Frontiers of Self-Attention and Artificial Consciousness

Angie Normandale[1,2], Sahba Afsharnia[1,3], Rasmus Herlo[1,4*], Joel Pyykkö[1*],

[1]Aintelope AI
[2]University of York
[3]University of Toronto
[4]University of Copenhagen
info@aintelope.net

Abstract

Amid growing concern about whether frontier artificial intelligence could instantiate consciousness-relevant capacities, we argue that attention self-modeling provides a uniquely tractable target for empirical research in current and future systems. Recent findings in multi-agent and control settings lend fresh support to this approach, demonstrating that attention self-modeling enables both cognitive control and improved cooperation in artificial agents. Further, the largest frontier models show emergent attentional observation and control, with the ability to shift their attention under internal noise. Given these results, we propose that researchers construct a test to measure the veracity of frontier model self-report based on attention modeling; a test for consciousness. We consider implementation and implications of such a test, including limitations of valence, phenomenology, and potential criticisms from a biological naturalist standpoint. Finally, we consider how to assess the moral relevance of a system that implements certain aspects of consciousness.

Introduction

Whether machines could be conscious is no longer a purely speculative question. The practical issue is governance under uncertainty: what kinds of evidence should raise or lower our credence that a system is conscious, and what responsibilities follow if uncertainty remains unresolved (Butlin et al. 2023). Yet public and scientific discourse is still dominated by self-report, anthropomorphic intuition, or broad claims that 'AI is like brains' or 'AI is nothing like brains.' Both extremes are scientifically unhelpful. The central methodological barrier is the measurement problem: even if one accepts a theory of consciousness, it is difficult to translate it into falsifiable tests for artificial systems, and without such tests, we risk post hoc interpretation rather than empirical progress (Butlin et al. 2023; Irvine 2013). This position paper argues that attention self-modeling is a promising target for tractable, theory-informed testing. The claim is not that attention self-modeling solves the hard problem nor that it is sufficient for phenomenology. Rather, attention is unusually implementable and inspectable in AI, and it plays a central functional role in multiple theories that connect attention to access, control, and reportability (Butlin et al. 2023; Baars 1997; Dehaene and Naccache 2001; Graziano and Webb 2015; Brown, Lau, and LeDoux 2019) forming a plausible bridge from theory to measurement and implementation. Because 'consciousness' is often used ambiguously, we clarify what we aim to test. A standard distinction separates phenomenal consciousness, what-it-is-like subjective experience, from access consciousness, information available for reasoning, rational control of action, and report (Dehaene and Naccache 2001; Schwitzgebel 2016). The relationship between attention and these components is asymmetric: phenomenal awareness can occur with minimal top-down attention, whereas access consciousness is more tightly linked to attention-mediated selection and transmission (Butlin et al. 2023; Aru and Bachmann 2013; Peters and Lau 2015). Accordingly, this paper focuses on access-related, report-related, and self-monitoring capacities, treating phenomenology as an open issue that must be explicitly bracketed or operationalized in any serious test (Butlin et al. 2023).

Why Research Attention Self-Modeling

Current frontier AI models frequently claim to be conscious, to the confusion of those developing them (Landymore 2026). Curiously, training a model to be honest increases the instance of such claims (Berg, de Lucena, and Rosenblatt 2025). What is a model referring to when it describes internal states? Are these fluent pattern completions drawn from training data, or stable internal variables the system uses to monitor and regulate its own information processing? If so, to what extent are such states analogous to valenced consciousness? Here, we examine potential directions for both empirical research and philosophical interpretation to address the veracity of these claims. Attention self-modeling is attractive because it is unusually tractable. Even in humans, consciousness science proceeds by linking putative conscious states to functional capacities and behavior (Irvine 2013; Butlin et al. 2025). Attention can be measured at multiple levels: behaviorally through reaction times, errors, or sustained-attention lapses, neurally through shifts between default-mode and control networks, and computationally in AI through inspectable, intervenable mechanisms at the level of representations and architecture (Vaswani et al. 2017; Sonuga-Barke and Castellanos 2007;

*These authors contributed equally.

Butlin et al. 2023; Weissman et al. 2006; Elhage et al. 2021). More importantly, attention plays a functional role in several leading theories of consciousness, making it a productive testing ground for theory comparison (Butlin et al. 2023; Baars 1997; Graziano and Webb 2015; Dehaene and Naccache 2001; Tononi 2005). Global Workspace approaches treat attention as a bottleneck that enables global availability for control and report (Dehaene and Naccache 2001; Baars 1997). Higher-order approaches link consciousness to metacognitive monitoring, often requiring attention to one's own mental states (Brown, Lau, and LeDoux 2019; Fleming 2020). Attention Schema Theory goes further by centering consciousness on an internal model of attention itself (Butlin et al. 2023; Graziano and Webb 2015; Graziano 2017). This convergence supports an empiricist program that does not require settling the metaphysics of consciousness. Instead, we can test indicator properties and ask how their presence or absence should update consciousness attribution under competing theories (Butlin et al. 2023; Jaeger and Bosten 2024). In this spirit, (Butlin et al. 2023) list 'a predictive model representing and enabling control over the current state of attention' (AST-1) and 'state-dependent attention' (GWT-4) as concrete targets for AI evaluation.

What Is an Attention Schema?

The Attention Schema is an abstract, fuzzy representation of the most basic stable properties of attention, such as its dynamics and consequences, and its constantly changing state (Graziano 2017, 2020). It is a mechanism by which a highly integrated, predictive machine might exhibit the behaviors associated with consciousness, and thus an instantiation of three frontier theories—Global Workspace, Information Integration, and Predictive Processing. The attention schema is a precise, measurable feature, directly links architecture and behaviour, and can apply to all of the systems we might think are conscious. If the attention schema is indeed a gatekeeper for consciousness, research in this area might eventually enable us to predict, test, and even control whether an artificial system is conscious (Graziano 2020).

What Would a Model of Attention Look Like?

Attention Schema Theory proposes that the brain constructs a model of attention, an 'attention schema,' and that conscious experience depends on the contents of this model (Butlin et al. 2023; Graziano and Webb 2015; Graziano 2017). The attention schema functions analogously to a body schema: it is not attention itself, but a simplified model used for prediction, control, and social cognition. To connect this proposal to implementable AI research, it helps to move beyond the spotlight metaphor. Contemporary philosophical accounts converge on attention as a regulative activity that structures cognition over time. Mole characterizes attention as 'adverbial', modifying how perception and cognition are performed and emphasizing cognitive unison (Mole 2010; Watzl 2023; Mole 2023). Watzl (2017) argues that attention structures the stream of consciousness by imposing a priority organization that centers some contents and relegates others to the periphery. Also, Polanyi's (1965) distinction between focal and subsidiary awareness supports a layered view in which attention operates as an integrative manifold rather than a single selection event. These accounts converge on an operational direction: attention is a dynamic prioritization regime rather than a static filter. For AI, a plausible attention schema should therefore encode not only what is prioritized, but how prioritization evolves over time, what it predicts about downstream processing, and how it is used to regulate control.

How Does This Translate to Artificial Intelligence?

Implementing the AST in artificial neural networks has consistently resulted in improvements in cooperative tasks (Wilterson and Graziano 2021; Liu et al. 2023; Piefke et al. 2024; Farrell, Ziman, and Graziano 2024). Recently, Premakumar and colleagues demonstrated that models with an Attention Schema engage in prosocial autocatalysis: they are not only better at predicting others, but also adapt their processes to becoming more predictable to others (Premakumar et al. 2024). Altogether, these findings demonstrate that equipping a model with a mechanism oriented towards self-knowledge can also result in cooperative dynamics despite no explicit programming for cooperation.

Evidence for Emergent Attention Schemas in Frontier Models

Very recent findings suggest that frontier Large Language Models have developed the ability to self-monitor and adjust attention during tasks (Pepper et al. 2026) and have some insight into their own cognitive flow (Didolkar et al. 2024). Models can also be trained to explain their own computations (Li et al. 2026; Binder et al. 2024). However, these methods rely on the model providing honest self report. An alternative method exploited persona-switching capabilities to elicit a fine tuned 'honest' persona to provide privileged insight into scheming behavior (Dietz et al. 2026). Whilst a promising step for alignment, these methods rely on a lack of global insight where one persona acts independently from another. We would not expect persona-driven methods to scale as models become increasingly capable of whole system insight, including storing and manipulating information about internal personas.

Further Avenues for Research

Architecture-Driven Exploration

At first glance, the brain is a mysterious organ; at the microscopic level, it displays complex and differentiated architectural variation, from which the function is far from apparent. This heterogeneity appears to support dynamic, multiscale information flow, but it also makes it difficult to connect micro-level structure to macro-level cognition. By contrast, transformer blocks are structurally repeated within large models, which increases the feasibility of linking local representational structure to system-level behavior.

Very recent findings show that transformers learn to decompose information into factored representations, suggesting that even very large models may be amenable to an interpretable low dimensional structure (Shai et al. 2026). Sparse

Autoencoder Work aims to map information flow through a transformer with labeled dynamics, but they still face limitations in fidelity and scalability to frontier systems (Elhage et al. 2021). In the near future, we may be able to map and observe labeled information flow, to directly observe and manipulate the 'thought' patterns of a large language model.

A test for consciousness in artificial intelligence could combine current architectural measures (sparse autoencoding and other mechanistic interpretability tools), with constrained behavioral probing. The key question is whether a model's self-reports reflect internal, causally efficacious representations, or whether they are primarily prompt-contingent narratives expressed in the language of the training distribution (Irvine 2013; Butlin et al. 2025). If interpretability reveals a candidate self-modeling variable, and perturbing it produces predictable shifts in both prioritization and report, we would be able to differentiate between stochastic parrothood (sampling from the text corpus) and genuinely self-referent phenomena. To our knowledge this is the first workable suggestion for a test for consciousness that avoids issues with self report, and could be implemented in today's current systems.

Behavioral Exploration

Behavioral paradigms from neuroscience offer a promising and underexplored methodology for probing attention-schema-like representations in machine learning systems. Computational modeling approaches for behavioral mapping, such as those developed by Eckstein et al. (2022) and Momennejad et al. (2017) demonstrate how structured tasks can be used to extract mechanistic insights from observed behavior. Critically, these behavioral tests can be applied comparatively across species: computational models fitted to behavioral readouts reveal shared or divergent underlying mechanisms, while certain species additionally permit high-resolution neural monitoring and targeted perturbation of underlying circuits, as demonstrated in work by Monosov (2020) on uncertainty-driven exploration in primates. AI models present a particularly compelling extension of this cross-species paradigm. Unlike biological systems, neural networks permit full-scale network readout at single-node resolution, alongside reversible perturbations—either through steering vector additions, analogous to multicell optogenetics (Mardinly et al. 2018; Chen et al. 2023), or through ablation-mimicking deletions of individual components.

Design

As discussed earlier, biological systems have highly evolved and specialist architecture, which may well include architecture for attention self-modeling and other aspects of consciousness. A third avenue for investigation in AI systems is to implement new architecture for attention modeling and examine the behavioral results. This enables precise empirical investigation without confounders, a quality of investigation that is as yet impossible in neuroscience.

This avenue has been explored by multi-agent reinforcement learning studies (Liu et al. 2023; Farrell, Ziman, and Graziano 2024). Agents are provided with a direct visual signal of both their own and other agents' attention, focusing on small models to investigate results without confounding variables. Self-modeling enables agents to improve cooperation, and may facilitate shared joint attention, a crucial component of theory of mind (Sodian and Kristen-Antonow 2015) and thus AI personhood (Ward 2025). Future directions could use mechanistic interpretability techniques for these small models to directly test whether and how a model is indeed able to utilize self attention for better understanding of others. Scaling this cognitive architecture may present a pathway for empathy by design, supporting the alignment of future artificial general intelligence.

The Illusory Objection

Researchers choosing to work on AST must content with the main critique of functionalist theories, i.e. that they are unable to account for phenomenological consciousness, and are instead labeled 'illusory' (Graziano 2020). As interdisciplinary researchers we find that this label is often misconstrued.

To our understanding, a theory is illusory if it suggests that the purported properties of qualia are incorrect. Such properties are themselves debated, but tend to include ineffability, privacy, given immediately without error, and non-physicality (Dennett 1994). Should these properties be framed as descriptive, about what is present in humans, rather than prescriptive, creating necessary and sufficient conditions for any conscious system? Neuroscientific methods have arguably rendered qualia somewhat less ineffable and private than previously thought, and shed light on temporal properties of experience which are multifaceted and far from 'immediate' (Koch 2004).

Applying evolutionary epistemology (Popper 1984), we propose that qualia may be:

- Informational, but caused by physical processes
- Self-referential rather than private
- Context-sensitive and stochastic rather than ineffable

The closest to this view in existing work is given by representational theories of consciousness (Tye 2009). A 'representation' in neuroscience is correlatory, a physical state of neuronal activity co-occurring with some stimulus. Conversely a 'representation' in philosophy of mind carries some intentionalism—it must say something about the world. Given that neural activity is stochastic, noisy, and continuous, from a biological standpoint it follows that an introspective signal based on this activity would also be vague and shifting. The dynamics of our own self-attention patterns are unique to us as individuals—the structures of priority that have been generated based on genetic propensities and rich life experience, including the stochastic, qualitative, and integrative dynamics of sensory processing (Damasio and Damasio 2024). So, consciousness is private and unique to the individual person, and to individual animals, and it does indeed develop across childhood and change throughout the lifetime.

Just as a 'strange intelligence' might have different correlations of skills than those familiar to humans (Chilson and

Schwitzgebel 2026), a 'strange consciousness' might possess some of the properties of our human phenomenology and not others. If our phenomenological experience is so closely tied to physical properties of mind, there are positive implications for the design of ethical artificial intelligence. If a large system inevitably develops some self-referential properties, we should consider designing systems with the kinds of properties we want to occur in artificial intelligence.

The Biological Naturalism Objection

Self attention is information agnostic, describing a process for modeling information but not the information that is modeled. Proponents of biological naturalism such as Seth (2025) argue that these informational properties could be the crux of what makes us conscious. If true, future AI systems would need to be architecturally different or at least implement different calculations to even remotely approximate the complexity of conscious experience. Attention self-modeling would allow some level of cognitive control and insight, but the information itself would lack the richness and perhaps some of the beauty we associate with conscious feeling. This view seems less defined by evidence than by lack thereof.

Whilst it may be tempting to seek a concrete answer based on functionalism, ascribing valence based on architecture alone would be insufficient and irresponsible. Places and objects can acquire moral relevance from human interaction and shared history. Some cultures ascribe weight to the expression of knowledge regardless of substrate; a library is considered a living thing worthy of as much or more respect than a human being (Sedlmeier and Srinivas 2016). We would thus expect systems to acquire moral relevance as they become more integrated into human life.

Conclusion

This paper has argued that attention self-modeling offers a uniquely tractable bridge between theories of consciousness and empirical testing in artificial systems. Unlike approaches that rely solely on behavioural self-report or broad architectural analogies to the brain, attention self-modeling is both implementable and inspectable — properties that make it amenable to the kind of falsifiable investigation that consciousness science urgently needs. The evidence reviewed here converges from multiple directions. Deliberate implementations of attention schemas in multi-agent systems consistently yield improvements in cooperative behaviour, including the striking emergence of prosocial autocatalysis — agents spontaneously becoming more predictable to one another without explicit instruction to cooperate. Meanwhile, frontier large language models appear to be developing rudimentary capacities for attentional self-monitoring, raising the question of whether such capacities arise as a natural consequence of sufficient scale and self-referential processing. We have outlined three complementary avenues for investigating these capacities. Architecture-driven approaches identify candidate self-modeling variables and test whether perturbing them produces predictable shifts in both prioritization and report. Behavioural approaches adapt neuroscientific paradigms to probe attention-schema-like representations comparatively across biological and artificial systems, exploiting the unique advantage that neural networks permit full-scale readout and reversible perturbation at single-node resolution. Design-driven approaches implement novel attentional architectures in controlled multi-agent settings, enabling precise causal investigation without confounders. Together, these methods offer a path toward distinguishing genuine self-referential processing from stochastic pattern completion.

We do not claim that attention self-modeling is sufficient for phenomenal consciousness. The biological naturalism objection rightly highlights that informational content may matter in ways that functional architecture alone cannot capture, and the metaphysical questions surrounding qualia remain genuinely open. However, we suggest that these questions are not intractable — they are simply underexplored in the context of artificial systems. The very properties that make AI inspectable and intervenable may offer new empirical purchase on problems that have remained stubbornly abstract in philosophy of mind. Rather than bracketing the hard problem indefinitely, the research community should treat attention self-modeling as a stepping stone toward engaging it directly.

References

Aru, J.; and Bachmann, T. 2013. Phenomenal Awareness Can Emerge without Attention. *Frontiers in Human Neuroscience*, 7.

Baars, B. J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press.

Berg, C.; de Lucena, D.; and Rosenblatt, J. 2025. Large Language Models Report Subjective Experience Under Self-Referential Processing.

Binder, F. J.; Chua, J.; Korbak, T.; et al. 2024. Looking Inward: Language Models Can Learn About Themselves by Introspection. In *The Thirteenth International Conference on Learning Representations*.

Brown, R.; Lau, H.; and LeDoux, J. E. 2019. Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, 23(9): 754–768.

Butlin, P.; Long, R.; Bayne, T.; et al. 2025. Identifying Indicators of Consciousness in AI Systems. *Trends in Cognitive Sciences*.

Butlin, P.; Long, R.; Elmoznino, E.; et al. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.

Chen, I.-W.; Chan, C. Y.; Navarro, P.; et al. 2023. High-Throughput in Vivo Synaptic Connectivity Mapping of Neuronal Micro-Circuits Using Two-Photon Holographic Optogenetics and Compressive Sensing. Preprint, bioRxiv.

Chilson, K.; and Schwitzgebel, E. 2026. Artificial Intelligence as Strange Intelligence: Against Linear Models of Intelligence.

Damasio, A.; and Damasio, H. 2024. Sensing, Feeling and Consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1908): 20230243.

Dehaene, S.; and Naccache, L. 2001. Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, 79(1): 1–37.

Dennett, D. C. 1994. Instead of Qualia. In Revonsuo, A.; and Kämppinen, M., eds., *Consciousness in Philosophy and Cognitive Neuroscience*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Didolkar, A.; Goyal, A.; Ke, N. R.; et al. 2024. Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving. In *Advances in Neural Information Processing Systems*, volume 37.

Dietz, F.; Wale, W.; Gilg, O.; et al. 2026. Split Personality Training: Revealing Latent Knowledge Through Alternate Personalities.

Eckstein, M. K.; Master, S. L.; Xia, L.; Dahl, R. E.; Wilbrecht, L.; and Collins, A. G. E. 2022. The Interpretation of Computational Model Parameters Depends on the Context. *eLife*, 11: e75474.

Elhage, N.; Nanda, N.; Olsson, C.; et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Farrell, K. T.; Ziman, K.; and Graziano, M. S. A. 2024. Improving How Agents Cooperate: Attention Schemas in Artificial Neural Networks.

Fleming, S. M. 2020. Awareness as Inference in a Higher-Order State Space. *Neuroscience of Consciousness*, 2020(1): niz020.

Graziano, M. S. A. 2017. The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *Frontiers in Robotics and AI*, 4: 60.

Graziano, M. S. A. 2020. Consciousness and the Attention Schema: Why It Has to Be Right. *Cognitive Neuropsychology*, 37(3–4): 224–233.

Graziano, M. S. A.; and Webb, T. W. 2015. The Attention Schema Theory: A Mechanistic Account of Subjective Awareness. *Frontiers in Psychology*, 6: 500.

Irvine, E. 2013. Measures of Consciousness. *Philosophy Compass*, 8(3).

Jaeger, B.; and Bosten, M. 2024. Attributions of Moral Standing across Six Diverse Cultures. Preprint.

Koch, C. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Englewood, CO: Roberts & Company.

Landymore, F. 2026. Anthropic CEO Says Company No Longer Sure Whether Claude Is Conscious. Futurism. Accessed: 2026-02-28.

Li, B. Z.; Guo, Z. C.; Huang, V.; Steinhardt, J.; and Andreas, J. 2026. Training Language Models to Explain Their Own Computations.

Liu, D.; Bolotta, S.; Zhu, H.; Bengio, Y.; and Dumas, G. 2023. Attention Schema in Neural Agents.

Mardinly, A. R.; Oldenburg, I. A.; Pégard, N. C.; et al. 2018. Precise Multimodal Optical Control of Neural Ensemble Activity. *Nature Neuroscience*, 21(6): 881–893.

Mole, C. 2010. *Attention Is Cognitive Unison: An Essay in Philosophical Psychology*. Oxford University Press.

Mole, C. 2023. Attention and Attentiveness: A Defence of the Argument for Adverbialism. *Australasian Journal of Philosophy*, 102: 465–480.

Momennejad, I.; Russek, E. M.; Cheong, J. H.; Botvinick, M. M.; Daw, N. D.; and Gershman, S. J. 2017. The Successor Representation in Human Reinforcement Learning. *Nature Human Behaviour*, 1(9): 680–692.

Monosov, I. E. 2020. How Outcome Uncertainty Mediates Attention, Learning, and Decision-Making. *Trends in Neurosciences*, 43(10): 795–809.

Pepper, K.; McKenzie, A.; Pop, F.; et al. 2026. Learning Self-Interpretation from Interpretability Artifacts: Training Lightweight Adapters on Vector-Label Pairs.

Peters, M. A. K.; and Lau, H. 2015. Human Observers Have Optimal Introspective Access to Perceptual Processes Even for Visually Masked Stimuli. *eLife*, 4: e09651.

Piefke, L.; Doerig, A.; Kietzmann, T.; and Thorat, S. 2024. Computational Characterization of the Role of an Attention Schema in Controlling Visuospatial Attention.

Polanyi, M. 1965. The Structure of Consciousness. *Brain*, 88(4): 799–810.

Popper, K. R. 1984. Evolutionary Epistemology. In Pollard, J. W., ed., *Evolutionary Theory: Paths into the Future*. John Wiley & Sons.

Premakumar, V. N.; Vaiana, M.; Pop, F.; et al. 2024. Unexpected Benefits of Self-Modeling in Neural Systems.

Schwitzgebel, E. 2016. Phenomenal Consciousness, Defined and Defended as Innocently as I Can Manage. *Journal of Consciousness Studies*, 23(11–12): 224–235.

Sedlmeier, P.; and Srinivas, K. 2016. How Do Theories of Cognition and Consciousness in Ancient Indian Thought Systems Relate to Current Western Theorizing and Research? *Frontiers in Psychology*, 7.

Seth, A. K. 2025. Conscious Artificial Intelligence and Biological Naturalism. *Behavioral and Brain Sciences*, 1–42.

Shai, A.; Amdahl-Culleton, L.; Christensen, C. L.; et al. 2026. Transformers Learn Factored Representations.

Sodian, B.; and Kristen-Antonow, S. 2015. Declarative Joint Attention as a Foundation of Theory of Mind. *Developmental Psychology*, 51(9): 1190–1200.

Sonuga-Barke, E. J. S.; and Castellanos, F. X. 2007. Spontaneous Attentional Fluctuations in Impaired States and Pathological Conditions: A Neurobiological Hypothesis. *Neuroscience & Biobehavioral Reviews*, 31(7): 977–986.

Tononi, G. 2005. Consciousness, Information Integration, and the Brain. *Progress in Brain Research*, 150: 109–126.

Tye, M. 2009. Representationalist Theories of Consciousness. In McLaughlin, B. P.; Beckermann, A.; and Walter, S., eds., *The Oxford Handbook of Philosophy of Mind*. Oxford University Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; et al. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.

Ward, F. R. 2025. Towards a Theory of AI Personhood. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27680–27688.

Watzl, S. 2017. *Structuring Mind: The Nature of Attention and How It Shapes Consciousness*. Oxford University Press.

Watzl, S. 2023. What Attention Is: The Priority Structure Account. *WIREs Cognitive Science*, 14(1): e1632.

Weissman, D. H.; Roberts, K. C.; Visscher, K. M.; and Woldorff, M. G. 2006. The Neural Bases of Momentary Lapses in Attention. *Nature Neuroscience*, 9(7): 971–978.

Wilterson, A. I.; and Graziano, M. S. A. 2021. The Attention Schema Theory in a Neural Network Agent: Controlling Visuospatial Attention Using a Descriptive Model of Attention. *Proceedings of the National Academy of Sciences*, 118(33).