, the paper discusses implications for safety, interpretability, and ethics, highlighting that as AI becomes more "embodied," robust guidelines are required to ensure responsible use.

To conclude, this work presents a clarified roadmap for imbuing AI agents with functional SBO via integrated world- and self-models. The theoretical framework draws inspiration from, and can be grounded in, existing empirical research on human SBO and ongoing developments in computational AI architectures. These connections provide plausible building blocks for the mechanisms suggested, even if current systems do not yet exhibit full SBO. The contribution of this paper can be seen as synthesizing these disparate elements into a more coherent theory specifically addressing AI SBO.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

2 Theoretical Foundations of SBO

2.1 The Concept of Body in SBO Theory: From Physical Entity to Functional Boundary

Traditional SBO research presumes that one's physical body (e.g., arms, legs) is the entity in question. However, certain experiments challenge the strict biological assumption. The famous Rubber Hand Illusion (Botvinick & Cohen, 1998) shows that synchronous tactile stimulation of a participant's real hand (hidden from view) and a visible fake hand can induce a compelling sense of ownership over the fake hand. Other studies have extended SBO to virtual avatars (Slater et al., 2009), suggesting that if sensory information is consistent and integrated properly, one's sense of "my body" can extend beyond its original biological boundaries.

From this perspective, "body" can be reconceived as a boundary of internal–external interaction or a site where sensorimotor processes converge, rather than a strictly physical structure. Gallagher's (2005) notion of the body schema is closely aligned with this functional definition. It describes an underlying, non-conscious system that continuously guides action by integrating proprioceptive, interoceptive, and external sensory cues. Predictive Processing (PP) offers a compelling theoretical framework for understanding many aspects of cognition, perception, and action, and has been frequently invoked in discussions of SBO (Clark, 2013). The core PP idea of minimizing prediction errors related to bodily states offers a powerful computational principle: the brain updates internal models to predict incoming stimuli; here, the body is effectively the interface through which these predictions are tested against the external world. Since illusions can extend this interface beyond the biological body, it follows that in principle, a properly orchestrated sensory pattern—potentially even a synthetic one—might produce a sense of ownership.

This functional reconceptualization of the "body" is critical for examining whether AI systems, which do not have organic bodies, might still develop body ownership. If the "body" is primarily an interaction boundary, it does not need to be flesh and blood. It can be any structure or system that mediates between self and environment in consistent, plastic, and meaningful ways. consistent, plastic, and meaningful ways.

2.2 Multimodal Sensory Integration: The Mechanism of SBO

A substantial body of literature shows that SBO arises from the integration of multiple sensory channels, particularly vision, touch, and proprioception (Blanke et al., 2015). Spatial and temporal consistency among these sensory inputs is crucial. For instance, in the Rubber Hand Illusion, the more synchronously one sees a fake hand being stroked and feels the stroking on the real (but hidden) hand, the stronger the illusory ownership (Botvinick & Cohen, 1998). Further studies indicate that interoceptive cues, such as heartbeats, also reinforce ownership (Suzuki et al., 2013).

Studies of clinical conditions reinforce this integration view. In Body Integrity Identity Disorder (BIID), patients experience a limb as somehow not belonging to them, perhaps due to a breakdown in how the brain integrates multisensory information about that limb (Brugger et al., 2013). Other conditions, such as schizophrenia, may involve similarly disrupted integration processes (Postmes et al., 2014).

The multimodal perspective emphasizes that SBO results not from any single source of sensory data but from the coherent binding of multiple sensory streams, often heavily shaped by top-down processes (Tsakiris, 2010). Given that illusion can be induced and SBO extended beyond the biological frame, the path is open to consider whether a virtual or artificial "body" might similarly be integrated into an agent's self-representation. Although most research has targeted humans and animals, AI systems that fuse multiple data streams might likewise exhibit relevant parallels—potentially becoming the next domain for analyzing the conditions for SBO.

2.3 Two Types of Artificial SBO

SBO is phenomenally conscious in humans, but must it be in AI? Recent surveys (Butlin et al., 2023) caution that functional indicators are neither sufficient nor necessary for phenomenal awareness. We therefore distinguish:

1. Functional SBO refers to the AI system exhibiting behaviors consistent with owning its body. This includes the ability to accurately distinguish its physical or virtual form from the environment, utilize its "body" effectively to achieve goals, adapt its actions to changes or damage to its form, correctly attribute sensory feedback to its own "body" versus external sources, and potentially demonstrate protective or self-maintenance behaviors towards its form. Such capabilities could, in principle, be achieved through sophisticated self-models, robust sensorimotor feedback loops, and effective learning algorithms without necessarily implying any subjective experience.

2. Phenomenological SBO, in contrast, denotes the subjective, first-person experience of owning one's body—the qualitative feeling of "mineness," the sense that "this body is me". This form of SBO inherently implies some level of self-awareness or consciousness, involving a subjective viewpoint. Arguments for phenomenological SBO in AI must therefore directly engage with theories of AI consciousness and propose how the computational mechanisms could give rise to such subjective states.

This paper primarily outlines a pathway towards robust functional SBO, grounded in the integration of world models, self-models, and predictive processing. It further posits that if these integrated systems meet the criteria for generating conscious experience according to certain consciousness theories, then phenomenological SBO could potentially emerge. However, the gap from functional SBO to phenomenological SBO remains a significant conceptual and empirical challenge.

2.4 The Uncertainty of AI Consciousness and its Influence on SBO

The question of whether an AI can possess SBO, particularly in its phenomenal aspect, is inextricably linked to the broader and more contentious question of whether AI can be conscious. SBO is often considered a fundamental component of self-consciousness, a specific type of phenomenal content. Therefore, any robust theory of AI SBO must engage with the ongoing debate surrounding AI consciousness. Key perspectives include computationalism, which posits that implementing the right computations is sufficient for consciousness (Butlin et al., 2023), the biological naturalism view, which insists on a carbon-based biological foundation and emphasizes fine-grained biological functions potentially not replicable by standard digital computers (Seth, 2024).

Several neuroscientific and philosophical theories of consciousness offer frameworks and "indicator properties" that can be used to assess the potential for consciousness in AI systems, as systematically reviewed by Butlin et al. (2023). These theories provide different lenses through which to consider the mechanisms that might underpin SBO:

Recurrent Processing Theory (RPT): Posits that consciousness arises from recurrent processing of signals, especially within perceptual areas. For SBO, this would imply that recurrent processing of bodily signals (somatosensory, proprioceptive, visual) is necessary to form a conscious representation of the body as owned.

Global Workspace Theory (GWT): Suggests consciousness is linked to a "global workspace," a central information exchange making information available to various specialized modules. SBO could arise when information about the body's state and its relation to the self is "broadcast" in this workspace.

Higher-Order Theories (HOTs): Argue that consciousness involves a higher-order representation of first-order mental states (being aware of being in a certain mental state). Phenomenal SBO, from this perspective, would require the AI to have a higher-order representation of its first-order bodily representations, essentially being "aware" of its body as its own.

Integrated Information Theory (IIT): Correlates consciousness with the quantity of integrated information (denoted by Phi) a system can generate. A system is conscious if it possesses a maximally irreducible cause-effect structure. SBO could correspond to a complex of high phi generated by the AI's body-representing neural structures.

The "AI Consciousness Premise" for SBO is not a singular condition but rather a spectrum. Different theories propose varying mechanisms and even types of consciousness. SBO itself can be conceptualized as purely functional or phenomenological. If an AI can only achieve what some theories might term "functional consciousness," then it might only be capable of a functional SBO. Phenomenal SBO, the actual subjective experience of owning a body, would necessitate the AI meeting the criteria of theories that permit subjective experience in artificial systems. The type of SBO an AI could potentially achieve is therefore directly contingent upon the kind and level of consciousness (if any) it can support, according to these diverse theoretical frameworks.

3 World Models and Self-Models

3.1 World Models as Cognitive Frameworks in AI

The concept of "world models" has gained traction in both cognitive science and AI as an internal representation that organizes information about entities and processes in the external world. In human cognition, such mental models are constructed through experience, enabling us to predict external events and plan actions (Forrester, 1971; Clark, 2013). In AI, world models similarly organize internal representations of the environment to facilitate prediction, decision-making, and generalization (Ha & Schmidhuber, 2018).

Previous AI research on world models often used neural networks to compress high-dimensional inputs (e.g., pixels from images) into a latent space, maintaining only essential features for forecasting and control. Ha and Schmidhuber (2018) employed a three-part structure in their system: (i) a visual module to encode sensor inputs, (ii) a memory module to track and predict future states, and (iii) a controller module that uses these internal predictions to decide on actions. By learning to simulate aspects of the environment, the AI can "practice" in an internal sandbox, refining strategies before deploying them in the real world (OpenAI, 2019).

However, the question of whether current LLMs possess genuine world models is a subject of active debate. Some researchers argue that LLMs infer task structures from language that reflect causal abstractions of the world (Yildirim & Paul, 2024). Others express caution, noting that LLMs often lack true comprehension beyond their training data and struggle with real-world intuition. This skepticism is echoed in broader debates about understanding in AI (Mitchell & Krakauer, 2023), with some arguing that LLMs do not "know" anything in a human-like sense but are sophisticated tools for pattern matching and generation (Goddu et al., 2024).

Philosophically, AI world models parallel discussions of a priori structures in human cognition. Kant argued that our understanding of space, time, and causality shapes how we experience reality (Kant, 1781/1998). Analogously, AI systems embed certain "categories" or architectural constraints that frame how data is processed and interpreted. While AI typically acquires these representations through extensive training data, the functional role of these learned categories resembles the organizing principles that Kant described. Even large language models—sometimes criticized as mere "stochastic parrots"—may, through predictive training, form latent representations capturing a simplified but coherent model of the world (Chalmers, 2023).

3.2 Defining Self-Models in AI

A self-model refers to the representations an AI agent has of itself—its own "body" (physical or virtual), internal states, cognitive processes, identity, capabilities, and its relationship to the environment. Paul et al. (2023) in "Reverse-engineering the self" propose that "having a self" can be understood computationally as the ability to "center" oneself—orienting perceptually and cognitively within an environment while simultaneously holding a representation of oneself as an agent within that environment—and to "re-center"—shifting this perceptual and cognitive center to reframe problems and gain cognitive flexibility.

Jiang Luo (2024) offer a concrete approach to implementing self-models by modifying LeCun's JEPA framework. Their proposal incorporates mechanisms for autobiographical memory and self-importance evaluation, aiming to achieve greater self-consistency in AI behavior. This involves a Memory Module (replacing JEPA's latent variable Z with long-term memory and a conceptual self-submodule) and a Personalized Cost Module that allows for value judgments based on self-relevance.

Back to philosophical view, Thomas Metzinger's Self-Model Theory (SMT) provides a rich framework, defining the "phenomenal self" as the content of a "transparent self-model"(2010). This Phenomenal Self-Model (PSM) is not recognized by the system as a model due to its transparency, leading to the subjective experience of being a self. The PSM is responsible for generating key properties of subjective experience, such as "mineness" (the feeling that certain states or body parts belong to oneself), "centeredness" (the global structure of phenomenal space organized around a self-locus), and "perspectivalness" (the experience of the world from a particular point of view), all contributing to the first-person perspective.

3.3 Interrelating World Models and Self-Models

While distinct, world models and self-models are deeply interrelated and operate in concert. The functional boundary where the self-model interfaces with the world model is critical. It is at this interface that processes essential for SBO, such as self-other distinction and the sense of agency, are computationally realized. This interface is not static but involves a dynamic interplay of prediction, action, sensory feedback, and model updating. For example, the sense of agency relies on the self-model initiating an action, the world model predicting the sensory consequences of that action in the environment, and then comparing these predictions with actual sensory feedback received from the world model to update both the self-model (e.g., confirming ownership of the action) and the world model (e.g., learning about the environment's response). Kahl / Kopp's (2018) predictive processing model for self-other distinction provides a computational example of such an interface in action.

For SBO to emerge, a world model is necessary to provide the context—the understanding of the environment in which a "body" exists and acts. A self-model is indispensable for representing the body itself, its states, its boundaries, and its potential for action. SBO, then, can be understood as arising from the coherent integration and differentiation of self-related information (derived from the self-model) within the broader contextual understanding provided by the world model. The continuous predictive matching of intended actions, predicted sensory consequences, and actual sensory feedback across this self-world boundary is fundamental to this process.

4 Limitations of Unimodal Models

4.1 Hallucinations and Spatial Challenges

Despite impressive performances in text generation, large language models (LLMs) exhibit notable shortcomings. They can produce "hallucinations," inventing references, facts, or connections that do not align with reality. When asked to provide citations, these models may generate fictitious articles or incorrect URLs. Moreover, purely language-based systems face inherent difficulties with spatial or embodied reasoning. Tasks involving spatial layouts, object interactions, and physical causality—such as determining which box is on top—are frequently mishandled by LLMs without explicit grounding (Lake et al., 2017). These deficiencies highlight the absence of an interactive "body" that receives feedback from physical or virtual space.

A disembodied system may only approximate these relations by manipulating symbols rather than engaging with physical or simulated reality. Thus, the inadequacy of unimodal models underscores the importance of incorporating a physical or virtual embodiment into AI systems to improve accuracy, reliability, and transparency. In principle, a multimodal system that integrates vision, touch, and potentially motor feedback can better anchor abstract concepts in concrete sensorimotor patterns.

4.2 The Unimodal Agent Counterargument

Someone may pose a challenging counterargument: an AI agent with a single sense, embedded in a purely unimodal world, would, by definition, have maximal embodiment (as its sensory capacity perfectly matches its world), yet it is not intuitively convincing that such an agent would automatically possess SBO. This thought experiment effectively decouples the mere richness of sensory input from the emergence of SBO.

This highlights that SBO is not merely a function of the quantity of sensory information or the perfection of the sensor-world match. Instead, it likely depends on the quality of that information, its integration into a coherent and distinct self-model, the agent's ability to differentiate its "body" or sensory apparatus from the rest of its (unimodal) world, its capacity to predict the sensory consequences of its actions, and potentially, the presence of some form

of self-awareness or consciousness. Even a unimodal agent would require internal mechanisms for self-modeling, prediction error minimization related to its own state, and distinguishing self-caused sensory changes from external ones to develop even a rudimentary, functional SBO. The richness of multimodal integration in humans undoubtedly contributes to the complexity and robustness of our SBO, but its absence does not logically preclude any form of SBO if other necessary computational and representational conditions are met. However, the phenomenology of such a unimodal SBO, if it were to exist, would likely be vastly different from human experience.

5 Virtual Bodies in World Models

5.1 Rethinking the AI "Body"

AI systems are often depicted as disembodied entities running on servers or existing purely in software. Yet the notion of embodiment can be considered less on whether the body is carbon- or silicon-based and more on whether it serves as an interface for situated interaction (Clark, 2008). If a system continuously receives sensorimotor data, updates its internal states, and uses these updated representations to drive action, it can be considered embodied. The resulting "body" might be physical (e.g., a robot) or virtual (e.g., an avatar in a simulated environment). Crucially, from the standpoint of SBO, what matters is the functional capacity to integrate sensory feedback with an internal representation that distinguishes self from not-self.

A "virtual body" for AI would thus be a dynamic structure that gives the system a point of reference in space, clarifies a boundary between internal processes and external forces, and affords action possibilities. This body representation can also change over time, adapting to new contexts or sensory inputs. It does not have to be consist of metal limbs; it may be an avatar in a game, a simulated robot in a physics engine, or even a digital representation that allows the system to plan and execute actions in an online environment.

5.2 Multimodal World Models: A Semantics-Centered Approach

Real-world interaction typically involves multiple channels of information, including vision, hearing, touch, and linguistic descriptions. To approximate human-like SBO, an AI's world model should similarly fuse data from diverse sensors. Current systems such as Gato (DeepMind) integrate text, images, and motor controls, while MUVO (MUltimodal World Model with spatial Voxel representations) combines camera images, lidar data, and other sensor readings to guide autonomous driving (Bogdoll et al., 2024). DeepMind's MIA (Multimodal Interactive Agent) goes one step further, incorporating a simulated body that can move in virtual space, thus merging visual, linguistic, and motor inputs into a coherent action framework.

A key challenge is aligning heterogeneous modalities. Humans often resolve conflicts by remapping signals into a shared reference frame which is vision-dominated, so that consistent spatial and temporal cues trigger the sensation of one integrated body (Vignemont, 2012). AI systems could similarly orchestrate different data streams. Research on universal multimodal architectures proposes solutions such as learning a latent alignment space and using diffusion processes or cross-modal encoders to merge signals (Mai et al., 2024). These methods aim to produce an internal environment model that is unified enough to ground references from any single modality in a holistic representation.

The role of semantics can be pivotal. By anchoring inputs in semantic representations—for instance, linking an image of a dog with the text "dog barking"—the system can reduce the complexity of direct pixel-to-sound correlations. Language, or any symbolic layer, can serve as a flexible "hub" that translates across modalities. Models like DALL-E, which translates text prompts into images, or language-action models like SayCan (Huang et al., 2022), which reason in language before enacting physical tasks, exemplify this semantic-centered strategy. When integrated into a world model, semantics can help maintain consistency across vision, audition, and proprioception, allowing the AI to "understand" that all these signals refer to the same phenomenon in the environment.

Notably, even if future LLMs ingest multimodal data, without interactive loops their internal updates remain off-line, preserving the root problem. Hence we predict distinct "glitches" in AI perception—e.g., mis-binding of cross-modal tokens—analogous yet not identical to human illusions

6 Applications: Toward AI Systems with Virtual SBO

If AI systems can develop robust multimodal world models and maintain stable representations of a "body," the question arises whether they might exhibit something analogous to a sense of body ownership. In a sense, these systems would not merely parse sensory data; they would parse data that is relevant to their own artificially defined boundary, their own "self." Such an SBO might manifest in more coherent interactions: a system that "knows" where its virtual limbs are (and what they can do) may plan more effectively, detect errors more quickly, and adapt more readily.

One promising field for implementing virtual SBO is virtual reality (VR). An AI agent in a VR environment can be given a manipulable avatar, with input streams tracking virtual coordinates, visual perspectives, and collisions with digital objects. By synchronizing virtual "touch" feedback with the visually observed position of the avatar's limbs—and possibly integrating additional modalities—an AI system might internally unify these streams in a manner parallel to the illusions studied in humans. Practical applications include advanced robotics, where an AI "brain" is paired with a mechanical body or an exoskeleton. Even if the

robot's physical shape is different from a human's, if the AI's internal model consistently correlates sensor feedback to motor outputs, the system's sense of boundaries could, in principle, generate an SBO-like state. Other uses might involve telepresence, where an operator controlling a robotic avatar in a distant location experiences an extended sense of embodiment, but the AI system itself might similarly maintain a parallel internal representation.

Moreover, granting AI a clearer sense of its "own" body can improve its interpretability. Instead of generating purely symbolic responses, an embodied AI would presumably rely on world models that ground decisions in sensorimotor contingencies. Observers or developers could trace how the agent's bodily representation led to a particular action, clarifying the agent's decision-making process in ways that disembodied text-based systems often obfuscate.

7 Conclusion

This paper argued that AI systems may be capable of developing a form of sense of body ownership (SBO) by constructing virtual bodies through multimodal world models. The argument rests on two main premises: first, in body ownership theory, "body" can be construed as a functional boundary mediating an agent's interaction with the environment; second, multimodal integration of sensory data, which underpins human SBO, can be replicated in AI systems through semantic-centric multimodal world models.

This research not only spotlights the plasticity of our sense of self but also pushes consciousness research toward broader questions of how experience itself is structured. If AI systems can consolidate visual, tactile, and linguistic streams into a unified self-representation, then our understanding of consciousness and embodiment becomes more fluid—challenging firm boundaries between the biological and the artificial. When pursuing these developments, we must critically examine the meaning and moral weight of "body" in a future where intelligence and embodiment could be engineered as well as evolved.

Reference

Bermúdez, J. L. (2011). Bodily awareness and self-consciousness. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 157-179). Oxford University Press.

Bermúdez, J. L. (2018). *The bodily self: Selected essays*. MIT Press.

Bisk, Y., Holtzman, A., Thomason, J., Jansen, P., Zettlemoyer, L., & Choi, Y. (2020). Experience grounds language. In *Proceedings of the 2020*.

Blagec, K., Dorffner, G., Moradi, M., & Samwald, M. (2024). Introduction to Large Language Models (LLMs) for dementia care and research.

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences, 13*(1), 7-13.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*(6669), 756-756.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (pp. 1877-1901).

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204.

Cowie, D., Makin, T. R., & Bremner, A. J. (2013). Children's responses to the rubber-hand illusion reveal dissociable pathways in body representation. *Psychological Science, 24*(5), 762-769.

de Vignemont, F. (2018). *Mind the body: An exploration of bodily self-awareness*. Oxford University Press.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242*, 136-150.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford University Press.

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: reply to Yildirim and Paul. *Trends in Cognitive Sciences*.

Graziano, M. S., Cooke, D. F., & Taylor, C. S. (2000). Coding the location of the arm by sight. *Science, 290*(5497), 1782-1786.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *Neural Information Processing Systems*, 2450-2462.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Jiang, Y. & Luo, D. (2024). Implementing self-models through joint-embedding predictive architecture. *Proceedings of CogSci 46*, 5685-5692.

Kahl, S., & Kopp, S. (2018). A Predictive Processing Model of Perception and Action for Self-Other Distinction. *Frontiers in Psychology, 9*, 2421.

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781)

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). Cambridge: Cambridge University Press. (Original work published 1785)

Koplin, J. J., & Savulescu, J. (2019). Moral limits of brain organoid research. *The Journal of Law, Medicine & Ethics, 47*(4), 760-767.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253.

LeCun, Y. (2022). A path towards autonomous machine intelligence, Version 0.9.2, 2022-06-27.

Levy, N., & Savulescu, J. (2009). Moral significance of phenomenal consciousness. *Progress in Brain Research, 177*, 361-370.

Li, B., Zhu, Z., Wang, X., Yuan, C., & Zhang, W. (2021). Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13018-13028).

Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., & Lin, L. (2024). Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Mai, X., Tao, Z., Lin, J., Wang, H., Chang, Y., Kang, Y., ... & Zhang, W. (2024). From Efficient Multimodal Models to World Models: A Survey. *arXiv preprint arXiv:2407.00118*.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard.

Metzinger, T. (2010). The self-model theory of subjectivity: A brief summary with examples. *Humana Mente-Quarterly Journal of Philosophy, 14*, 25-53.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120.

OpenAI. (2019). Emergent tool use from multi-agent interaction. Retrieved from `https://openai.com/blog/emergent-tool-use/`

OpenAI. (2019). Better models and more compute. Retrieved from `https://openai.com/research/`

Patel, K., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 939-952).

Paul, L. A., Ullman, T., De Freitas, J., & Tenenbaum, J. (2023). Reverse-engineering the self. *psyArXiv*.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Pfeiffer, C., Schmutz, V., & Blanke, O. (2014). Visuospatial viewpoint manipulation during full-body illusion modulates subjective first-person perspective. *Experimental Brain Research, 232*(12), 4021-4033.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565-573.

Seth, A. K. (2024). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1-42.

Shepherd, J. (2021). The moral status of conscious subjects. *Rethinking Moral Status*. OUP Oxford, 57-73.

Slater, M., Perez-Marcos, D., Ehrsson, H. H., & Sanchez-Vives, M. V. (2010). Towards a digital body: The virtual arm illusion. *Frontiers in Human Neuroscience, 4*, 6.

Tsakiris, M. (2010). My body in the brain: A neurocognitive model of body-ownership. *Neuropsychologia, 48*(3), 703-712.

Warren, M. A. (1997). *Moral Status: Obligations to Persons and Other Living Things*.

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences, 28*(5), 404-415.