Integrated World Modeling Theory (IWMT) and the Human Consciousness Hypothesis (HCH)

Adam Safron[1,3], **Victoria Klimaj** [2], **Zahra Sheikhbahaee**[3]
[1]Allen Discovery Center, Tufts University
[2]Indiana University
[3]Institute for Advanced Consciousness Studies
Adam.Safron@tufts.edu, Victoria.Klimaj@gmail.com, Sheikhbahaee@gmail.com

Abstract

Here, we explore points of convergence between the Human Consciousness Hypothesis (HCH) and Integrated World Modeling Theory (IWMT). HCH posits that consciousness is defined by three fundamental principles: Genesis (conscious functions constituting an early-stage learning algorithm), Coherence (maximization of representational consistency), and Second-Order Perception (synchronous meta-awareness of perceptual processes). IWMT serves as a unifying model that reconciles major theories of consciousness with a particular focus on Global Neuronal Workspace Theory, Integrated Information Theory, and the Free Energy Principle and Active Inference framework. Central to IWMT is the proposal that phenomenal consciousness is "what it feels like" to be the spatiotemporally and causally coherent functioning of a probabilistic generative world model for an embodied agent. Mechanistically, IWMT identifies "self-organizing harmonic modes" (SOHMs) as synchronous neural complexes implementing iterative Bayesian inference to generate consciousness as maximum a posteriori estimates of embodied sensorium states. Nested heterarchies of SOHMs are proposed as biophysical substrates for consciousness, acting as dynamic cores of integrated information that facilitate the synchronous combination of multimodal sense data into a unified field of experience to promote intelligent/adaptive (active) inference and learning. Critically, IWMT requires (body-)world models to be capable of both informing and being informed by action-perception cycles at behaviorally relevant timescales. This architecture suggests consciousness could potentially be realized in artificial systems with appropriate recurrent dynamics and sufficient degrees (and kinds) of embodied grounding.

Introduction

The California Institute for Machine Consciousness (CIMC) is a new research initiative dedicated to understanding what it might take to recapitulate the conscious functions of biological intelligences in artificial systems. Towards this end, the CIMC has introduced the Human Consciousness Hypothesis (HCH), which defines conscious systems in terms of three fundamental principles: 1) Genesis: Consciousness emerges from self-organizing systems as an early-stage learning algorithm and prerequisite for complex intelligence; 2) Coherence: Consciousness functions as a coherence-maximizing process (or "cortical conductor") that minimizes constraint violations across mental representations by orchestrating parallel processing and directing attention to conflicts/inconsistencies; 3) Second-order perception: Consciousness implements perception of perceptual processes, with this meta-awareness happening in synchrony with the registration of perceptual content. Here, we will summarize a recently introduced synthetic framework for consciousness that is highly consonant with the HCH, and which may provide useful direction in working towards realizing CIMC's goal of creating conscious machines: Integrated World Modeling Theory (IWMT) (Safron 2020b,a, 2022, 2023).

Integrated World Modeling Theory (IWMT)

Integrating Across Theories of Consciousness

IWMT is a unifying model of phenomenal consciousness and conscious access (Wiese 2020), explained across computational/functional, algorithmic, and implementational/mechanistic (supervenient) levels of analysis (Marr 1983). IWMT primarily focuses on Global Neuronal Workspace Theory (GNWT) (Dehaene 2014), Integrated Information Theory (IIT) (Tononi 2015), and the Free Energy Principle and Active Inference framework (Friston 2010; Friston et al. 2017), but is also compatible with other computational and biophysical models of consciousness such as Recurrent Processing (Grossberg 2017), Predictive Processing (Hohwy and Seth 2020), Dynamic Core (Edelman, Gally, and Baars 2011), Temporo-Spatial Sentience (Northoff et al. 2023), Higher Order Thought (Brown, Lau, and LeDoux 2019), and Attention Schema (Graziano 2013, 2019) theories. While not necessarily endorsing all claims made by these various theories, IWMT attempts to show how multiple (but not all) perspectives on consciousness may be coherently brought together for synergistic understanding. The overarching goal of IWMT is to identify precise physical and computational substrates of consciousness, so that they may be more skillfully understood and intervened upon, and potentially developed by artificial means.

Consciousness as the Coherent Modeling of Self and World

IWMT suggests that consciousness can be functionally understood as a kind of coherently integrated (self-)world modeling. World models are internal representations that a cognizing system constructs of its environment (including itself) that allow it to predict (or simulate) likely states based on either sensed or imagined observations and actions. These models expand intelligent and adaptive functioning by allowing systems to infer dynamics of the world (and self) beyond immediately-observable sense data, so enabling capacities for creative insight, planning, and causal reasoning based on counterfactual mental simulations (Pearl and Mackenzie 2018). World models afford forms of cognitive sophistication that Dennett (1996) described with respect to "Popperian creatures" that can select which actions are most likely to produce value (or avoid harm) based on imagined consequences, and so can let their "hypotheses die in their stead" (Klimaj and Safron 2025). From this point of view, consciousness has a clear adaptive significance by enhancing intelligence via coherent modeling of system-world relations.

World modeling is also increasingly suggested to be key to solving the enduring problems of AI, such as in proposals for designing systems centered on spatial intelligence (Li, 2025), or with Joint Embedding Predictive Architectures (i.e., JEPA) (Assran et al. 2023, 2025). Some have suggested that foundation models may also acquire such world modeling capacities in their latent spaces when trained on vast quantities of training data (Gwern 2020). However, it is a matter of fierce debate whether such systems have representations with sufficient degrees of coherent integration that they can engage in world modeling (or "sentience") of similar kinds as underlying the remarkable abilities of animal minds to efficiently learn and flexibly adapt to novel situations (Safron et al., in press). For reasons described in more detail below, IWMT suggests that without sufficiently rich physical or virtual embodiments, systems such as large language models are not likely to be able to generate the kinds of world modeling underlying robust general(izing) intelligence, nor consciousness (Zahavy 2026).

According to IWMT, phenomenal consciousness is "what it is like" (Nagel 1974) to be the spatiotemporally and causally coherent functioning of a probabilistic generative model for the sensorium of an embodied agent, entailing either perception or imagination when respectively coupled to or decoupled from sense-data. This synthetic framework draws upon multiple theories to describe how subnetworks of the brain can act as shared latent (work)spaces amongst heterogeneous sensory modalities and arenas for Bayesian model selection (Jaegle et al. 2021, 2022; Juliani, Kanai, and Sasai 2022), generating system-world estimates with sufficient rapidity to both inform and be informed by action-perception cycles on the timescales of their formation, so affording more flexibly intelligent and adaptive behavior. IWMT is distinct from the models it attempts to integrate in suggesting that no pre-existing theory taken in isolation identifies sufficient conditions for explaining why there should be "something that it feels like" to be a physical/computational system.

Spatiotemporal and Causal Coherence and Minimal Embodied Selfhood as Preconditions for Consciousness

IWMT argues that phenomenal consciousness emerges from processes capable of generating integrated models of system and world with coherence with respect to space (as organized locality), time (as proportional changes in space), and cause (as predictable/modellable regularities in these changes). These coherence-making properties are inspired by the preconditions for judgment suggested by Immanuel Kant (1781) (i.e., "synthetic a priori categories"), with spatiotemporal and causal coherence (and potentially integration into unified self-models [c.f. transcendental unity of apperception]) also being necessary for there to be coherently modeled compositional properties (Greff, van Steenkiste, and Schmidhuber 2020), and so any kind of subjective experience whatsoever (De Kock 2016; Northoff 2012). [Please note: while some degree of basic self-modeling and agency may be required for coherent subjectivity, this selfhood may potentially be extremely minimal (Metzinger 2020), and indeed cannot be overly complex if it is to serve the function of a foundational inductive bias that emerges early in development (i.e., the Genesis principle).]

IWMT is distinct from the theories it draws upon in that it is fundamentally cybernetic and embodied in its approach. That is, IWMT suggests that consciousness is "what it feels like" to generate coherent system-world estimates with sufficient rapidity that they can both inform and be informed by action-perception cycles on the timescales over which they evolve as embodied agents interact with their environments. While inspired by IIT as a means of identifying biophysical systems that are more or less likely to be capable of entailing consciousness, IWMT suggests that systems could generate (or be) arbitrarily large amounts of integrated information and still not be conscious without embodied grounding. While inspired by GNWT with respect to the computational principles underlying consciousness, IWMT suggests that systems could instantiate functional global workspaces without entailing any kind of subjective experience (unless workspace dynamics also entail spatiotemporally- and causally-coherent world modeling with embodied grounding). IWMT shares the enactive approach of Thompson and Varela (2001) in insisting that consciousness is inseparable from embodied cybernetic action–perception cycles, but it diverges from strong enactivism in retaining brain-internal representations and (probabilistic generative world) models as essential explanatory constructs (Safron 2021a; Safron, Hipólito, and Clark 2023).

Learning to Model/Infer the World for the Sake of Adaptive Action

What are the necessary and sufficient conditions for coherently generating a stream of experience, organized according to coherently structured perspectival reference frames? In brief, IWMT views consciousness as a kind of mathemat-

ical object (Tegmark 2015), and in particular an iteratively computed maximum a posteriori estimate—or compressed representation (Safron 2023)—of likely system-world relations. More informally, phenomenal consciousness is understood as a kind of computational object akin to a "deep fake" created by generative AI, but rather than pixel arrays, the information generated is iteratively inferred sensorium states for embodied organisms with various combinations of modal features (e.g., sight, touch, sound, interoception). By focusing on the generation of likely patterns of sense data that correspond to all the various aspects/qualities of embodied experience—including the interoceptive inferences contributing to emotional/affective feelings (Safron 2021a,b; Seth and Friston 2016)—IWMT may help answer the (often begged) question as to why computational processes would entail consciousness, rather than happening "in the dark."

Functionally speaking, phenomenal consciousness is proposed to constitute a kind of evolutionary adaptation and data structure that combines different sensory modalities into an iteratively-estimated unified field of experience. It would be highly adaptive—in terms of more intelligent action selection and learning—for an agent to have a spatiotemporally and causally coherent model of self and world, organized by egocentric perspective, with organism-salient/relevant features being given greater attention based on histories of (valenced) experience. Functionally speaking, conscious simulations of body-world relations could be thought of as "digital twins," or virtual representations/simulations of physical systems that generate/predict likely states for the sake of optimizing operations without requiring costly and potentially risky interactions with the real world. The generation of coherently organized streams of experience further promotes various forms of conscious access and metacognitive reflection, so enhancing capacities for both reasoning, episodic memory, and lifelong (meta-)learning (c.f. Second-order perception and Genesis principles). This developmental emphasis is also consistent with learning-based accounts in which conscious access is progressively acquired through plasticity and the increasing higher-order knowledge of an agent's own internal states (Cleeremans 2011).

Physical and Computational Substrates of Consciousness and Conscious Access

IWMT identifies points of convergence between IIT and GNWT and attempts to reconcile conflicting claims regarding physical substrates of consciousness (Ferrante et al. 2025). Mechanistically, IWMT introduces "self-organizing harmonic modes" (SOHMs) as synchronous complexes of neural activity that emerge as metastable attractors within brain networks—c.f. "communication through coherence" (Deco and Kringelbach 2016; Fries 2015)—functioning as dynamic cores of integrated information and workspaces. Streams of experience emerge as an evolving generation of sensorimotor predictions, with the precise composition of conscious contents depending on the extent to which patterns of effective connectivity from various modalities couple with integrative-workspace-enabling SOHMs on the timescales of their formation. These integrating dynamics are suggested to be particularly likely to occur via richly connected subnetworks that afford body-centric sources of phenomenal binding and executive control. Along these connectivity backbones (for workspace dynamics), SOHMs are proposed to implement turbo coding via loopy message-passing over predictive (autoencoding) networks (McEliece, MacKay, and Cheng 1998), thus generating moments of consciousness as maximum a posteriori estimates of likely system-world configurations (Figure 1).

These subjective experiences are also suggested to be causally efficacious with respect to agency—c.f. ideomotor theory (Safron 2021a; Shin, Proctor, and Capaldi 2010)—functioning as sources of control energy and enslaving order parameters governing neural evolution (Haken 2007), with alpha frequencies generating multimodal binding into coherent sensorium-states by posterior cortices, and cross-frequency phase-coupling within theta frequencies affording involvement of frontal cortices—as well as the hippocampal/entorhinal system—potentially affording various forms of higher-order processing, conscious access, and volitional control (Safron, Çatal, and Verbelen 2022; Safron 2021a; Gershman 2019). These dynamic cores of integrated information also function as global workspaces, centered on posterior cortices, but capable of being entrained by frontal cortices and interoceptive hierarchies (Domenech and Koechlin 2015; Seth, Suzuki, and Critchley 2012), thus affording (varying degrees of) agentic causation. Therefore, with respect to the inconclusive adversarial collaboration between IIT and GNWT, both theories are proposed to be correct in their claims about whether consciousness is realized by either a "posterior hot zone" (as SOHMs over occipitotemporal and parietal cortices forming at alpha frequencies) or by a broader functional network involving the frontal lobes (as larger and more slowly evolving SOHMs potentially requiring synchronization at theta frequencies). In these ways, IIT and GNWT are suggested to make differing and potentially complementary claims about the systemic properties required to generate different aspects of experience and awareness (i.e., phenomenal consciousness and conscious access).

Implications for Machine Consciousness and Future Directions for IWMT

In proposing correspondences between biophysical mechanisms and machine learning algorithms/architectures (Figure 1), IWMT suggests consciousness could potentially be realized in non-biological systems. Speculatively, if these functional mappings were realized in the form of human-mimetic, neuromorphic AI, then they may allow for both flexible general intelligence and conscious experience. Indeed, a fully developed theory of consciousness and its algorithmic- and implementational-level realizations could be considered to be 'pseudocode' for (partially human-interpretable) artificial general intelligence with "System 2" capacities (Bengio 2019; Kahneman 2011), and possibly also phenomenal consciousness.

In contrast to assertions that a computer simulation of water cannot entail real "wetness" (Koch, 2019; Seth, 2026),

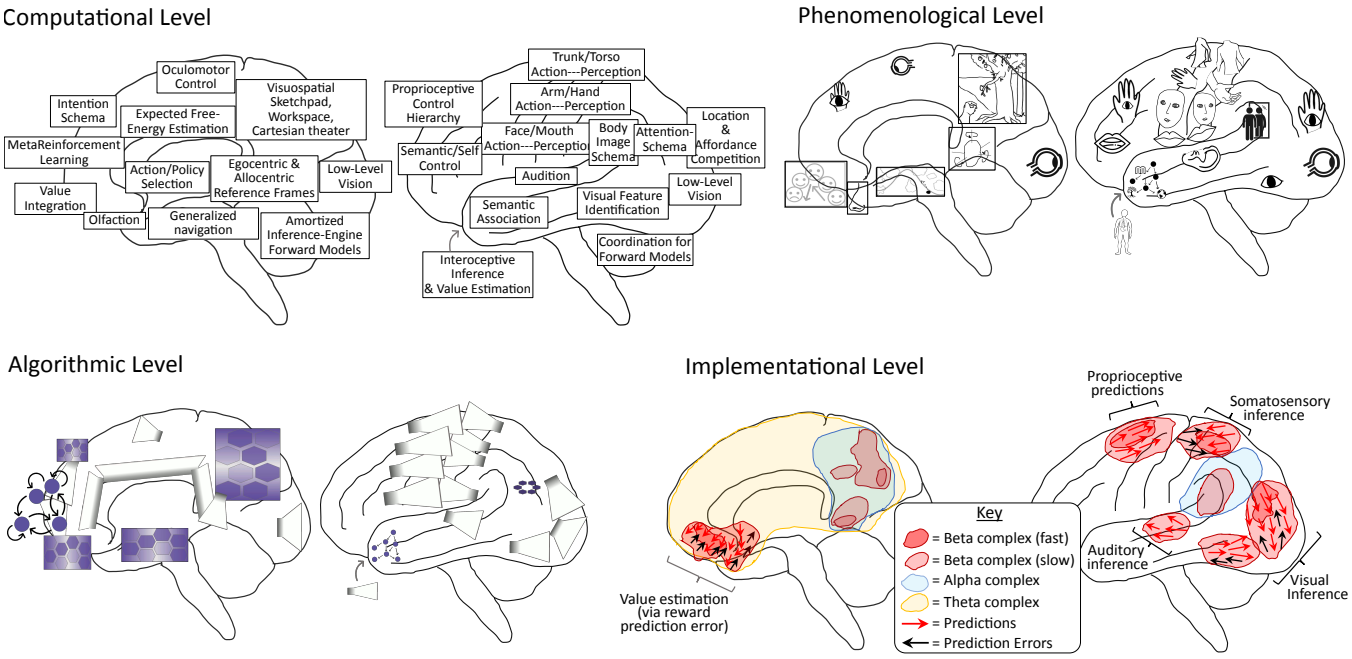

Figure 1: For a more detailed description, please see (Safron 2021b,a). Depiction of the human brain in terms of entailed aspects of experience (i.e., phenomenology), as well as computational (or functional), algorithmic, and implementational levels of analysis (Marr 1983). A phenomenological level is specified to provide mappings between consciousness and these complementary/supervenient levels of analysis. Modal depictions connote the radically embodied nature of mind, but not all images are meant to indicate conscious experiences. On the algorithmic level, these functions are mapped onto variants of machine learning architectures—e.g., autoencoders and generative adversarial networks, graph neural networks, recurrent reservoirs and liquid state machines—organized according to potential realization by neural systems. While the modal character of prefrontal computation is depicted at the phenomenological level of analysis, IWMT proposes frontal cortices might only indirectly contribute to consciousness via influencing dynamics in posterior cortices. On the implementational level, biological realizations of algorithmic processes are depicted as corresponding to flows of activity and interactions between neuronal populations, canalized by the formation of metastable synchronous complexes (i.e., "self-organizing harmonic modes").

processes involving virtual machines may be just as real/-causal as physical mechanisms, and potentially even more substantial (Albantakis 2017; Hoel et al. 2016; Hoel 2025). That is, a simulation capturing the emergent properties of interacting water molecules would recapitulate the phenomenology of wetness within the simulated world, and such properties could even be experienced externally if sufficient information can be transmitted to subjects outside the simulation (e.g. via fully immersive multimodal virtual reality (Chalmers 2022)). Even if one grants that simulated water cannot produce experienceable wetness, this is precisely the same situation we face with respect to consciousness. That is, an individual's consciousness cannot directly make someone else feel what that being is experiencing unless there are sufficient efficient causal processes for transmitting the necessary information to generate such experiences (e.g. via detailed linguistic descriptions).

However, the realization of "artificial" consciousness may depend upon particular implementational details, and may practically (and perhaps also in principle) require the parallelism and physical recurrence of neuromorphic and/or analogue computing systems. As described in greater detail elsewhere (Safron 2022), while feedforward systems can theoretically mimic the outputs of recurrent architectures through "unrolling" (Albantakis 2017; Albantakis et al. 2017), non-recurrent networks (such as transformer-based large language models) lack the functional closure, robustness, and efficiency required for realizing the computational properties that generate consciousness. According to IWMT, recurrent processes can afford "dynamic cores" of effective connectivity that facilitate flexible binding and iterative Bayesian model selection of likely system-world states, given past experiences and present/hypothetical observations/actions. A feedforward system could theoretically instantiate workspace functions (e.g. "broadcasting") (Doerig et al. 2019), but this would be divorced from the environmental couplings and historicity by which adaptive predictive (world-)models are learned.

IWMT has been further developed to inform models of intentional goal-oriented behavior (as enacted imaginings (Safron 2021a)), high-level control for robotic systems (as generalized navigation (Safron, Çatal, and Verbelen 2022)), and to explain a diverse range of psychedelic phenomena (as kinds of waking dreams (Safron et al. 2025)). Future directions for advancing the theory include updating the algorithmic mapping of biophysical processes to modern machine learning architectures—as imperfect-but-useful models of computational/functional properties—such as transformers and recurrent state space models (Chen et al. 2024; Zhu et al. 2025), as well as methods from physics-informed machine learning (Champion et al. 2019; Hirsh et al. 2021), and geometric deep learning (Bronstein, Bruna, and Cohen 2021; Greff, van Steenkiste, and Schmidhuber 2020; Yao et al. 2025). If consciousness is potentially computable in these ways, then advances in hardware and algorithms may eventually allow us to test IWMT by examining whether various conscious functions are realized when we construct systems consistent with the theory. For example, future neuromorphic systems could provide opportunities to test whether binding via recurrent workspaces facilitates multimodal data fusion and enhances adaptive/intelligent control-of and learning-from action-perception cycles for embodied agents.

Conclusion: IWMT, HCH, and the Future of Consciousness Studies

Thus, there appears to be substantial convergence between IWMT and the HCH. IWMT supports the "Genesis" and "Coherence" principles in that while conscious functions may support high-level reasoning, consciousness is suggested to emerge early in development from self-organizing learning processes once modeling capacities achieve sufficient degrees of (spatiotemporal and causal) coherence, and also provides a means of enhancing the coherence of subsequent inference and learning. IWMT also supports the "Second-order perception" principle in that agent-centered deep temporal models are realized via synchronous recurrent dynamics such that high-level beliefs contextualize and stabilize low-level perceptual processes as they co-evolve. These hierarchical priors take a variety of forms, including body and attention schemas wherein covert/mental actions—c.f. "Premotor" and "Biased Competition" theories (Desimone 1998; Rizzolatti and Craighero 2010; Safron 2021a)—allow for coherence-maximization via the top-down (potentially agentic) control of attentional selection. In working towards "IWMT 2.0," it will undoubtedly be fruitful to more deeply explore points of conceptual intersection with the HCH, as well as the broader idea(/value) community growing around the California Institute for Machine Consciousness.

Acknowledgments

Adam Safron would like to thank the Survival and Flourishing Fund in providing support for his position working with Michael Levin (to whom he is also immensely grateful) at the Allen Discovery Center. We would also like to thank Zan Huang for helpful feedback on earlier versions of this manuscript, as well as Karl Friston for the inspiration and generous guidance he provided as these ideas were being developed.

References

Albantakis, L. 2017. A Tale of Two Animats: What does it take to have goals. *arXiv preprint arXiv:1705.10854*.

Albantakis, L.; Marshall, W.; Hoel, E.; and Tononi, G. 2017. What caused what? A quantitative account of actual causation using dynamical causal networks. *arXiv preprint arXiv:1708.06716*.

Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Mojtaba, K.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; Arnaud, S.; Gejji, A.; Martin, A.; Hogan, F. R.; Dugas, D.; Bojanowski, P.; Khalidov, Y., V. LeCun; Rabbat, M.; and Ballas, N. 2025. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985*.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Bengio, Y. 2019. The Consciousness Prior. *arXiv preprint arXiv:1709.08568*.

Bronstein, M. M.; Bruna, J.; and Cohen, P., T. ad Veličković. 2021. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*.

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

Chalmers, D. J. 2022. *Reality+: Virtual Worlds and the Problems of Philosophy*. Penguin Books Limited.

Champion, K.; Lusch, B.; Kutz, J. N.; and Brunton, S. L. 2019. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45): 22445–22451.

Chen, C.; Wu, Y.-F.; Yoon, J.; and Ahn, S. 2024. TransDreamer: Reinforcement Learning with Transformer World Models. *arXiv preprint arXiv:2202.09481*.

Cleeremans, A. 2011. The radical plasticity thesis: how the brain learns to be conscious. *Frontiers in Psychology*, 2: 86.

De Kock, L. 2016. Helmholtz's Kant revisited (Once more). The all-pervasive nature of Helmholtz's struggle with Kant's Anschauung. *Studies in History and Philosophy of Science*, 56: 20–32.

Deco, G.; and Kringelbach, M. L. 2016. Metastability and Coherence: Extending the Communication through Coherence Hypothesis Using A Whole-Brain Computational Perspective. *Trends in Neurosciences*, 39(3): 125–135.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.

Dennett, D. 1996. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.

Desimone, R. 1998. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 353(1373): 1245–1255.

Doerig, A.; Schurger, A.; Hess, K.; and Herzog, M. H. 2019. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, 72: 49–59. https://doi.org/10.1016/j.concog.2019.04.002.

Domenech, P.; and Koechlin, E. 2015. Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, 1: 101–106. https://doi.org/10.1016/j.cobeha.2014.10.007.

Edelman, G.; Gally, J. A.; and Baars, B. J. 2011. Biology of consciousness. *Frontiers in Psychology*, 2: 4.

Ferrante, O.; Gorska-Klimowska, U.; Henin, S.; Hirschhorn, R.; Khalaf, A.; Lepauvre, A.; Liu, L.; Richter, D.; Vidal, Y.; Bonacchi, N.; Brown, T.; Sripad, P.; Armendariz, M.; Bendtz, K.; Ghafari, T.; Hetenyi, D.; Jeschke, J.; Kozma, C.; Mazumder, D. R.; and Melloni, L. 2025. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066): 133–142.

Fries, P. 2015. Rhythms For Cognition: Communication Through Coherence. *Neuron*, 88(1): 220–235.

Friston, K. J. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138.

Friston, K. J.; FitzGerald, T.; Rigoli, F.; Schwartenbeck, P.; and Pezzulo, G. 2017. Active Inference: A Process Theory. *Neural Computation*, 29(1): 1–49.

Gershman, S. J. 2019. The Generative Adversarial Brain. *Frontiers in Artificial Intelligence*, 2.

Graziano, M. S. A. 2013. *Consciousness and the Social Brain*. Oxford University Press.

Graziano, M. S. A. 2019. *Rethinking Consciousness: A Scientific Theory of Subjective Experience*. W. W. Norton & Company, first edition.

Greff, K.; van Steenkiste, S.; and Schmidhuber, J. 2020. On the Binding Problem in Artificial Neural Networks. *arXiv preprint arXiv:2012.05208*.

Grossberg, S. 2017. Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Networks*, 87: 38–95.

Gwern. 2020. The Scaling Hypothesis. https://gwern.net/scaling-hypothesis.

Haken, H. 2007. Synergetics. *Scholarpedia*, 2(1): 1400.

Hirsh, S. M.; Ichinaga, S. M.; Brunton, S. L.; Kutz, J. N.; and Brunton, B. W. 2021. Structured Time-Delay Models for Dynamical Systems with Connections to Frenet-Serret Frame. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2254): 20210097.

Hoel, E. 2025. Causal Emergence 2.0: Quantifying emergent complexity. *arXiv preprint arXiv:2503.13395*.

Hoel, E. P.; Albantakis, L.; Marshall, W.; and Tononi, G. 2016. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, 2016(1).

Hohwy, J.; and Seth, A. 2020. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Preprint, PsyArXiv:https://osf.io/preprints/psyarxiv/nd82g_v1.

Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; Hénaff, O.; Botvinick, M. M.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv preprint arXiv:2107.14795*.

Jaegle, A.; Gimeno, F.; Brock, A.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2021. Perceiver: General Perception with Iterative Attention. In *Proceedings of the 38 th International Conference on Machine Learning*.

Juliani, A.; Kanai, R.; and Sasai, S. S. 2022. The Perceiver Architecture is a Functional Global Workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Allen Lane.

Kant, I. 1781. *Critique of Pure Reason*. Cambridge University Press.

Klimaj, V.; and Safron, A. 2025. The natural history of intelligent systems: Toward understanding major transitions in cognitive evolution. OSF Preprint. https://doi.org/10.31234/osf.io/vhz56_v2.

Marr, D. 1983. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company.

McEliece, R. J.; MacKay, D. J. C.; and Cheng, J.-F. 1998. Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications*, 16(2): 140–152.

Metzinger, T. 2020. Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of "pure" consciousness. *Philosophy and the Mind Sciences*, 1(I): 1–44.

Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4): 435–450.

Northoff, G. 2012. Immanuel Kant's mind and the brain's resting state. *Trends in Cognitive Sciences*, 16(7): 356–359.

Northoff, G.; Klar, P.; Bein, M.; and Safron, A. 2023. As without, so within: How the brain's temporo-spatial alignment to the environment shapes consciousness. *Interface Focus*, 13(3): 20220076.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Rizzolatti, G.; and Craighero, L. 2010. Premotor theory of attention. *Scholarpedia*, 5(1): 6311.

Safron, A. 2020a. Integrated World Modeling Theory (IWMT) Implemented: Towards Reverse Engineering Consciousness with the Free Energy Principle and Active Inference. In Verbelen, T.; Lanillos, P.; Buckley, C. L.; and De Boom, C., eds., *Active Inference*, 135–155. Springer International Publishing.

Safron, A. 2020b. An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation. *Frontiers in Artificial Intelligence*, 3.

Safron, A. 2021a. The Radically Embodied Conscious Cybernetic Bayesian Brain: From Free Energy to Free Will and Back Again. *Entropy*, 23(6).

Safron, A. 2021b. World Models and the Physical Substrates of Consciousness: Hidden Sources of the Stream of Experience? *Journal of Consciousness Studies*, 28(11–12): 210–221.

Safron, A. 2022. Integrated world modeling theory expanded: Implications for the future of consciousness. *Frontiers in Computational Neuroscience*, 16.

Safron, A. 2023. AIXI, FEP-AI, and Integrated World Models: Towards a Unified Understanding of Intelligence and Consciousness. In Buckley, C. L.; Cialfi, D.; Lanillos, P.; Ramstead, M.; Sajid, N.; Shimazaki, H.; and Verbelen, T., eds., *Active Inference*, 251–273. Springer Nature Switzerland.

Safron, A.; Çatal, O.; and Verbelen, T. 2022. Generalized Simultaneous Localization and Mapping (G-SLAM) as unification framework for natural and artificial intelligences: Towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition. *Frontiers in Systems Neuroscience*, 16.

Safron, A.; Hipólito, I.; and Clark, A. 2023. Editorial: Bio A.I. - from embodied cognition to enactive robotics. *Frontiers in Neurorobotics*, 17.

Safron, A.; Juliani, A.; Reggente, N.; Klimaj, V.; and Johnson, M. 2025. On the varieties of conscious experiences: Altered Beliefs Under Psychedelics (ALBUS). *Neuroscience of Consciousness*, 2025(1): niae038.

Seth, A. K.; and Friston, K. J. 2016. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B*, 371(1708): 20160007.

Seth, A. K.; Suzuki, K.; and Critchley, H. D. 2012. An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2.

Shin, Y. K.; Proctor, R. W.; and Capaldi, E. J. 2010. A review of contemporary ideomotor theory. *Psychological Bulletin*, 136(6): 943–974.

Tegmark, M. 2015. Consciousness as a State of Matter. *Chaos, Solitons & Fractals*, 76: 238–270.

Thompson, E.; and Varela, F. J. 2001. Radical embodiment: neural dynamics and consciousness. *Trends in Cognitive Sciences*, 5(10).

Tononi, G. 2015. Integrated information theory. *Scholarpedia*, 10(1): 4164.

Wiese, W. 2020. The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness*, 2020(1): niaa013.

Yao, S.; Ping, Y.; Yue, X.; and Chen, H. 2025. Graph Convolutional Networks for multi-modal robotic martial arts leg pose recognition. *Frontiers in Neurorobotics*, 18.

Zahavy, T. 2026. LLMs can't jump. *preprint philsci-archive*. https://philsci-archive.pitt.edu/28024/.

Zhu, X.; Ruan, Q.; Qian, S.; and Zhang, M. 2025. A hybrid model based on transformer and Mamba for enhanced sequence modeling. *Scientific Reports*, 15(1): 11428.