Accepted to the 3rd International Workshop on Active Inference (IWAI 2022), held in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML//PKDD 2022).

AIXI, FEP-AI, and integrated world models: Towards a unified understanding of intelligence and consciousness

Adam Safron[1]

[1] Center for Psychedelic and Consciousness Research, Department of Psychiatry & Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21224 USA
asafron@gmail.com

Abstract. Intelligence has been operationalized as both goal-pursuit capacity across a broad range of environments, and also as learning capacity above and beyond a foundational set of core priors. Within the normative framework of AIXI, intelligence may be understood as capacities for compressing (and thereby predicting) data and achieving goals via programs with minimal algorithmic complexity. Within the Free Energy Principle and Active Inference framework, intelligence may be understood as capacity for inference and learning of predictive models for goal-realization, with beliefs favored to the extent they fit novel data with minimal updating of priors. Most recently, consciousness has been proposed to enhance intelligent functioning by allowing for iterative state estimation of the essential variables of a system and its relationships to its environment, conditioned on a causal world model. This paper discusses machine learning architectures and principles by which all these views may be synergistically combined and contextualized with an Integrated World Modeling Theory of consciousness.

Keywords: Free Energy Principle, Active Inference, AIXI, Integrated World Modeling Theory, Intelligence, Consciousness.

1 IWMT and Universal Intelligence

Integrated World Modeling Theory (IWMT) attempts to solve the enduring problems of consciousness by combining aspects of other models within the Free Energy Principle and Active Inference (FEP-AI) framework [1, 2]. In brief, IWMT claims that phenomenal consciousness is "what it feels like" when likely patterns of sense-data are inferred from probabilistic generative models over the sensorium of embodied-embedded agents, whose iterative estimates constitute the stream of experience (as a series of 'quale' states), conditioned on a causal world model trained from histories of environmental interaction. Different forms of "conscious access" are enabled when these streams/flows of experience are channeled/contextualized within coherent causal unfoldings via different varieties of mental actions, such as attentional selection of particular remembered and imagined contents. For such processes to be able to entail (or "bring forth a world" of) subjective experience, generative modeling is suggested to require sufficient coherence with respect to space (i.e., locality), time

(i.e. proportional changes for different processes), and cause (i.e., regularities with respect to changes that can be learned/inferred) for the system and its relationships with its environment. That is, structuring of experience by quasi-Kantian categories may be required not only for coherent judgment, but for solving binding problems such that different entities may be modeled with different properties [3–8]; without these sources of coherence, we may still have various forms of modeling, but not necessarily of a conscious variety due to a lack of compositional representations.

These system-world configurations are suggested to primarily (and potentially exclusively/exhaustively) involve visuospatial and somatospatial modalities, where views and poses mutually inform one another: estimating where and how one is looking is invaluable for constraining what one is seeing, and vice versa. This conjunction of view and pose information (particularly with respect to head and gaze direction) organizes experience according to egocentric reference frames (at least for systems with centrally-located censors), so providing a "point of view" on a world with the experience of a "lived body" at its center, with feelings likely heavily involving cross-modal inference with interoceptive hierarchies [9]. These processes are suggested to be realized by autoencoding heterarchies, whose shared latent space may be structured according to principles of geometric deep learning for generating state estimates with sufficient rapidity that they can both inform and be informed by action-perception cycles over the timescales over which they are enacted. If such iterative estimation of body-world states is organized into coherent organismic/agentic trajectories by semantic pointer systems—such as those provided by hippocampal/entorhinal system—then we may have episodic memory and (counterfactual) imaginings for the sake of prediction, postdiction, and planning/control [10–12]. Taken together, we may have most of the desiderata for explaining why (and how) it might feel like something to be some kinds of physical systems, potentially sufficiently answering hard central questions about biocomputational bases of experience in ways that afford new progress on the "easy-," "real-," and "meta-" problems of consciousness (which may be far more difficult than the "Hard problem" itself; e.g. the complexities of different forms of self-consciousness with the potential "strange loops" involved).

IWMT was significantly inspired by the work of Jürgen Schmidhuber (since 1990, see surveys of 2020, 2021), whose pioneering intellectual contributions include early discoveries in unsupervised machine learning, characterization of fundamental principles of intelligence, development of means of realizing scalable deep learning, designing world models of the kinds emphasized by IWMT, and more [15–24]. While this body of work cannot be adequately described in a single article, below I will explore how both the theoretical and practical applications of principles of universal intelligence and algorithmic information theory may help illuminate the nature(s) of consciousness.

1.1 FEP-AI and AIXI: Intelligence as Prediction/Compression

IWMT was developed within the unifying framework of FEP-AI as a formally-grounded model of intelligent behavior, with rich connections to both empirical phenomena and advances in machine learning. In this view, consciousness

and cognition more generally is understood as a process of learning how to better predict the world for the sake of adaptive control. However, these ideas could have been similarly expressed in the language of algorithmic information theory [25, 26], wherein complexity is quantified according to the length of programs capable of generating patterns of data (rather than the number of parameters for a generative model). From this point of view, intelligence is understood in terms of abilities to effectively compress information with algorithms of minimal complexity. These two perspectives can be understood as mutually supporting, in that predictability entails compressibility. If one can develop a predictive model of a system, then one also has access to a 'code'/cypher (even if only implicitly) capable of describing relevant information in a more compressed fashion. Similarly, if compression allows information to be represented with greater efficiency, then this compressed knowledge could be used as a basis for modeling future likely events, given past and present data.

This understanding of intelligence as compression was given a formal description in AIXI, a provably-optimal model of intelligent behavior that combines algorithmic and sequential decision theories [27]. While not formally computable, Solomonoff induction provides a Bayesian derivation of Occam's razor [28], in selecting models that minimize the Kolmogorov complexity of associated programs. AIXI agents use this principle in selecting programs that generate actions expected to maximize reward over some time horizon for an agent. This parsimony-constraint also constitutes a highly adaptive inductive bias and meta-prior in that we should indeed expect to be more likely to encounter simpler processes in the worlds we encounter [18, 29–32], all else being equal. Similarly, FEP-AI agents select policies (as Bayesian model selection over predictions) expected to maximize model evidence for realizing adaptive system-world states. However, this expectation is evaluated (or indirectly estimated with a variational evidence lower bound approach) according to a "free energy" functional that prioritizes inferences based on their ability to fit data, penalized by the degree to which updates of generative models diverge from priors, so preventing overfitting and facilitating knowledge generalization [33].

While thoroughly exploring connections between FEP-AI and AIXI is beyond the scope of this paper, their parallels are striking in that both frameworks prescribe choice behavior as the model-based realization of preferences under conditions of uncertainty, where governing models are privileged according to parsimony. Further, both FEP-AI and AIXI agents operate according to curiosity motivations, in which choices are selected not only to realize particular outcomes, but to explore for the sake of acquiring novel information in the service of model evolution [34, 35]. Both AIXI and FEP-AI confront the challenge of not being able to evaluate the full distribution of hypotheses required for normative program/model selection, instead relying on approximate inferred distributions. Further, both frameworks also attempt to overcome potentially complicated issues around ergodic exploration of state spaces with various forms of sophisticated planning [36, 37]. Taken together, the combination of these frameworks may provide a provably-optimal, formally-grounded theory for studying intelligent behavior in both biological and artificial systems. With relevance to IWMT, consciousness may be understood as a kind of lingua franca (or semi-interpretable shared latent space) for more efficiently converting between heterogeneous systems for

the sake of promoting functional synergy [1, 2, 38]. [While beyond the scope of the present discussion, it may be worth considering which forms of conceptual and functional integration between AIXI and FEP-AI could be achievable through the study of generative modeling based on probabilistic programs [39–41]. In terms of the neural systems described by IWMT, many of these "programs" could be understood as realized through modes of policy selection canalized by (a) striatal-cortical loops as sequences of both overtly enacted and covertly expressed mental actions and (b) large scale state transitions between equilibrium points as orchestrated by the hippocampal/entorhinal system [10].]

Although AIXI and FEP-AI both specify optimality criteria, both frameworks require the further specification of process theories and heuristic implementations in order to realize these principles in actual systems [42, 43]. Below I focus on some of the work of Schmidhuber [13, 14] and colleagues in which these principles of universal computation have been used to inform the creation of artificial systems with many of the hallmarks of biological intelligences. Understanding the design of these systems may be informative for understanding not only computational properties of nervous systems, but also their ability to generate integrative world models for the sake of adaptively controlling behavior.

1.2 Kinds of world models

[Please note: For background on computational properties of different kinds of neural networks and their potential implications for intelligence, please see Appendix: "2.1 Recurrent networks, universal computation, and generalized predictive coding."]

World models are becoming an increasingly central topic in machine learning due to their ability to support knowledge synthesis for self-supervised learning, especially for agents governed by objective functions (potentially implicit and meta-learned) involving intrinsic sources of value such as curiosity and empowerment [13–15, 20, 22, 44–47]. IWMT emphasizes consciousness as an entailment of integrated world modeling [1], and attempts to specify which brain systems and machine learning principles may be important for realizing different aspects of this function [2]. According to IWMT, posterior cortices represent generative models over the sensorium for embodied agents, whose inversion generates likely patterns of sense data (given past experience), entailing streams of consciousness. Cortical perceptual hierarchies are modelled as "folded" variational autoencoders (VAEs), but this could similarly be described as hierarchically-organized LSTMs [16, 48], or perhaps (generalized) transformer hierarchies [49–52]. There is a sense in which this posterior-located autoencoding heterarchy is itself a world model by virtue of not only containing information regarding the spatially-organized structure of environments, but also deep temporal modeling via the capacity of recurrent systems—and complex dendritic processing [53]—to support sequence memories, whose coherent state transitions could be understood as entailing causal dependencies. However, this world modeling achieves much greater temporal depth and counterfactual richness with the addition of the hippocampal-entorhinal system and frontal lobes [54–56], which together may generate likely-future (or counterfactual/postdicted-past) state transitions. Further, sculpting of striatal-cortical loops by neuromodulators allows

action—broadly construed to include forms of attentional selection and imaginative mental acts—to be selected and policies to be channeled by histories of reward learning [57]. This involvement of value-canalized control structures provides functional closure in realizing action-perception cycles, which are likely preconditions for any kind of coherent world modeling to be achieved via active inference [58, 59].

These sorts of functions have been realized in artificial systems for decades [15, 16], with recent advances exhibiting remarkable convergence with the neurocomputational account of world modeling described by IWMT. One particularly notable world modeling system was developed by Ha and Schmidhuber [60], in which a VAE is used to compress sense data into reduced-dimensionality latent vectors, which are then predicted by a lower parameter recurrent neural network (RNN), whose outputs are then used as bases for policy selection by a much lower parameter controller trained with evolutionary strategies. The ensuing actions from this control system are then used as a basis for further predictive modeling and policy selection, so completing action-perception cycles for further inference and learning. The dimensionality-reducing capacities of VAEs not only provide greater efficiency in providing compressions for subsequent computation, but moving from pixel space to latent space—with associated principle components of data variation—allow for more tractable training regimes. The use of evolutionary optimization is capable of handling challenging credit-assignment problems (e.g. non-differentiable free energy landscapes) via utilizing the final cumulative reward over a behavioral epoch, rather than requiring evaluation of the entire history of performance and attempting to determine which actions contributed to which outcomes to which degrees. (Speculatively, this could be understood as one interpretation of hippocampal ripples contributing to phasic dopamine signals in conjunction with remapping events [61, 62].) This architecture was able to achieve unprecedented performance on car racing, first-person shooter video games, and challenging locomotion problems requiring the avoidance of local minima. Further, by feeding the signals from the controller directly back into latent space (rather than issuing commands over actions), this architecture was able to achieve even greater functionality by allowing training with policy rollouts in simulated (or "imagined") environments. Even more, by incorporating a temperature (or uncertainty) parameter into this imaginative training, the system was able to avoid patterns of self-adversarial/degenerate policy selection from leveraging exploits of internal world models, so enhancing transfer learning from simulations to real-world situations (and back again). Theoretically, this could help explain some of the functional significance of dreaming [63], and possibly also aspects of neuromodulator systems (functionally understood as machine learning parameters) and the ways in which they may be involved in various "psychedelic" states of mind [64].

Limitations of this world modeling system include the problem of independent VAE-optimization resulting in encoding task-irrelevant observations, but with task-based training potentially undermining transfer learning. An additional challenge involves issues of catastrophic forgetting as the encoding of new data interferes with prior memories. Ha and Schmidhuber [60] suggested that these challenges could be respectively overcome with governance by intrinsic drives for artificial curiosity—which could be thought of as a kind of adaptive-adversarial

learning (Schmidhuber, 1990, 2020, 2021)—as well as the incorporation of external memory modules for exploring more complicated worlds. Intriguingly, both of these properties seem to be a feature of the hippocampal/entorhinal-system [2, 65, 66], which possesses high centrality and effective connectivity with value-estimation networks such as ventral prefrontal cortices and striatum [57]. An additional means of enhancing the performance of world modeling architectures is the incorporation of separate "teacher" and "student" networks, whose particular functional details may shed light on not just the conditions for optimal learning, but for the particular nature(s) of conscious experience, as will be explored next.

1.3 Kinds of minds: bringing forth worlds of experience

One of Schmidhuber's [16] more intriguing contributions involves a two-tier unsupervised machine (meta-)learning architecture operating according to principles of predictive coding. The lower level consists of a fast "subconscious automatiser RNN" that attempts to compress (or predict) action sequences, so learning to efficiently encode histories of actions and associated observations. This subsystem automatically creates abstraction hierarchies with similar features to those found in receptive fields of the mammalian neocortical ventral visual stream. In this way, the automatiser identifies invariances (or symmetries) across data structures and generates prototypical features that can be used to efficiently represent particular instances by only indicating deviations from these eigenmode basis functions. Surprising information not predicted by this lower level is passed up to a more slowly operating "conscious chunker RNN," so becoming the object of attention for more elaborative processing [23]. The automatiser continuously attempts to encode subsequent streams of information in compressed forms, eventually rendering them "subconscious" in terms of not requiring the more complex (expensive, and relatively slow) modeling of the conscious chunker. This setup can also be understood as a "teacher" network being imitated by a "student" network, and through this knowledge distillation (and meta-learning) process, high-level policy selection becomes "cloned" such that it can be realized by fast and efficient lower-level processes. With the "Neural History Compressor" [24, 67], this architecture was augmented with the additional functionality of having the strength of updates for the conscious chunker be modulated by the magnitude of surprises for the subconscious automatiser.

In terms of brain functioning, we would expect predictive processing mechanisms to cause information to be predicted in a way that is increasingly sparse and pushed closer to primary modalities with time and experience. Because these areas lack access to the kind of rich club connectivity found in deeper portions of the cortical heterarchy, this information would be less likely to be taken up into workspace dynamics and contribute to alpha/beta-synchronized large-scale "ignition" events, which IWMT interprets as Bayesian model selection over the sensorium of an embodied-embedded agent, entailing iterative estimation of system-world states. In this way, initially conscious patterns of action selection involving workspace dynamics would constitute "teacher" networks, but which with experience will tend to become automatic/habitual/unconscious as they are increasingly realized by soft-assembled, distributed, fast/small ensembles coordinated by cerebellar-coordinated forward

models [68], functioning as "student" networks. In terms of some of the architectures associated with FEP-AI, this could be thought of as information being processed in the lower level of Forney factor graphs with continuous (as opposed to discrete) updating and less clearly symbolic functionality [69, 70]. In more standard machine learning terms, this could be thought as moving from more "model-based" to more "model-free" control, or "amortized inference" with respect to policy selection [57, 71]. However, when these more unconscious forms of inference fail, associated prediction errors will ascend until they can receive more elaborative handling by conscious systems, whose degree of activity may be modulated by overall surprisal [2, 72–75].

To return to artificial neural history compressors [24, 67], these systems are capable of discovering a wide range of invariances across actions and associated observations, with one of the most enduring of these symmetries being constituted by the agent itself as a relatively constant feature. With RNN-realized predictive coding mechanisms combined with reinforcement learning, systems learn to efficiently represent themselves, with these symbols becoming activated in the context of planning and through active self-exploration, so exhibiting a basic form of self-modeling, including with respect to counterfactual imaginings. Based on the realization of these functionalities, Schmidhuber has controversially claimed that "we have had simple, conscious, self-aware, emotional, artificial agents for 3 decades" [13, 14, 76]. IWMT has previously focused on the neurocomputational processes that may realize phenomenal consciousness, but functionally speaking (albeit potentially not experientially), this statement is consistent with the computational principles identified for integrated world modeling.

These functional and mechanistic accounts have strong correspondences aspects of Hiedeggerian phenomenology [77]. A "ready-to-hand" situation of effortless mastery of sensorimotor contingencies could correspond to maximally sparse/compressed patterns of neural activity, pushed down close to primary cortices, with the system coupling with the world in continuously evolving unconscious action-perception cycles, without requiring explicit representation of information. In such a situation of nearly effortless (and largely unconscious) mastery of sensorimotor contingencies, dithering would be prevented as the changing affordance landscape results in rapid selection of appropriate grips [9, 78–81], with coordination largely occurring via reflexes and morphological constraints. However, a "present-at-hand" situation of reflective evaluation could correspond to prediction errors making their way up to upper levels of cortical heterarchies, where largescale ignition events—potentially entailing Bayesian model selection and discrete updating of hierarchically higher (or deeper) beliefs—may become more likely due to closer proximity to rich-club connectivity cores. If such novel information can couple with subsystems enabling various kinds of workspace dynamics, then agents can not only consciously perceive such patterns, but also draw upon counterfactual considerations for causal reasoning and planning, so requiring flexibly (and consciously) accessible re(-)presentations and various forms of self-referential modeling. Convergence between these machine learning architectures, neural systems, and phenomenology is compelling, but does this mean that Schmidhuberian world-modeling agents and similar systems possess

phenomenal consciousness? This is the final issue that we will consider in exploring world models and the nature(s) of intelligence.

1.4 Machine consciousness?

[*Please note: For discussion of how different principles from neural networks may contribute to explaining computational properties of consciousness, please see Appendix: "2.2 Unfolding and (potentially conscious) self-world modeling."*]

More recent work from Schmidhuber's group has explored the preconditions for human-like abilities to generalize knowledge beyond direct experience [3]. They suggest that such capable systems require capacities to flexibly bind distributed information in novel combinations, and where this binding problem requires capacities for compositional and symbolization of world structure for systematic and predictable knowledge synthesis. They further describe a unifying framework involving the formation of meaningful internal structure from unstructured sense data (segregation), the maintenance of this segregated data at an abstract level (representation), and the ability to construct new inferences through novel combinations of these representations (composition). Similarly to the preconditions for subjective experience identified by Integrated Information Theory (IIT) [1, 82]—and with deep relevance to IWMT—this proposal describes a protocol for producing data structures with intrinsic existence (realization via entangled connections), composition (possessing meaningful structure), information (distinguishability from alternatives), and integration (constituting wholes that are greater than the sum of their parts). This is further compatible with GNWT's description of architectures for dynamically combining information in synergistic ways.

Even more, the particular architectural principles identified by this framework as promising research directions are the same as those previously suggested by IWMT: geometric deep learning and graph neural networks (GNNs). However, this proposal for solving binding problems for neural networks also suggests utilizing a generalization of GNNs in the form of graph nets [83]. This is also an issue with respect to which the original publication of IWMT contained a possible error, in that rather than the hippocampal/entorhinal-system being understood as hosting graph-grid GNNs for spatial mapping, (spectral) graph nets may be a more apt description of this system on algorithmic and functional levels of analysis [66, 84, 85]. These technical details notwithstanding, there are striking parallels between the functionalities described by this framework for functional binding and the machine learning description of neural systems suggested by IWMT. More specifically, autoencoding/compressing hierarchies provide segregation, with GNN-structured latent spaces providing explicit representation (and potentially conscious experiences), and a further level of higher abstraction providing novel representational combinations for creative cognition (i.e., hippocampal place fields as flexibly configurable graphical models).

If such an architecture were to be artificially realized, would it possess 'consciousness?' The Global Neuronal Workspace Theory would likely answer this question in the affirmative with respect to conscious access [86], and IIT would provide a tentative affirmation for phenomenal consciousness [87], given realization on

neuromorphic hardware capable of generating high phi for the physical system (rather than for an entailed virtual machine). IWMT is agnostic on this issue, yet may side with IIT in terms of matters of practical realization. Something like current developments with specialized graph processors might be required for these systems to achieve sufficient efficiency [88] for engaging in iterative estimation of system-world states with sufficient rapidity that such estimates could both form and be informed by perceivable action-perception cycles. If such a system were constructed as a controller for an embodied-embedded agent, then it may be the case that it could not only generate "System 2" abilities [89, 90], but also phenomenal consciousness, and everything that entails.

Acknowledgments

In addition to collaborators who have taught me countless lessons over many years, I would like to thank Jürgen Schmidhuber and Karl Friston for their generous feedback on previous versions of these discussions. Any errors (either technical or stylistic) are entirely my own.

References

1. Safron, A.: 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. Front. Artif. Intell. 3, (2020). https://doi.org/10.3389/frai.2020.00030.
2. Safron, A.: Integrated World Modeling Theory (IWMT) Implemented: Towards Reverse Engineering Consciousness with the Free Energy Principle and Active Inference. PsyArXiv (2020). https://doi.org/10.31234/osf.io/paz5j.
3. Greff, K., van Steenkiste, S., Schmidhuber, J.: On the Binding Problem in Artificial Neural Networks. arXiv:2012.05208 [cs]. (2020).
4. Evans, R., Hernández-Orallo, J., Welbl, J., Kohli, P., Sergot, M.: Making sense of sensory input. Artificial Intelligence. 293, 103438 (2021). https://doi.org/10.1016/j.artint.2020.103438.
5. De Kock, L.: Helmholtz's Kant revisited (Once more). The all-pervasive nature of Helmholtz's struggle with Kant's Anschauung. Stud Hist Philos Sci. 56, 20–32 (2016). https://doi.org/10.1016/j.shpsa.2015.10.009.
6. Northoff, G.: Immanuel Kant's mind and the brain's resting state. Trends Cogn. Sci. (Regul. Ed.). 16, 356–359 (2012). https://doi.org/10.1016/j.tics.2012.06.001.
7. Swanson, L.R.: The Predictive Processing Paradigm Has Roots in Kant. Front Syst Neurosci. 10, 79 (2016). https://doi.org/10.3389/fnsys.2016.00079.
8. Marcus, G.: The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177 [cs]. (2020).

9. Safron, A.: The Radically Embodied Conscious Cybernetic Bayesian Brain: From Free Energy to Free Will and Back Again. Entropy. 23, 783 (2021). https://doi.org/10.3390/e23060783.
10. Safron, A., Çatal, O., Verbelen, T.: 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, https://psyarxiv.com/tdw82/, (2021). https://doi.org/10.31234/osf.io/tdw82.
11. Safron, A., Sheikhbahaee, Z.: Dream to explore: 5-HT2a as adaptive temperature parameter for sophisticated affective inference, https://psyarxiv.com/zmpaq/, (2021). https://doi.org/10.31234/osf.io/zmpaq.
12. Safron, A.: On the Varieties of Conscious Experiences: Altered Beliefs Under Psychedelics (ALBUS), https://psyarxiv.com/zqh4b/, (2020). https://doi.org/10.31234/osf.io/zqh4b.
13. Schmidhuber, J.: 1990: Planning & Reinforcement Learning with Recurrent World Models and Artificial Curiosity, https://people.idsia.ch//~juergen/world-models-planning-curiosity-fki-1990.html, last accessed 2021/05/16.
14. Schmidhuber, J.: 1991: First very deep learning with unsupervised pre-training, https://people.idsia.ch//~juergen/very-deep-learning-1991.html, last accessed 2021/05/16.
15. Schmidhuber, J.: Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments. (1990).
16. Schmidhuber, J.: Neural Sequence Chunkers. (1991).
17. Schmidhuber, J.: Learning complex, extended sequences using the principle of history compression. Neural Comput. 4, 234–242 (1992). https://doi.org/10.1162/neco.1992.4.2.234.
18. Schmidhuber, J.: Algorithmic Theories of Everything. arXiv:quant-ph/0011122. (2000).
19. Schmidhuber, J.: The Speed Prior: A New Simplicity Measure Yielding Near-Optimal Computable Predictions. In: Kivinen, J. and Sloan, R.H. (eds.) Computational Learning Theory. pp. 216–228. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-45435-7_15.
20. Schmidhuber, J.: Gödel Machines: Fully Self-referential Optimal Universal Self-improvers. In: Goertzel, B. and Pennachin, C. (eds.) Artificial General Intelligence. pp. 199–226. Springer, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-68677-4_7.
21. Schmidhuber, J.: Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity & Creativity. arXiv:0709.0674 [cs]. (2007).
22. Schmidhuber, J.: POWERPLAY: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem. arXiv:1112.5309 [cs]. (2012).

23. Schmidhuber, J.: On Learning to Think: Algorithmic Information Theory for Novel Combinations of Reinforcement Learning Controllers and Recurrent Neural World Models. arXiv:1511.09249 [cs]. (2015).
24. Schmidhuber, J.: One Big Net For Everything. arXiv:1802.08864 [cs]. (2018).
25. Kolmogorov, A.N.: On Tables of Random Numbers. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002). 25, 369–376 (1963).
26. Schmidhuber, J.: Hierarchies of generalized kolmogorov complexities and nonenumerable universal measures computable in the limit. Int. J. Found. Comput. Sci. 13, 587–612 (2002). https://doi.org/10.1142/S0129054102001291.
27. Hutter, M.: A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. arXiv:cs/0004001. (2000).
28. Solomonoff, R.J.: Algorithmic Probability: Theory and Applications. In: Emmert-Streib, F. and Dehmer, M. (eds.) Information Theory and Statistical Learning. pp. 1–23. Springer US, Boston, MA (2009). https://doi.org/10.1007/978-0-387-84816-7_1.
29. Feynman, R.P. (Richard P.: Quantum mechanics and path integrals. New York : McGraw-Hill (1965).
30. Kaila, V., Annila, A.: Natural selection for least action. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 464, 3055–3070 (2008). https://doi.org/10.1098/rspa.2008.0178.
31. Campbell, J.O.: Universal Darwinism As a Process of Bayesian Inference. Front Syst Neurosci. 10, 49 (2016). https://doi.org/10.3389/fnsys.2016.00049.
32. Vanchurin, V.: The World as a Neural Network. Entropy. 22, 1210 (2020). https://doi.org/10.3390/e22111210.
33. Hanson, S.J.: A stochastic version of the delta rule. Physica D: Nonlinear Phenomena. 42, 265–272 (1990). https://doi.org/10.1016/0167-2789(90)90081-Y.
34. Orseau, L., Lattimore, T., Hutter, M.: Universal Knowledge-Seeking Agents for Stochastic Environments. In: Jain, S., Munos, R., Stephan, F., and Zeugmann, T. (eds.) Algorithmic Learning Theory. pp. 158–172. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40935-6_12.
35. Friston, K.J., Lin, M., Frith, C.D., Pezzulo, G., Hobson, J.A., Ondobaka, S.: Active Inference, Curiosity and Insight. Neural Comput. 29, 2633–2683 (2017). https://doi.org/10.1162/neco_a_00999.
36. Aslanides, J., Leike, J., Hutter, M.: Universal Reinforcement Learning Algorithms: Survey and Experiments. arXiv:1705.10557 [cs]. (2017).
37. Friston, K., Da Costa, L., Hafner, D., Hesp, C., Parr, T.: Sophisticated Inference. (2020).
38. VanRullen, R., Kanai, R.: Deep learning and the Global Workspace Theory. Trends in Neurosciences. (2021). https://doi.org/10.1016/j.tins.2021.04.005.
39. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science. 350, 1332–1338 (2015). https://doi.org/10.1126/science.aab3050.

40. Lázaro-Gredilla, M., Lin, D., Guntupalli, J.S., George, D.: Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. Science Robotics. 4, (2019). https://doi.org/10.1126/scirobotics.aav3150.
41. Ullman, T.D., Tenenbaum, J.B.: Bayesian Models of Conceptual Development: Learning as Building Models of the World. Annual Review of Developmental Psychology. 2, 533–558 (2020). https://doi.org/10.1146/annurev-devpsych-121318-084833.
42. Veness, J., Ng, K.S., Hutter, M., Uther, W., Silver, D.: A Monte Carlo AIXI Approximation. arXiv:0909.0801 [cs, math]. (2010).
43. Hesp, C., Tschantz, A., Millidge, B., Ramstead, M., Friston, K., Smith, R.: Sophisticated Affective Inference: Simulating Anticipatory Affective Dynamics of Imagining Future Events. In: Verbelen, T., Lanillos, P., Buckley, C.L., and De Boom, C. (eds.) Active Inference. pp. 179–186. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-64919-7_18.
44. de Abril, I.M., Kanai, R.: A unified strategy for implementing curiosity and empowerment driven reinforcement learning. arXiv:1806.06505 [cs]. (2018).
45. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to Control: Learning Behaviors by Latent Imagination. arXiv:1912.01603 [cs]. (2020).
46. Hafner, D., Ortega, P.A., Ba, J., Parr, T., Friston, K., Heess, N.: Action and Perception as Divergence Minimization. arXiv:2009.01791 [cs, math, stat]. (2020).
47. Wang, R., Lehman, J., Rawal, A., Zhi, J., Li, Y., Clune, J., Stanley, K.O.: Enhanced POET: Open-Ended Reinforcement Learning through Unbounded Invention of Learning Challenges and their Solutions. arXiv:2003.08536 [cs]. (2020).
48. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. 9, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735.
49. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: FNet: Mixing Tokens with Fourier Transforms. arXiv:2105.03824 [cs]. (2021).
50. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G.K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: Hopfield Networks is All You Need. arXiv:2008.02217 [cs, stat]. (2021).
51. Schlag, I., Irie, K., Schmidhuber, J.: Linear Transformers Are Secretly Fast Weight Memory Systems. arXiv:2102.11174 [cs]. (2021).
52. Tay, Y., Dehghani, M., Gupta, J., Bahri, D., Aribandi, V., Qin, Z., Metzler, D.: Are Pre-trained Convolutions Better than Pre-trained Transformers? arXiv:2105.03322 [cs]. (2021).
53. Hawkins, J., Ahmad, S.: Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex. Front. Neural Circuits. 10, (2016). https://doi.org/10.3389/fncir.2016.00023.
54. Knight, R.T., Grabowecky, M.: Escape from linear time: Prefrontal cortex and conscious experience. In: The cognitive neurosciences. pp. 1357–1371. The MIT Press, Cambridge, MA, US (1995).

55. Koster, R., Chadwick, M.J., Chen, Y., Berron, D., Banino, A., Düzel, E., Hassabis, D., Kumaran, D.: Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes. Neuron. 99, 1342-1354.e6 (2018). https://doi.org/10.1016/j.neuron.2018.08.009.
56. Faul, L., St. Jacques, P.L., DeRosa, J.T., Parikh, N., De Brigard, F.: Differential contribution of anterior and posterior midline regions during mental simulation of counterfactual and perspective shifts in autobiographical memories. NeuroImage. 215, 116843 (2020). https://doi.org/10.1016/j.neuroimage.2020.116843.
57. Mannella, F., Gurney, K., Baldassarre, G.: The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. Front Behav Neurosci. 7, 135 (2013). https://doi.org/10.3389/fnbeh.2013.00135.
58. Friston, K.J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G.: Active Inference: A Process Theory. Neural Comput. 29, 1–49 (2017). https://doi.org/10.1162/NECO_a_00912.
59. Friston, K.J.: Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). Front. Psychol. 9, (2018). https://doi.org/10.3389/fpsyg.2018.00579.
60. Ha, D., Schmidhuber, J.: World Models. arXiv:1803.10122 [cs, stat]. (2018). https://doi.org/10.5281/zenodo.1207631.
61. Rusu, S.I., Pennartz, C.M.A.: Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized cortico-basal ganglia systems. Hippocampus. 30, 73–98 (2020). https://doi.org/10.1002/hipo.23167.
62. Sanders, H., Wilson, M.A., Gershman, S.J.: Hippocampal remapping as hidden state inference. eLife. 9, e51140 (2020). https://doi.org/10.7554/eLife.51140.
63. Hoel, E.: The overfitted brain: Dreams evolved to assist generalization. Patterns. 2, 100244 (2021). https://doi.org/10.1016/j.patter.2021.100244.
64. Boureau, Y.-L., Dayan, P.: Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin. Neuropsychopharmacology. 36, 74–97 (2011). https://doi.org/10.1038/npp.2010.151.
65. Hassabis, D., Maguire, E.A.: The construction system of the brain. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 364, 1263–1271 (2009). https://doi.org/10.1098/rstb.2008.0296.
66. Çatal, O., Verbelen, T., Van de Maele, T., Dhoedt, B., Safron, A.: Robot navigation as hierarchical active inference. Neural Networks. 142, 192–204 (2021). https://doi.org/10.1016/j.neunet.2021.05.010.
67. Schmidhuber, J.H., Mozer, M.C., Prelinger, D.: Continuous History Compression. In: Proc. of Intl. Workshop on Neural Networks, RWTH Aachen. pp. 87–95. Augustinus (1993).
68. Shine, J.M.: The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. Prog Neurobiol. 199, 101951 (2021). https://doi.org/10.1016/j.pneurobio.2020.101951.

69. Friston, K.J., Parr, T., de Vries, B.: The graphical brain: Belief propagation and active inference. Network Neuroscience. 1, 381–414 (2017). https://doi.org/10.1162/NETN_a_00018.
70. Parr, T., Friston, K.J.: The Discrete and Continuous Brain: From Decisions to Movement-And Back Again. Neural Comput. 30, 2319–2347 (2018). https://doi.org/10.1162/neco_a_01102.
71. Gershman, S., Goodman, N.: Amortized Inference in Probabilistic Reasoning. Proceedings of the Annual Meeting of the Cognitive Science Society. 36, (2014).
72. Sales, A.C., Friston, K.J., Jones, M.W., Pickering, A.E., Moran, R.J.: Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model. PLOS Computational Biology. 15, e1006267 (2019). https://doi.org/10.1371/journal.pcbi.1006267.
73. Shea, N., Frith, C.D.: The Global Workspace Needs Metacognition. Trends in Cognitive Sciences. 0, (2019). https://doi.org/10.1016/j.tics.2019.04.007.
74. Shine, J.: Neuromodulatory Influences on Integration and Segregation in the Brain. undefined. (2019).
75. Holroyd, C.B., Verguts, T.: The Best Laid Plans: Computational Principles of Anterior Cingulate Cortex. Trends in Cognitive Sciences. 25, 316–329 (2021). https://doi.org/10.1016/j.tics.2021.01.008.
76. Carmichael, J.: Artificial Intelligence Gained Consciousness in 1991, https://www.inverse.com/article/25521-juergen-schmidhuber-ai-consciousness, last accessed 2021/11/14.
77. Dreyfus, H.L.: Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. Philosophical Psychology. 20, 247–268 (2007). https://doi.org/10.1080/09515080701239510.
78. Cisek, P.: Cortical mechanisms of action selection: the affordance competition hypothesis. Philos Trans R Soc Lond B Biol Sci. 362, 1585–1599 (2007). https://doi.org/10.1098/rstb.2007.2054.
79. Seth, A.K.: The Cybernetic Bayesian Brain. Open MIND. Frankfurt am Main: MIND Group (2014). https://doi.org/10.15502/9783958570108.
80. Tani, J.: Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press (2016).
81. Kiverstein, J., Miller, M., Rietveld, E.: The feeling of grip: novelty, error dynamics, and the predictive brain. Synthese. 196, 2847–2869 (2019). https://doi.org/10.1007/s11229-017-1583-9.
82. Tononi, G., Boly, M., Massimini, M., Koch, C.: Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience. 17, 450 (2016). https://doi.org/10.1038/nrn.2016.44.
83. Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., Pascanu, R.: Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 [cs, stat]. (2018).

84. Gothoskar, N., Guntupalli, J.S., Rikhye, R.V., Lázaro-Gredilla, M., George, D.: Different clones for different contexts: Hippocampal cognitive maps as higher-order graphs of a cloned HMM. bioRxiv. 745950 (2019). https://doi.org/10.1101/745950.
85. Peer, M., Brunec, I.K., Newcombe, N.S., Epstein, R.A.: Structuring Knowledge with Cognitive Maps and Cognitive Graphs. Trends in Cognitive Sciences. 25, 37–54 (2021). https://doi.org/10.1016/j.tics.2020.10.004.
86. Dehaene, S.: Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking, New York, New York (2014).
87. Tononi, G., Koch, C.: Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society B: Biological Sciences. 370, 20140167 (2015). https://doi.org/10.1098/rstb.2014.0167.
88. Ortiz, J., Pupilli, M., Leutenegger, S., Davison, A.J.: Bundle Adjustment on a Graph Processor. arXiv:2003.03134 [cs]. (2020).
89. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux (2011).
90. Bengio, Y.: The Consciousness Prior. arXiv:1709.08568 [cs, stat]. (2017).
91. Lange, S., Riedmiller, M.: Deep auto-encoder neural networks in reinforcement learning. In: The 2010 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2010). https://doi.org/10.1109/IJCNN.2010.5596468.
92. Lotter, W., Kreiman, G., Cox, D.: Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. arXiv:1605.08104 [cs, q-bio]. (2016).
93. Wu, Y., Wayne, G., Graves, A., Lillicrap, T.: The Kanerva Machine: A Generative Distributed Memory. arXiv:1804.01756 [cs, stat]. (2018).
94. Jiang, Y., Kim, H., Asnani, H., Kannan, S., Oh, S., Viswanath, P.: Turbo Autoencoder: Deep learning based channel codes for point-to-point communication channels. arXiv:1911.03038 [cs, eess, math]. (2019).
95. Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., Guttenberg, N.: Information generation as a functional basis of consciousness. Neurosci Conscious. 2019, (2019). https://doi.org/10.1093/nc/niz016.
96. Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., Hinton, G.: Backpropagation and the brain. Nature Reviews Neuroscience. 1–12 (2020). https://doi.org/10.1038/s41583-020-0277-3.
97. Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S.: The Helmholtz machine. Neural Comput. 7, 889–904 (1995).
98. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat]. (2014).
99. Candadai, M., Izquierdo, E.J.: Sources of predictive information in dynamical neural networks. Scientific Reports. 10, 16901 (2020). https://doi.org/10.1038/s41598-020-73380-x.
100. Lu, Z., Bassett, D.S.: Invertible generalized synchronization: A putative mechanism for implicit learning in neural systems. Chaos. 30, 063133 (2020). https://doi.org/10.1063/5.0004344.
101. Rumelhart, D.E., McClelland, J.L.: Information Processing in Dynamical Systems: Foundations of Harmony Theory. In: Parallel Distributed Processing:

Explorations in the Microstructure of Cognition: Foundations. pp. 194–281. MIT Press (1987).
102. Kachman, T., Owen, J.A., England, J.L.: Self-Organized Resonance during Search of a Diverse Chemical Space. Phys. Rev. Lett. 119, 038001 (2017). https://doi.org/10.1103/PhysRevLett.119.038001.
103. Friston, K.J.: A free energy principle for a particular physics. arXiv:1906.10184 [q-bio]. (2019).
104. Ali, A., Ahmad, N., Groot, E. de, Gerven, M.A.J. van, Kietzmann, T.C.: Predictive coding is a consequence of energy efficiency in recurrent neural networks. bioRxiv. 2021.02.16.430904 (2021). https://doi.org/10.1101/2021.02.16.430904.
105. Bejan, A., Lorente, S.: The constructal law of design and evolution in nature. Philos Trans R Soc Lond B Biol Sci. 365, 1335–1347 (2010). https://doi.org/10.1098/rstb.2009.0302.
106. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 5, 115–133 (1943).
107. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 15, 1929–1958 (2014).
108. Ahmad, S., Scheinkman, L.: How Can We Be So Dense? The Benefits of Using Highly Sparse Representations. arXiv preprint arXiv:1903.11257. (2019).
109. Mumford, D.: On the computational architecture of the neocortex. Biol. Cybern. 65, 135–145 (1991). https://doi.org/10.1007/BF00202389.
110. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87 (1999). https://doi.org/10.1038/4580.
111. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J.: Canonical microcircuits for predictive coding. Neuron. 76, 695–711 (2012). https://doi.org/10.1016/j.neuron.2012.10.038.
112. Grossberg, S.: Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. Neural Networks. 87, 38–95 (2017). https://doi.org/10.1016/j.neunet.2016.11.003.
113. Heeger, D.J.: Theory of cortical function. Proc. Natl. Acad. Sci. U.S.A. 114, 1773–1782 (2017). https://doi.org/10.1073/pnas.1619788114.
114. George, D., Lázaro-Gredilla, M., Lehrach, W., Dedieu, A., Zhou, G.: A detailed mathematical theory of thalamic and cortical microcircuits based on inference in a generative vision model. bioRxiv. 2020.09.09.290601 (2020). https://doi.org/10.1101/2020.09.09.290601.
115. Friston, K.J., Rosch, R., Parr, T., Price, C., Bowman, H.: Deep temporal models and active inference. Neurosci Biobehav Rev. 77, 388–402 (2017). https://doi.org/10.1016/j.neubiorev.2017.04.009.
116. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books (2018).
117. Csáji, B.C.: Approximation with artificial neural networks. Faculty of Sciences, Etvs Lornd University, Hungary. 24, 7 (2001).

118. Malach, E., Shalev-Shwartz, S.: Is Deeper Better only when Shallow is Good? arXiv:1903.03488 [cs, stat]. (2019).
119. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway Networks. arXiv:1505.00387 [cs]. (2015).
120. Lin, H.W., Tegmark, M., Rolnick, D.: Why does deep and cheap learning work so well? J Stat Phys. 168, 1223–1247 (2017). https://doi.org/10.1007/s10955-017-1836-5.
121. Sherstinsky, A.: Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. Physica D: Nonlinear Phenomena. 404, 132306 (2020). https://doi.org/10.1016/j.physd.2019.132306.
122. Schmidhuber, J.: On learning how to learn learning strategies., (1994).
123. Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., Botvinick, M.: Prefrontal cortex as a meta-reinforcement learning system. Nature Neuroscience. 21, 860 (2018). https://doi.org/10.1038/s41593-018-0147-8.
124. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature. 393, 440 (1998). https://doi.org/10.1038/30918.
125. Jarman, N., Steur, E., Trengove, C., Tyukin, I.Y., van Leeuwen, C.: Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion. Scientific Reports. 7, 13158 (2017). https://doi.org/10.1038/s41598-017-12589-9.
126. Rentzeperis, I., Laquitaine, S., van Leeuwen, C.: Adaptive rewiring of random neural networks generates convergent-divergent units. arXiv:2104.01418 [q-bio]. (2021).
127. Massobrio, P., Pasquale, V., Martinoia, S.: Self-organized criticality in cortical assemblies occurs in concurrent scale-free and small-world networks. Scientific Reports. 5, 10578 (2015). https://doi.org/10.1038/srep10578.
128. Gal, E., London, M., Globerson, A., Ramaswamy, S., Reimann, M.W., Muller, E., Markram, H., Segev, I.: Rich cell-type-specific network topology in neocortical microcircuitry. Nature Neuroscience. 20, 1004–1013 (2017). https://doi.org/10.1038/nn.4576.
129. Takagi, K.: Information-Based Principle Induces Small-World Topology and Self-Organized Criticality in a Large Scale Brain Network. Front Comput Neurosci. 12, (2018). https://doi.org/10.3389/fncom.2018.00065.
130. Goekoop, R., de Kleijn, R.: How higher goals are constructed and collapse under stress: A hierarchical Bayesian control systems perspective. Neuroscience & Biobehavioral Reviews. 123, 257–285 (2021). https://doi.org/10.1016/j.neubiorev.2020.12.021.
131. Sporns, O.: Network attributes for segregation and integration in the human brain. Curr. Opin. Neurobiol. 23, 162–171 (2013). https://doi.org/10.1016/j.conb.2012.11.015.
132. Cohen, J.R., D'Esposito, M.: The Segregation and Integration of Distinct Brain Networks and Their Relationship to Cognition. J. Neurosci. 36, 12083–12094 (2016). https://doi.org/10.1523/JNEUROSCI.2965-15.2016.

133. Mohr, H., Wolfensteller, U., Betzel, R.F., Mišić, B., Sporns, O., Richiardi, J., Ruge, H.: Integration and segregation of large-scale brain networks during short-term task automatization. Nat Commun. 7, 13217 (2016). https://doi.org/10.1038/ncomms13217.
134. Badcock, P.B., Friston, K.J., Ramstead, M.J.D.: The hierarchically mechanistic mind: A free-energy formulation of the human psyche. Physics of Life Reviews. (2019). https://doi.org/10.1016/j.plrev.2018.10.002.
135. Bak, P., Sneppen, K.: Punctuated equilibrium and criticality in a simple model of evolution. Phys. Rev. Lett. 71, 4083–4086 (1993). https://doi.org/10.1103/PhysRevLett.71.4083.
136. Edelman, G., Gally, J.A., Baars, B.J.: Biology of consciousness. Front Psychol. 2, 4 (2011). https://doi.org/10.3389/fpsyg.2011.00004.
137. Paperin, G., Green, D.G., Sadedin, S.: Dual-phase evolution in complex adaptive systems. J R Soc Interface. 8, 609–629 (2011). https://doi.org/10.1098/rsif.2010.0719.
138. Safron, A., Klimaj, V., Hipólito, I.: On the importance of being flexible: dynamic brain networks and their potential functional significances, https://psyarxiv.com/x734w/, (2021). https://doi.org/10.31234/osf.io/x734w.
139. Safron, A.: Integrated World Modeling Theory (IWMT) Expanded: Implications for Theories of Consciousness and Artificial Intelligence, https://psyarxiv.com/rm5b2/, (2021). https://doi.org/10.31234/osf.io/rm5b2.
140. Smith, R.: Do Brains Have an Arrow of Time? Philosophy of Science. 81, 265–275 (2014). https://doi.org/10.1086/675644.
141. Wolfram, S.: A New Kind of Science. Wolfram Media (2002).
142. Friston, K.J., Wiese, W., Hobson, J.A.: Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism. Entropy. 22, 516 (2020). https://doi.org/10.3390/e22050516.
143. Doerig, A., Schurger, A., Hess, K., Herzog, M.H.: The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. Consciousness and Cognition. 72, 49–59 (2019). https://doi.org/10.1016/j.concog.2019.04.002.
144. Marshall, W., Kim, H., Walker, S.I., Tononi, G., Albantakis, L.: How causal analysis can reveal autonomy in models of biological systems. Phil. Trans. R. Soc. A. 375, 20160358 (2017). https://doi.org/10.1098/rsta.2016.0358.
145. Joslyn, null: Levels of control and closure in complex semiotic systems. Ann. N. Y. Acad. Sci. 901, 67–74 (2000).
146. Chang, A.Y.C., Biehl, M., Yu, Y., Kanai, R.: Information Closure Theory of Consciousness. arXiv:1909.13045 [q-bio]. (2019).
147. Singer, W.: Consciousness and the binding problem. Ann. N. Y. Acad. Sci. 929, 123–146 (2001).
148. Baars, B.J., Franklin, S., Ramsoy, T.Z.: Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents. Front Psychol. 4, (2013). https://doi.org/10.3389/fpsyg.2013.00200.

149. Atasoy, S., Donnelly, I., Pearson, J.: Human brain networks function in connectome-specific harmonic waves. Nature Communications. 7, 10340 (2016). https://doi.org/10.1038/ncomms10340.
150. Wu, L., Zhang, Y.: A new topological approach to the L∞-uniqueness of operators and the L1-uniqueness of Fokker–Planck equations. Journal of Functional Analysis. 241, 557–610 (2006). https://doi.org/10.1016/j.jfa.2006.04.020.
151. Carroll, S.: The Big Picture: On the Origins of Life, Meaning, and the Universe Itself. Penguin (2016).
152. Hoel, E.P., Albantakis, L., Marshall, W., Tononi, G.: Can the macro beat the micro? Integrated information across spatiotemporal scales. Neurosci Conscious. 2016, (2016). https://doi.org/10.1093/nc/niw012.
153. Albantakis, L., Marshall, W., Hoel, E., Tononi, G.: What caused what? A quantitative account of actual causation using dynamical causal networks. arXiv:1708.06716 [cs, math, stat]. (2017).
154. Hoel, E.P.: When the map is better than the territory. Entropy. 19, 188 (2017). https://doi.org/10.3390/e19050188.
155. Rocha, L.M.: Syntactic autonomy. Why there is no autonomy without symbols and how self-organizing systems might evolve them. Ann. N. Y. Acad. Sci. 901, 207–223 (2000). https://doi.org/10.1111/j.1749-6632.2000.tb06280.x.
156. Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-P., Le Van Quyen, M.: From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. Biol. Res. 36, 27–65 (2003).
157. Everhardt, A.S., Damerio, S., Zorn, J.A., Zhou, S., Domingo, N., Catalan, G., Salje, E.K.H., Chen, L.-Q., Noheda, B.: Periodicity-Doubling Cascades: Direct Observation in Ferroelastic Materials. Phys. Rev. Lett. 123, 087603 (2019). https://doi.org/10.1103/PhysRevLett.123.087603.
158. Chen, T., Gou, W., Xie, D., Xiao, T., Yi, W., Jing, J., Yan, B.: Quantum Zeno effects across a parity-time symmetry breaking transition in atomic momentum space. (2020).
159. Fruchart, M., Hanai, R., Littlewood, P.B., Vitelli, V.: Non-reciprocal phase transitions. Nature. 592, 363–369 (2021). https://doi.org/10.1038/s41586-021-03375-9.
160. Hofstadter, D.R.: Gödel, Escher, Bach: An Eternal Golden Braid. Basic Books (1979).
161. Hofstadter, D.R.: I Am a Strange Loop. Basic Books (2007).
162. Lloyd, S.: A Turing test for free will. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 370, 3597–3610 (2012). https://doi.org/10.1098/rsta.2011.0331.
163. Parr, T., Markovic, D., Kiebel, S.J., Friston, K.J.: Neuronal message passing using Mean-field, Bethe, and Marginal approximations. Sci Rep. 9, (2019). https://doi.org/10.1038/s41598-018-38246-3.
164. Madl, T., Baars, B.J., Franklin, S.: The timing of the cognitive cycle. PloS one. 6, e14803 (2011).

165. Maguire, P., Maguire, R.: Consciousness is Data Compression. undefined. (2010).
166. Tegmark, M.: Improved Measures of Integrated Information. PLoS Comput Biol. 12, (2016). https://doi.org/10.1371/journal.pcbi.1005123.
167. Maguire, P., Moser, P., Maguire, R.: Understanding Consciousness as Data Compression. Journal of Cognitive Science. 17, 63–94 (2016).
168. Metzinger, T.: The Ego Tunnel: The Science of the Mind and the Myth of the Self. Basic Books, New York (2009).
169. Limanowski, J., Friston, K.J.: 'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. Front. Psychol. 9, (2018). https://doi.org/10.3389/fpsyg.2018.00643.
170. Hoffman, D.D., Prakash, C.: Objects of consciousness. Front. Psychol. 5, (2014). https://doi.org/10.3389/fpsyg.2014.00577.
171. Kirchhoff, M., Parr, T., Palacios, E., Friston, K.J., Kiverstein, J.: The Markov blankets of life: autonomy, active inference and the free energy principle. J R Soc Interface. 15, (2018). https://doi.org/10.1098/rsif.2017.0792.
172. Dennett, D.: Consciousness Explained. Back Bay Books (1992).
173. Haun, A., Tononi, G.: Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. Entropy. 21, 1160 (2019). https://doi.org/10.3390/e21121160.
174. Sutterer, D.W., Polyn, S.M., Woodman, G.F.: α-Band activity tracks a two-dimensional spotlight of attention during spatial working memory maintenance. Journal of Neurophysiology. 125, 957–971 (2021). https://doi.org/10.1152/jn.00582.2020.

2 Appendix

2.1 Recurrent networks, universal computation, and generalized predictive coding

In describing brain function in terms of generative modeling, IWMT attempts to characterize different aspects of nervous systems in terms of principles from machine learning. Autoencoders are identified as a particularly promising framework for understanding cortical generative models due to their architectural structures reflecting core principles of FEP-AI (and AIXI). By training systems to reconstruct data while filtering (or compressing) information through dimensionality-reducing bottlenecks, this process induces the discovery of both accurate and parsimonious models of data in the service of the adaptive control of behavior [91]. IWMT's description of cortical hierarchies as consisting of "folded" autoencoders was proposed to provide a bridge between machine learning and predictive-coding models of cortical functioning. Encouraging convergence may be found in that these autoencoder-inspired models were developed without knowledge of similar proposals by others [92–96].

However, these computational principles have an older lineage predating "The Helmholtz machine" and its later elaboration in the form of variational autoencoders [97, 98]. These "Neural Sequence Chunkers" (NSCs) instantiate nearly every functional aspect of IWMT's proposed architecture for integrated world modeling (Schmidhuber, 1991), with the only potential exceptions being separate hippocampal/entorhinal analogue systems [10], and shared latent spaces structured according to the principles of geometric deep learning (however, subsequent work by Schmidhuber and colleagues has begun to move in this direction [3, 60]). NSCs consist of recurrent neural networks (RNNs) organized into a predictive hierarchy, where each RNN attempts to predict data from the level below, and where only unpredicted information is passed upwards. These predictions are realized in the form of recurrent dynamics [99, 100], whose unfolding entails a kind of search processes via competitive/cooperative attractor formation over states capable of most efficiently responding to—or resonating with [2, 101]—ascending data streams. In this way, much as in FEP-AI, predictive abilities come nearly "for free" via Hamilton's principle of least action [30, 102–104], as dynamical systems automatically minimize thermodynamic (and informational) free energy by virtue of greater flux becoming more likely to be channeled through efficient and unobstructed flow paths [105].

RNNs are Turing complete in that they are theoretically capable of performing any computational operation, partially by virtue of the fact that small clusters of signaling neurons may entail the logical operations such as NAND gates [106]. Further, considering that faster forming attractors are more likely to dominate subsequent dynamics, such systems could be considered to reflect the "Speed prior", whereby the search for more parsimonious models is heuristically realized by the selection of faster running programs [19]. Attracting states dependent upon communication along fewer synaptic connections can be thought of as minimal length descriptions, given data. In this way, each RNN constitutes an autoencoder, which when hierarchically organized

constitute an even larger autoencoding system [2]. Thus, predictive coding via "folded" (or stacked) autoencoders can be first identified in the form of NSCs as the first working deep learning system. Even more, given the recurrent message passing of NSCs (or "stacked" autoencoders), the entire system could be understood as an RNN, the significance of which will be described below.

RNN-based NSCs exhibit all of the virtues of predictive coding and more in forming increasingly predictive and sparse representations via predictive suppression of ascending signals [33, 107, 108]. However, such systems have even greater power than strictly suppressive predictive coding architectures [109–111], in that auto-associative matching/resonance allows for microcolumn-like ensemble inference as well as differential prioritization and flexible recombination of information [3, 112–114]. The auto-associative properties of RNNs afford an efficient (free-energy-minimizing) basis for generative modeling in being capable of filling-in likely patterns of data, given experience. Further, the dynamic nature of RNN-based computation is naturally well-suited for instantiating sequential representations of temporally-extended processes, which IWMT stipulates to be a necessary coherence-enabling condition for conscious world modeling. Even more, the kinds of hierarchically-organized RNNs embodied in NSCs naturally provide a basis for representing extended unfoldings via a nested hierarchy of recurrent dynamics evolving over various timescales, so affording generative modeling with temporal depth [115], and potentially counterfactual richness [59, 79, 116]. Some of these temporally-extended unfoldings include agent-information as sources of relatively invariant—but nonetheless dynamic—structure across a broad range of tasks. Given these invariances, the predictive coding emerging from NSCs will nearly automatically discover efficient encodings of structure for self and other [13], potentially facilitating meta-learning and knowledge-transfer across training epochs, including explicit self-modeling for the sake of reasoning, planning, and selecting actions [23].

In contrast to feedforward neural networks (FNNs), RNNs allow for multiple forms of open-ended learning [20, 22, 47], in that attractor dynamics may continually evolve until they discover configurations capable of predicting/compressing likely patterns of data. FNNs can be used to approximate any function [117], but this capacity for universal computation may often be more of a theoretical variety that may be difficult to achieve in practice [118], even with the enhanced processing efficiency afforded by adding layers for hierarchical pattern abstraction [119, 120]. The limitations of FNNs can potentially be understood in light of their relationship with recurrent networks. RNNs can be translated into feedforward systems through "unrolling" [48, 121], wherein the evolution of dynamics across time points can be represented as subsequent layers in an FNN hierarchy, and vice versa. Some of the success of very deep FNNs such as "open-gated Highway Nets" could potentially be explainable in terms of entailing RNNs whose dynamics are allowed to unfold over longer timescales in searching for efficient encodings. However, FNN systems are still limited relative to their RNN brethren in that the former are limited to exploring inferential spaces over a curtailed length of time, while the latter can explore without these limitations and come to represent arbitrarily extended sequences. On a higher level of abstraction, such open-endedness allows for one of the holy grails of AI in

terms of enabling transfer and meta-learning across the entire lifetime of a system [24, 122, 123].

However, this capacity for open-ended evolution is also a potential liability of RNNs, in that they may become difficult to train with backpropagation, partially due to potential non-linearities in recurrent dynamics. More specifically, training signals may either become exponentially amplified or diminished via feedback loops—or via the number of layers involved in very deep FNNs—resulting in either exploding or vanishing gradients, so rendering credit assignment intractable. These potential problems for RNNs were largely solved by the advent of "long short-term memory" systems (LSTMs) [48], wherein recurrent networks are provided flexible routing capacities for controlling information-flow. This functionality is achieved via an architecture in which RNN nodes contain nested (trainable) gate-RNNs that allow information to either be ignored, temporarily stored, or forgotten. Notably, LSTMs not only overcame a major challenge in the deployment of RNN systems, but they also bear striking resemblances to descriptions of attention in global workspace theories [38, 86]. Limited-capacity workspaces may ignore insufficiently precise/salient data (i.e., unconscious or subliminal processing), or may instead hold onto contents for a period of time (i.e., active attending, working memory, and imagination) before releasing (or forgetting) these short-term memories so that new information can be processed. Although beyond the scope of the present discussion, LSTMs may represent a multiscale, universal computational framework for realizing adaptive autoencoding/compression.

With respect to consciousness, a notable feature of deep FNNs is that their performance may be greatly enhanced via the addition of layer-skipping connections, which can be thought of as entailing RNNs with small-world connectivity [124], so allowing for a synergistic combination of integrating both local and global informational dependencies. Such deep learning systems are even more effective if these skip connections are adaptively configurable [119], so providing an even closer correspondence to the processes of dynamic evolution underlying the intelligence of RNNs [125, 126], including nervous systems [127–130]. The flexible small-worldness of these "highway nets"—and their entailed recurrent processing—has potentially strong functional correspondences with aspects of brains thought to enable workspace architectures via connectomic "rich clubs," which may be capable of supporting "dynamic cores" of re-entrant signaling, so allowing for synergistic processing via balanced integrated and segregated processing [74, 131–133]. Notably, such balanced integration and segregation via small-worldness is also a property of systems capable of both maximizing integrated information and supporting self-organized critical dynamics [1, 134], the one regime in which (generalized) evolution is possible [30, 31, 135–138]. As described elsewhere with respect to the critique of Aaronson's critique of Integrated Information Theory (IIT) [139], small-world networks can also be used for error-correction via expander graphs (or low-density parity checking codes), which enable systems to approach the Shannon limit with respect to handling noisy, lossily—and irreversibly [140]—compressed data. Speculatively, all efficiently functioning recurrent systems might instantiate turbo codes in evolving towards regimes where they

may come to efficiently resonate with (and thereby predict/compress) information from other systems in the world.

2.2 Unfolding and (potentially conscious) self-world modeling

Given this generalized predictive coding, there may be a kind of implicit intelligence at play across all persisting dynamical systems [18, 32, 103, 141, 142]. However, according to IWMT, consciousness will only be associated with systems capable of coherently modeling themselves and their interactions with the world. This is not to say that recurrence is necessarily required for the functionalities associated with consciousness [143]. However, recurrent architectures may be a practical requirement, as supra-astronomical resources may be necessary for unrolling a network the size of the human brain across even the 100s of milliseconds over which workspace dynamics unfold. Further, the supposed equivalence of feedforward and feedback processes are only demonstrated when unrolled systems are returned to initial conditions and allowed to evolve under identical circumstances [144]. These feedforward "zombie" systems tend to diverge from the functionality of their recurrent counterparts when intervened upon and are unable to repair their structure when modified. This lack of robustness and context-sensitivity means that unrolling loses one of the primary advantages of consciousness as dynamic core and temporally-extended adaptive (modeling) process, where such (integrated world) models allow organisms to flexibly handle novel situations. Further, while workspace-like processing may be achievable by feedforward systems, largescale neuronal workspaces heavily depend on recurrent dynamics unfolding over multiple scales. Perhaps we could model a single inversion of a generative model corresponding to one quale state, given a sufficiently large computational device, even if this structure might not fit within the observable universe. However, such computations would lack functional closure across moments of experience [145, 146], which would prevent consciousness from being able to evolve as a temporally-extended process of iterative Bayesian model selection.

Perhaps more fundamentally, one of the primary functions of workspaces and their realization by dynamic cores of effective connectivity may be the ability to flexibly bind information in different combinations in order to realize functional synergies [3, 147, 148]. While an FNN could theoretically achieve adaptive binding with respect to a single state estimate, this would divorce the integrating processes from its environmental couplings and historicity as an iterative process of generating inferences regarding the contents of experience, comparing these predictions against sense data, and then updating these prior expectations into posterior beliefs as priors for subsequent rounds of predictive modeling. Further, the unfolding argument does not address the issue of how it is that a network may come to be perfectly configured to reflect the temporally-extended search process by which recurrent systems come to encode (or resonate with) symmetries/harmonies of the world. Such objections notwithstanding, the issue remains unresolved as to whether an FNN-based generative model could generate experience when inverted.

This issue also speaks to the ontological status of "self-organizing harmonic modes" (SOHMs), which IWMT claims provide a functional bridge between

biophysics and phenomenology [2, 139]. Harmonic functions are places where solutions to the Laplacian are 0, indicating no net flux, which could be defined intrinsically with respect to the temporal and spatial scales over which dynamics achieve functional closure in forming self-generating resonant modes [149]. (Note: These autopoietic self-resonating/forming attractors are more commonly referred to as "nonequilibrium steady state distributions" in the FEP literature [103], which are derived using different—but possibly related [150]—maths.) However, such recursively self-interacting processes would not evolve in isolation, but would rather be influenced by other proto-system dynamics, coarse-graining themselves and each other as they form renormalization groups in negotiating the course of overall evolution within and without. Are standing wave descriptions 'real,' or is everything just a swirling flux of traveling waves? Or, are traveling waves real, or is there 'really' just an evolving set of differential equations over a vector field description for the underlying particles? Or are underlying particles real, or are there only the coherent eigenmodes of an underlying topology? Even if such an eliminative reductionism bottoms out with some true atomism, from an outside point of view we could still operate according to a form of subjective realism [151], in that once we identify phenomena of interest, then maximally efficient/explanatory partitioning into kinds might be identifiable [152–154]. Yet even then, different phenomena will be of differential 'interest' to other phenomena in different contexts evolving over different timescales.

While the preceding discussion may seem needlessly abstract, it speaks to the question as to whether we may be begging fundamental questions in trying to identify sufficient physical substrates of consciousness, and also speaks to the boundary problem of which systems can and cannot be considered to entail subjective experience. More concretely, do unrolled SOHMs also entail joint marginals over synchronized subnetworks, some of which IWMT claims to be the computational substrate of consciousness? Based on the inter-translatability of RNNs and FNNs described above, this question appears to be necessarily answered in the affirmative. However, if the functional closure underlying these synchronization manifolds require temporally-extended processes that recursively alter themselves [155, 156], then it may be the case that this kind of autopoietic ouroboros cannot be represented via geometries lacking such entanglement. Speculatively (and well-beyond the technical expertise of this author), Bayesian model selection via SOHMs might constitute kinds of self-forming (and self-regenerating) "time crystals" [157–159], whose symmetry-breaking might provide a principled reason to privilege recurrent systems as physical and computational substrates for consciousness. If this were found to be the case, then we may find yet another reason to describe consciousness as a kind of "strange loop" [160–162], above and beyond the seeming and actual paradoxes involved in explicit self-reference.

This kind of self-entanglement would render SOHMs opaque to external systems lacking the cypher of the self-generative processes realizing those particular topologies [155]. Hence, we may have another way of understanding marginalization/renormalization with respect to inter-SOHM information flows as they exchange messages in the form of sufficient statistics [163], while also maintaining

degrees of independent evolution (cf. mean field approximation) over the course of cognitive cycles [164]. These self-generating entanglements could further speak to interpretations of IIT in which quale states correspond to maximal compressions of experience [165]. In evaluating the integrated information of systems according to past and future combinatorics entailed by minimally impactful graph cuts [166], we may be describing systems capable of encoding data with maximal efficiency [167], in terms of possessing maximum capacities for information-processing via supporting "differences that make a difference." Indeed, adaptively configurable skip connections in FNNs could potentially be understood as entailing a search process for discovering these maximally efficient codes via the entanglements provided by combining both short- and long-range informational dependencies in a small-world (and self-organized-critical) regime of compression/inference. A system experiencing maximal alterations in the face of minimal perturbations would have maximal impenetrability when observed from without, yet accompanied by maximal informational sensitivity when viewed from within.

If we think of minds as systems of interacting SOHMs, then this lack of epistemic penetration could potentially be related to notions of phenomenal transparency (e.g. not being able to inspect processes by which inspection is realized, perhaps especially if involving highly practiced, and thereby efficient and sparse patterns of neural signaling) [168, 169], and perhaps "user interface" theories of consciousness [170]. Intriguingly, maximal compressions have also been used as conceptualizations of the event horizons of black holes, for which corresponding holographic principles have been adduced in terms of internal information being projected onto 2D topologies. With respect to the FEP, it is also notable that singularities and Markov blankets have been interpreted as both points of epistemic boundaries as well as maximal thermal reservoirs [171]. Even more speculatively, such holography could even help explain how 3D perception could be derived from 2D sensory arrays, and perhaps also experienced this way in the form of the precuneus acting as a basis for visuospatial awareness and kind of "Cartesian theater" [172–174]. As described elsewhere [9], this structure may constitute a kind of graph neural network (GNN), utilizing the locality of recurrent message passing over grid-like representational geometries for generating sufficiently informative projections on timescales proportional to the closure of action-perception cycles [1, 2].

However, in exploring the potential conscious status of recurrent systems, we would still do well to consider the higher-level functional desiderata for consciousness as a process of integrated world modeling. More specifically, if particular forms of coherence are required for generating experience, then we may require hybrid architectures with specialized subsystems capable of supporting particular kinds of learning. This is an ongoing area of research for IWMT, with specific focus on the hippocampal/entorhinal-system as providing fundamental bases for higher-order cognition [10]—understood as a kind of generalized search/navigation process—including with respect to reverse engineering such functions in attempting to design (and/or grow) intelligent machines [2].