Integrated World Modeling Theory (IWMT) Implemented: Towards Reverse Engineering Consciousness with the Free Energy Principle and Active Inference

Adam Safron[1,2]

[1] Indiana University, Kinsey Institute, Bloomington IN 47404, USA
[2] Indiana University, Cognitive Science Program, Bloomington IN 47404, USA
asafron@gmail.com

Abstract. Integrated World Modeling Theory (IWMT) is a synthetic model that attempts to unify theories of consciousness within the Free Energy Principle and Active Inference framework, with particular emphasis on Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT). IWMT further suggests predictive processing in sensory hierarchies may be well-modeled as (folded, sparse, partially disentangled) variational autoencoders, with beliefs discretely-updated via the formation of synchronous complexes—as self-organizing harmonic modes (SOHMs)—potentially entailing maximal a posteriori (MAP) estimation via turbo coding. In this account, alpha-synchronized SOHMs across posterior cortices may constitute the kinds of maximal complexes described by IIT, as well as samples (or MAP estimates) from multimodal shared latent space, organized according to egocentric reference frames, entailing phenomenal consciousness as mid-level perceptual inference. When these posterior SOHMs couple with frontal complexes, this may enable various forms of conscious access as a kind of mental act(ive inference), affording higher order cognition/control, including the kinds of attentional/intentional processing and reportability described by GNWT. Across this autoencoding heterarchy, intermediate-level beliefs may be organized into spatiotemporal trajectories by the entorhinal/hippocampal system, so affording episodic memory, counterfactual imaginings, and planning.

Keywords: World models, Global workspaces, Integrated information, Autoencoders, Turbo codes, Phenomenal consciousness, Conscious access

"*The formal distinction between the FEP and IIT is that the free energy principle is articulated in terms of probabilistic beliefs about some (external) thing, while integrated information theory deals with probability distributions over the states of some system… On the other hand, both the FEP and IIT can be cast in terms of information theory and in particular functionals (e.g., variational free energy and 'phi'). Furthermore, they both rest upon partitions (e.g., Markov blankets that separate internal from external states and complexes that constitute conscious entities and can be distinguished from other entities). This speaks to the possibility of, at least, numerical analyses that show that minimising variational free energy maximises 'phi' and vice versa… This supports the (speculative) hypothesis that adding further constraints on generative models—entailed by systems possessing a Markov blanket—might enable us to say which systems are conscious, and which are not.*"-Friston et al. [1], Sentience and the Origins of Consciousness: From Cartesian Duality to Markovian Monism

1 Integrated World Modeling Theory (IWMT) summarized: Combining the Free Energy Principle and Active Inference (FEP-AI) framework with Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT)

The Hard problem of consciousness asks, how can it be that there is "something that it is like" to be a physical system [2]? This is usually distinguished from the "easy problems" of addressing why different biophysical and computational phenomena correspond to different qualities of experience. IWMT attempts to address consciousness' enduring problems with the Free Energy Principle [3] and Active Inference [4] (FEP-AI) framework. FEP-AI begins with the understanding that persisting systems must regulate environmental exchanges and prevent entropic accumulation (cf. Good Regulator Theorem from cybernetics) [5]. In this view, minds and brains are predictive controllers for autonomous systems, where action-driven perception is realized via probabilistic inference. FEP-AI has been used to address consciousness in multiple ways [1, 6], with IWMT representing one such attempt. Below I briefly summarize the major claims of IWMT via modified excerpts from the original publication of the theory in *Frontiers in Artificial Intelligence* [7], as well as the accompanying preprint, "IWMT Revisited" [8]. Please see these longer works for further discussion.

IWMT's primary claims are as follows *(originally published in [7])*:

1. Basic phenomenal consciousness is what it is like to be the functioning of a probabilistic generative model for the sensorium of an embodied–embedded agent [9].
2. Higher order and access consciousness are made possible when this information can be integrated into a world model with spatial, temporal, and causal coherence. Here, coherence is broadly understood as sufficient consistency to enable functional closure and semiotics/sense-making [10–12]. That is, for there to be the experience of a world, the things that constitute that world must be able to be situated and contrasted with other things in some kind of space, with relative changes constituting time, and with regularities of change constituting cause. These may also be preconditions for basic phenomenality, especially if consciousness (as subjectivity) requires an experiencing subject with a particular point of view on the world [13–15].
3. Conscious access, or awareness/knowledge of experience—and possibly phenomenal consciousness—likely requires generative processes capable of counterfactual modeling with respect to selfhood and self-generated actions [16, 17].

IIT begins with considering the preconditions for systems to exist intrinsically from their own perspectives, as is observed with the privately-experienced 1st person ontology of consciousness as subjectivity [18]. IIT speaks to the Hard problem by grounding itself in phenomenological axioms, and then goes on to postulate mechanisms that could realize such properties. While IWMT focuses on explaining the functional, algorithmic, and implementational properties that may give rise to consciousness—or experience as a subjective point of view—it also considers ways in which FEP-AI and IIT can be combined as general systems theories and models of causal emergence [19–21]. In brief, IWMT argues that complexes of integrated information (as irreducible self-cause-effect power) are also Markov-blanket-bound networks of effective connectivity associated with high marginal likelihoods and capacity for "self-evidencing" [22].

GNWT has a more restricted scope than IIT and FEP-AI, instead focusing on the properties of computational systems that could realize the functions of consciousness as a means of globally integrating and broadcasting information from otherwise disconnected mental systems [23]. GNWT suggests that workspaces help to select particular interpretations of events, potentially understandable as Bayesian model selection [23, 24], which is highly compatible with IWMT. However, IWMT also potentially differs from the theories it attempts to combine, suggesting that complexes of integrated information and global workspaces only entail subjective experience when applied to systems capable of functioning as Bayesian belief networks and cybernetic controllers for embodied agents [25]. That is, IWMT argues that integration and widespread availability of information are necessary, but not sufficient, preconditions for enabling consciousness. Specifically, *IWMT claims that consciousness is what integrated world-modeling is like, when generative processes are capable of jointly integrating information into models with coherence with respect to space, time, and cause for systems and their relationships with their environments.* These coherences are stipulated to be required for situating modeled entities relative to each other with specific properties, without which there would be no means of generating an experienceable world. IWMT further introduces a mechanism for generating complexes of integrated information and global workspaces via (Markov-blanket-bound) meta-stable synchronous complexes—or "self-organizing harmonic modes" (SOHMs)—wherein synchrony both emerges from and facilitates the integration of information via "communication-through-coherence" [26, 27]. IWMT further suggests that the stream of experience (Figure 1) is constituted by a series of SOHM-formation events, computationally understood as entailing loopy belief propagation, so generating joint posterior distributions (or maximal estimates derived thereof) over sensoriums of embodied agents as they engage with the environments in which they are embedded.

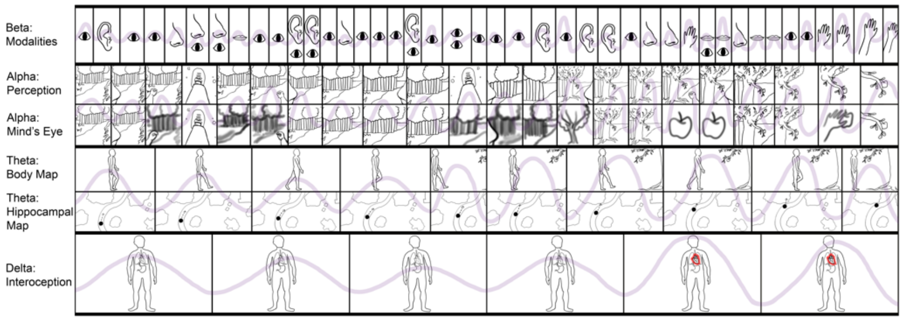

Fig. 1. Depiction of experience with components mapped onto EEG frequency bands.

While parallels can be identified between GNWT and IIT, present discussions and ongoing adversarial collaborations emphasize their differences, such as IIT's claim that consciousness is primarily located in a "posterior hot zone" [28], and GNWT's claim that consciousness requires frontal-lobe engagement for realizing "global availability" of information. IWMT considers both positions to be accurate, but with respect to phenomenal consciousness and conscious access (and other higher-order forms of conscious experience), respectively. That is, frontal cortices are likely required for mental processes such as manipulating and reporting on the contents of consciousness, and so modifying these phenomena in qualitatively/functionally important ways. However, the stream of experience itself may always be generated within hierarchies centered on the posterior cortices, as described in greater detail below.

2 Integrated World Modeling Theory (IWMT) implemented

2.1 Mechanisms of predictive processing; folded variational autoencoders (VAEs) and self-organizing harmonic modes (SOHMs)

IWMT understands cortex using principles from predictive coding [29–31], specifically viewing cortical hierarchies as analogous to VAEs [32, 33], where encoding and generative decoding networks have been folded over at their reduced-dimensionality bottlenecks such that corresponding hierarchical levels are aligned. In this view, hierarchies of superficial pyramidal neurons constitute encoding networks, whose bottom-up observations would be continually suppressed (or "explained away") by predictions from hierarchies of deep pyramidal neurons (and thalamic relays) [34, 35], with only prediction-errors being passed upwards. This is similar to other recent proposals [16], except beliefs are specifically communicated and updated via synchronous dynamics, wherein prediction-errors may be quantized via fast gamma-synchronized complexes [36, 37], and where predictions may take the form of a nested hierarchy of more slowly evolving synchronization manifolds, so affording hierarchical modeling of spatiotemporal events in the world [38] (Figure 1). More specifically, self-organizing harmonic modes (SOHMs) are suggested to implement loopy belief propagation for approximate inference (cf. turbo coding) [39–41], as well as marginalization over synchronized subnetworks, so instantiating marginal message passing regimes [42].

Combined with mechanisms of divisive normalization and spike-timing dependent plasticity [43–45], this predictive coding setup should induce increasingly sparse connectivity with experience, with all of the functional benefits sparsity provides [46]. These mechanisms (and entailed algorithms) may converge on near-optimal training protocols [47]. With respect to suggestions that the brain may indirectly realize backprop-like computations [48] (Appendix 2), these models view cortical hierarchies as "stacked autoencoders" [49], as opposed to being constituted by a single (folded) VAE. These interpretations of neural computation may be non-mutually exclusive, depending on the granularity with which relevant phenomena evolve. That is, we could think of separate VAEs for each cortical region (e.g. V1, V2, V4, IT), or each cortical macrocolumn [50, 51], and perhaps even each cortical minicolumn. Depending on the timescales over which we are evaluating the system, we might coarse-grain differently [19], with consciousness representing a single joint belief at the broadest level of integration, with the perceptual heterarchy considered as a single VAE. However, a more fine-grained analysis might allow for further factorization, where component subnetworks could be viewed as entailing separate VAEs, with separate Bayesian beliefs. In this view, the cortical heterarchy could be viewed as a single VAE (composed of nested VAEs), as well as a single autoregressive model, where latent beliefs between various VAEs are bound together via synchronous activity, potentially entailing normalizing flows across coupled latent space dynamics [52, 53].

2.2 A model of episodic memory and imagination

With respect to consciousness, SOHM-formation involving deep pyramidal neurons is suggested to correspond to both “ignition” events as described by GNWT [54], as well as implementation of semi-stochastic sampling from the latent space of VAEs (cf. the “reparameterization trick”) [55], including via latent (work)spaces shared by multiple VAEs. If these samples are sequentially orchestrated according to spatiotemporal trajectories of the entorhinal/hippocampal system [56], this may generate a coherent stream of experience. However, coherent sequence transitions between quale states may also potentially be realizable even in individuals without functioning medial temporal lobes, if prior histories of experience allow frontal lobes to enable coherent action-selection and action-driven perception—including with respect to mental acts— in which posterior dynamics may be driven either through overt enaction or via efference copies accompanying covert partial deployment of “forward models” [25].

In this view of the brain in terms of machine learning architectures, the hippocampal complex could be thought of as the top of the cortical heterarchy [57, 58] and spatiotemporally-organized memory register [59]. IWMT suggests this spatial and temporal organization may be essential for coherent world modeling. With respect to grid/place cells of the entorhinal/hippocampal system [60], this organization appears to take the form of 2D trajectories through space, wherein organisms situate themselves according to a kind of simultaneous localization and mapping via Kalman filtering [61]. Anatomically speaking, this dynamic (and volatile) memory system has particularly strong bidirectional linkages with deeper portions of cortical generative models (i.e., reduced-dimensionality latent feature spaces), so being capable of both storing information and shaping activity for these core auto-associative networks. Because of the predictive coding setup—and biasing via neuromodulatory value signals [62, 63]—only maximally informative, novel, unexplained observations will tend to be stored in this spatiotemporally organized memory register. Indeed, this may be one of the primary functions of the hippocampus: temporarily storing information that could not be predicted elsewhere, and then using replay to train relevant subnetworks to be more successfully predictive of likely observations.

As the hippocampus—and cortical systems with which it couples via “big loop recurrence” [64, 65]—re-instantiates trajectories of the organism through space, pointers to prediction errors will be sequentially activated, with the generative model inferring a more complete sensorium based on its training from a lifetime of experience. Computationally speaking, this setup would correspond to a Kalman variational auto-encoder [66]. Experientially speaking, this integration of organismic spatiotemporal trajectories with auto-associative filling-in could provide not just a basis for forming episodic memories, but also the imagination of novel scenarios [67, 68]. Importantly, memories and imaginings can be generated by cortex on its own—given a lifetime of experience with a functioning entorhinal/hippocampal system—but medial temporal lobe involvement appears to be required for these dynamics to be shaped in novel directions that break free of past experience [69–72]. The hippocampal system may further allow for contrasting of anticipated and present estimated states in the process of orchestrating goal-oriented behavior [25, 73] (Appendix 1).

2.3 Brains as hybrid machine learning architectures

Figure 2 provides a depiction of the human brain in terms of phenomenological correspondences, as well as Marr's computational (or functional), algorithmic, and implementational levels of analysis [74]. On the computational level, various brain functions are identified according to their particular modal character, either with respect to perception (both unconscious and conscious) or action (both unconscious and potentially conscious, via perceptual generative models). On the algorithmic level, these functions are mapped onto variants of machine learning architectures—e.g. autoencoders and generative adversarial networks (Appendix 2), graph neural networks, recurrent reservoirs and liquid state machines—organized according to their potential realization by various systems in the brain. On the implementational level, realizations of algorithmic processes are depicted as corresponding to flows of activity and interactions between neuronal populations, canalized by the formation of SOHMs as metastable synchronous complexes. While the language of predictive processing is used here to help provide bridges to the algorithmic level, descriptions such as vector/tensor fields and attracting manifolds could have alternatively been used in order to remain agnostic as to which algorithms may be entailed by physical dynamics.

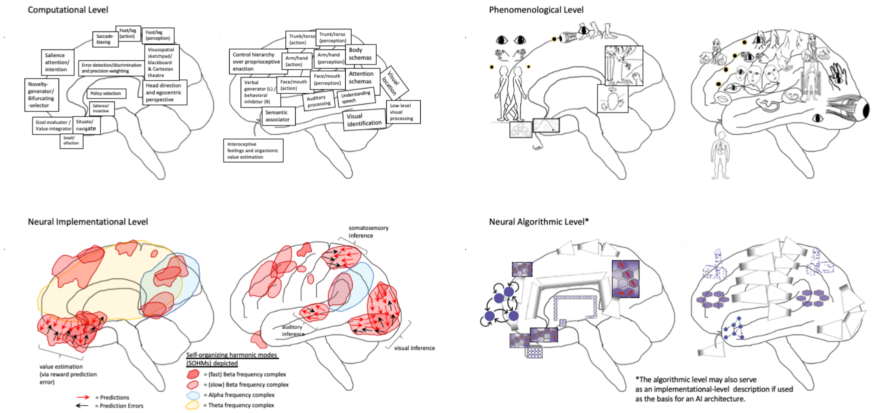
Computational Level
Saccade-biasing
Foot/leg (action)
Foot/leg (perception)
Salience attention/ intention
Error detection/discrimination and precision-weighting
Visuospatial sketchpad/ blackboard & Cartesian theatre
Novelty-generator/ Bifurcating -selector
Policy selection
Head direction and egocentric perspective
Salience/ incentive
Goal evaluator / Value-integrator
Situate/ navigate
Smell/ olfaction
Trunk/torso (action)
Trunk/torso (perception)
Control hierarchy over proprioceptive enaction
Arm/hand (action)
Arm/hand (perception)
Body schemas
Verbal generator (L) / behavioral inhibitor (R)
Face/mouth (action)
Face/mouth (perception)
Attention schemas
Visual location
Auditory processing
Understanding speech
Semantic associator
Visual identification
Low-level visual processing
Interoceptive feelings and organismic value estimation
Phenomenological Level
Neural Implementational Level
value estimation (via reward prediction error)
= Predictions
= Prediction Errors
Self-organizing harmonic modes (SOHMs) depicted
= (fast) Beta frequency complex
= (slow) Beta frequency complex
= Alpha frequency complex
= Theta frequency complex
somatosensory inference
auditory inference
visual inference
Neural Algorithmic Level*
*The algorithmic level may also serve as an implementational-level description if used as the basis for an AI architecture.

Fig. 2. Depiction of the human brain in terms of phenomenological correspondences, as well as computational (or functional), algorithmic, and implementational levels of analysis.

A phenomenological level is specified to provide mappings between consciousness and these complementary/supervenient levels of analysis. These modal depictions are meant to connotate the inherently embodied nature of experience, but not all images are meant to correspond to the generation of consciousness. That is, it may be the case that consciousness is solely generated by posterior hierarchies centered on the precuneus, lateral parietal cortices, and temporoparietal junction (TPJ) as respective visuospatial (cf. consciousness as projective geometric modeling) [13, 14], somatic (cf. grounded cognition and intermediate level theory) [75–77], and attentional/intentional phenomenology (cf. Attention Schema Theory) [78].

Graph neural networks (GNNs) are identified as a potentially important machine learning architectural principle [79], largely due to their efficiency in emulating physical processes [80–82], and also because the message passing protocols during training and inference may have correspondences with loopy belief propagation and turbo codes suggested by IWMT. Further, grid graphs—potentially hexagonally organized, possibly corresponding to cortical macrocolumns [50], with nested microcolumns that may also be organized as hexagonal grid GNNs, or "capsule networks") [83]—are adduced for areas contributing to quasi-Cartesian spatial modeling (and potentially experience) [84, 85], including the posterior medial cortices, dorsomedial and ventromedial prefrontal cortices, and the hippocampal complex. With respect to AI systems, such representations could be used to implement not just modeling of external spaces, but of consciousness as internal space (or blackboard), which could potentially be leveraged for reasoning processes with correspondences to category theory, analogy making via structured representations, and possibly causal inference.

Neuroimaging evidence suggests these grids may be dynamically coupled in various ways [68], with these aspects of higher-order cognition being understood as a kind of generalized navigation/search process [86, 87]. A further GNN is speculatively adduced in parietal cortices as a mesh grid placed on top of a transformed representation of the primary sensorimotor homunculus (cf. body schemas for the sake of efficient motor control/inference), which is here suggested to have some correspondence/scaling to the body as felt from within, but which may potentially be further morphed to better correspond with externally viewed embodiments (potentially both resulting from and enabling "mirroring" with the bodies of other agents for the sake of coordination and inference) [88]. This partial translation into an allocentric coordinate system is suggested to provide more effective couplings (or information-sharing) with semi-topographically organized representations in posterior medial cortices. The TPJ is depicted as containing a ring-shaped GNN to reflect a further level of abstraction and hierarchical control over action-oriented body schemas—which may influence more somatic-like geometries—functionally entailing vectors/tensors over attentional/intentional processes [89].

Frontal homologues to posterior GNNs are also depicted, which may provide a variety of higher-order modeling abilities, including epistemic access for extended/distributed self-processes and intentional control mechanisms. These higher-order functionalities may be achieved via frontal cortices being more capable of temporally-extended generative modeling [90], and also potentially by virtue of being located further from primary sensory cortices, so affording ("counterfactually rich") dynamics that are more decoupled from immediate sensorimotor contingencies. Further, these frontal control hierarchies afford multi-scale goal-oriented behavior via bidirectional effective connectivity with the basal ganglia (i.e., winner-take-all dynamics and facilitation of sequential operations) and canalization via diffuse neuromodulator nuclei of the brainstem (i.e., implicit policies and value signals) [91–95]. Finally, the frontal pole is described as a highly non-linear recurrent system capable of shaping overall activity via bifurcating capacities [96, 97]—with potentially astronomical combinatorics—providing sources of novelty and rapid adaptation via situation-specific attractor dynamics.

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 [8]. Speculatively, functional analogues for ring-shaped GNN salience/relevance maps may potentially be found in the central complexes of insects and the tectums of all vertebrates [98], although it is unclear whether those structures would be associated with any kind of subjective experience. Even more speculatively, if these functional mappings were realized in a human-mimetic, neuromorphic AI, then it may have both flexible general intelligence and consciousness. In this way, this figure can be considered to be a sort of pseudocode for potentially conscious (partially human-interpretable) AGI with "System 2" capacities [99–101].

2.4 Conclusion: Functions of basic phenomenal consciousness?

According to IWMT, whenever we have self-organizing harmonic modes (SOHMs), then we also have entailed joint marginal probability distributions (where synchrony selects or discretely updates Bayesian beliefs), some of which may entail consciousness. Functionally speaking, potentially experience-entailing SOHMs—as Markov-blanket-bound subnetworks of effective connectivity and complexes with high integrated information, functioning as workspaces—over coupled visuospatial, attentional/intentional, and somatic hierarchies could provide holistic discriminations between different classes of events in ways that would greatly facilitate coherent action selection and credit assignment. That is, a series of coherently estimated system-world states (even without higher-order awareness or explicit/reflexive knowledge) would be extremely adaptive if it could generate these joint posteriors (or MAP estimates derived thereof) on timescales allowing this information to shape (and be shaped by) action-perception cycles. Since there should be substantial auto-associative linkages across visuospatial, attentional/intentional, and somatic modalities, then the consistency of this mutual information may accelerate the formation of SOHMs, such that beliefs can be updated quickly and coherently enough to have actual organismic-semiotic content (i.e., relevance for the organism and its environment). Further, reentrant signaling across different sources of data may provide a) inferential synergy via knowledge fusion from combining modalities, b) enhanced transfer learning and representational invariance via perspectival diversity (i.e., flexible representation from multiple modalities), and c) sensitivity to higher-order relational information, potentially including causal and contextual factors identified by comparing and contrasting constancies/inconstancies across modalities [102]. Even more, these sources of (mutual) information have natural correspondences with subjectivity in terms of providing a particular point of view on a world, centered on the experience of having/being a body.

Thus, when we identify the kinds of information that could enable adaptive functional synergy in 'processing' sensory data, it becomes somewhat less surprising that there might be "something that it is like." However, such inferential dynamics might require a multi-level hierarchy, with a higher (or deeper) inner-loop capable of iteratively forming and vitiating attracting states [103], so instantiating a kind of "dual phase evolution" [104]. A shallow hierarchy might be overly enslaved to immediate environmental couplings/contingencies [105], and would potentially constitute unconscious inference, with consciousness-entailing states never being generated on any level of abstraction. However, the precise functional boundaries of phenomenal consciousness remain unclear, and is a direction for future work for IWMT.

3 Appendices

3.1 Appendix 1: A model of goal-oriented behavior with hippocampal orchestration

Figure 3 depicts memory and planning (as inference) via predictive processing, orchestrated via the spatiotemporal trajectories of the entorhinal/hippocampal system. In this model, precision-weighting/gain-amplification takes place via "big loop recurrence" with the frontal lobes [64], with the more specific suggestion that selection/biasing of policies over forward models cause efference copies to be projected to posterior generative models. In line with recent proposals [106], the hippocampus can operate with either "predictive-suppressive" or "fictive prediction error" modes, which are here suggested to correspond to degree of coupling with respective posterior vs. frontal cortices, with the former corresponding to direct suppression of observations, and the latter facilitating the 'reinstatement' of memories, and novel imaginings for the sake of planning and causal reasoning [68, 107]. This frontal coupling is hypothesized to be a source of "successor representations" (i.e., population vectors forming predictive anticipatory sweeps of where the organism is likely to go next) via integration of likely policies and action models (via dorsal prefrontal cortex) and evaluations of likely outcomes (via ventral prefrontal cortex).

In this model of generalized navigation, the hippocampal system iteratively contrasts predictive representations with (either sensory-coupled or imaginative) present state-estimates (from coupling with posterior cortices), where prediction-errors both modify future paths, and also allow for encoding of novel information within likely (generalized) spatiotemporal trajectories, given the meta-prior (or inductive bias) that organisms are likely to be pursuing valued goals as they navigate/forage-through physical and conceptual spaces. This alternation may occur at different phases of theta oscillations [108], so affording iterative contrasting of desired and current-estimated states, so canalizing neural activity for goal realization [109], potentially including the formation of complex action sequences (either physical or virtual) via (conscious) back-chaining from potential desired states (i.e., goals) to presently-inferred realities [25]. Theoretically, this kind of iterated contrasting mechanism may also provide a source of high-level analogical (and potentially causal) reasoning [107, 110–112].

By orchestrating alternating counterfactual simulations [73], the hippocampal system may allow for evaluation of possible futures, biased on a moment-to-moment basis by integrated (spatialized) value representations from ventromedial prefrontal cortex [113], and also via "as-if-body loops" with interoceptive hierarchies [114, 115]. In this view of thought as generalized navigation, alternating exploration/sampling of counterfactuals could also be understood as implementing Markov chain Monte Carlo tree search over policy/value space for planning [109, 116]. Theoretically, similar processes could be involved in generating particular actions, if visualization of acts is accompanied by a critical mass of model-evidence (as recurrent activity) accumulating in interoceptive/salience hierarchies [117]. Speculatively, this threshold crossing (or phase transition) may represent a source of readiness potentials [118–120], potentially understood as a kind of "ignition" event and driver of workspace dynamics (understood as high-level Bayesian model selection), corresponding to explosive percolation, triggered by the accumulation of recurrent-activity/model-evidence within upper levels of frontoparietal control hierarchies for enacting/inferring a particular (virtual or physical) proprioceptive state, or pose [121].

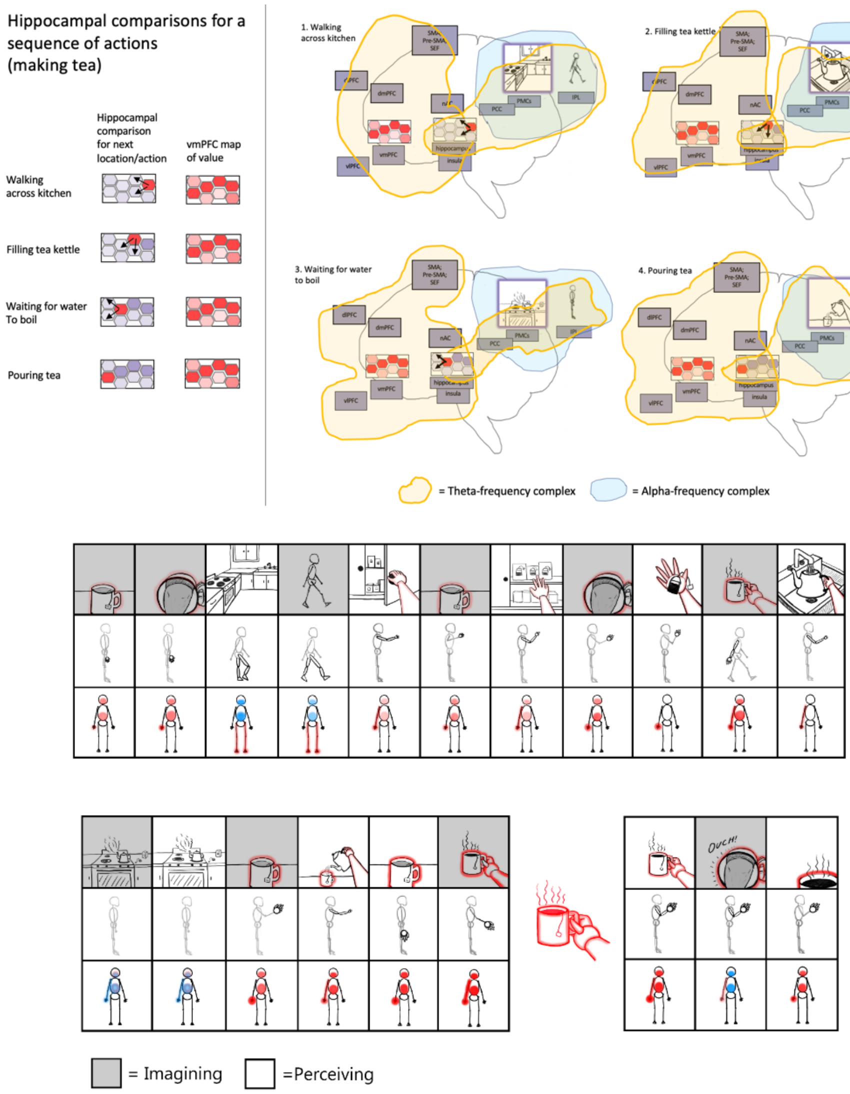
Hippocampal comparisons for a sequence of actions (making tea)
Hippocampal comparison for next location/action
vmPFC map of value
Walking across kitchen
Filling tea kettle
Waiting for water To boil
Pouring tea
1. Walking across kitchen
2. Filling tea kettle
3. Waiting for water to boil
4. Pouring tea
SMA; Pre-SMA; SEF
dlPFC
dmPFC
nAC
vmPFC
vlPFC
hippocampus
insula
PCC
PMCs
IPL
= Theta-frequency complex
= Alpha-frequency complex
OUCH!
= Imagining
=Perceiving

Fig. 3. Hippocampally-orchestrated imaginative planning and action selection via generalized navigation/search.

3.2 Appendix 2: The VAE-GAN brain?

Gershman [122] has presented an intriguing account of neural functioning in terms of a powerful class of generative models known as generative adversarial networks (GANs). GANs have many similar use cases to variational-encoders (VAEs) and can even be used in combination for enhanced training as in the case of VAE-GANs [123]. In Gershman's proposal, sensory cortices act as generators which are trained via Turing learning with frontal cortex, which functions as a discriminator and source of (higher-order) consciousness.

IWMT, in contrast, suggests that predictive coding can be understood as generating (basic phenomenal) consciousness via folded VAEs in the ways described above. From this perspective, the ascending and descending streams for each modality constitute respective encoding and generative decoding networks. This is not necessarily inconsistent with Gershman's proposal, in that a sensory hierarchy as a whole can be viewed as a generative network, which relative to the entire brain may provide a VAE-GAN setup. Alternatively, the ascending stream could be interpreted as acting as a discriminator in the GAN sense, in that it is attempting to evaluate the degree to which the descending stream generates veridical images. In this view, folded autoencoders might also be understood as folded GANs, but with folds taking place at output layers of generative decoders and input layers of discriminative encoders. The ascending stream is well-poised to serve this kind of discriminative function in terms of being more directly in touch with the ground truth of sensation and the generative processes of the world, which are the ultimate referents and selection criteria for neural dynamics. This is somewhat different from Gershman's proposal, in that consciousness (as experience) would correspond to generative processes in posterior sensory areas (including multi-modal association cortices), trained via embodied-embedded interaction with the world, with frontal cortices functioning as an elaboration of the generative process in multiple ways, including conscious access via the stabilization and alteration of dynamics within posterior networks, and also via simulated actions and inter-temporal modeling [124].

However, frontal cortex could also be viewed as serving a discriminator function in terms of attentional biasing based on the reliability of information (i.e., precision weighting), mechanistically achieved by altering the gain on excitatory activity from ascending ('discriminative') encoding networks. Thus, frontal cortex could provide a discriminatory function via tuning the sensitivity of ascending perceptual streams. Other non-mutually exclusive possibilities could also be envisioned for a discriminator-like role for frontal cortices:

1. Comparison with memory: "Is what I am perceiving consistent with what I have previously experienced?" This might be particularly important early in development for teaching patterns of attention that promote perceptual coherence.
2. Comparison with causal reasoning: "Is what I am perceiving consistent with what is plausible?" This is closer to Gershman's proposal, wherein failing to establish this discriminatory capacity could increase the probability of delusions and possibly lowered hallucination thresholds in some conditions.
3. Comparison with goal-attainment (a combination of 1 and 2): "Is what I am perceiving consistent with my normal predictions of working towards valued goals?" This could have the effect of adaptively shaping conscious states in alignment with

personal (and ultimately organismic) value. According to FEP-AI, all discriminator-like functionality may represent special cases of this highest-level objective: to reduce uncertainty (or accumulate model evidence) with respect to organismic value via perceptual and active inference. Mechanistically, cingulate cortices may have the greatest contributions to generating discrimination signals with respect to overall value [125–130], both through the integrative properties of the cingulum bundle [131, 132], as well as via close couplings with allostatic-organismic interoceptive insular hierarchies [133].

In all of these cases, frontal cortices (broadly construed to include the anterior cingulate) could be viewed as being in an adversarial (but ultimately cooperative) relationship with sensory hierarchies, whose recognition densities would optimize for minimizing perceptual prediction error (i.e., what is likely to be, given data), and where frontally-informed generative densities would optimize for future-oriented (counterfactual) adaptive policy selection (i.e., what ought to be, given prior preferences). In these ways, action and perceptual hierarchies would compete with respect to the ongoing minimization of free energy, while at the same time being completely interdependent for overall adaptive functioning, with both competitive and cooperative dynamics being required for adaptively navigating the world via action-perception cycles. An interesting hybrid of competitive and cooperative dynamics may be found in “learning to learn” via creative imagination and play (including self-play), in which learners may specifically try to maximize surprise/information-gain [134].

With respect to frontal predictions, these may be productively viewed with a 3-fold factorization:

1. A ventral portion representing affectively-weighted sensory outcomes associated with various actions.
2. A dorsal portion representing forward models for enacting sequences that bring about desirable outcomes.
3. A recurrent anterior pole portion that mediates between affect and action selection via its evolving/bifurcating/non-linear attractor dynamics [96, 135, 136].

(1) and (2) would be frontal analogues to the “what” and “where” pathways for vision [90, 137]—with macroscale connectivity reflecting these functional relationships—except here we are dealing with (1) what-where (via coupling with the hippocampal complex) and (2) how-where (via coupling with the parietal lobes). Taken together (which is how these systems are likely to work under most circumstances), these different parts of frontal cortices could all be understood in a unified sense as implementing policy selection via predictions and precision-weighting.

In Gershman’s proposal, he further suggests that predictive coding can be viewed as an efficient way of passing predictions up the cortical hierarchy while removing redundant information. This is consistent with proposals in which the descending stream is interpreted as constituting a means for communicating the backwards propagation of error signals to apical dendrites in cortical layer 1 [138]. Although this (potentially insightfully) inverts the way predictive coding is normally understood, with prediction errors being communicated via the ascending stream, these accounts could potentially be reconciled if we understand perception as involving a circular-causal process of iterative Bayesian model selection. When we consider the capacity for looping effects in networks on the scale of nervous systems—for even the largest deep learning systems, the number of parameters is dwarfed (for now) by those found in a cubic centimeter of cortex—with potentially multiple levels of qualitatively different ‘beliefs’ (e.g. unconscious sensorimotor, conscious embodiment, and implicit schemas), then it can be difficult to straightforwardly interpret the flow of inference in terms of a clear distinction between predictions and prediction errors. Indeed, hierarchical predictive processing can be viewed as converging on optimal backprop-like functionality via proposals such as “target propagation” and “natural gradient descent” [47, 139]. However, we would also do well to not be overly ecumenical with respect to this potential reconciliation, as more classical accounts of predictive coding induce sparsity on multiple levels, so creating many highly desirable computational properties such as energy efficiency, robustness, and sensitivity to coincidence detection [46, 140]. As such, framing the descending stream as a backpropagation signal may be an account that is both misleading and impoverished with respect to biological realities.

In terms of the potential complexity of cortical generative models, we may want to think of at least three coupled systems that are ultimately integrated as parts of a unified control hierarchy, but which can temporally evolve independently:

1. Unconscious/preconscious lower-level sensorimotor hierarchies with fast fine-grained dynamics for coupling with the environment [96].
2. Conscious mid-level sensorimotor representations with more coarse-grained spatial and temporal dynamics [75].
3. Higher-level abstract re-representations over recognition and generative densities, with unconscious/preconscious dynamics [141], and which may bidirectionally couple with lower and middle levels.

In this way, we could potentially dissect the brain into multiple competing and cooperating generative models, whose synergistic interactions may be productively considered as implementing GAN-type setups. Very speculatively, it may even be the case that perception-action cycles, hemispheric interactions, and interpersonal communication could all be understood as implementing CycleGAN-like dynamics. That is, to what extent could relationships between hemispheres (or between individuals) be analogous to a paired GAN setup, where each system may evaluate the output of the other, so promoting the formation of usefully disentangled representations of features in reduced dimensionality latent spaces [100, 101], thereby promoting controllability and combinatorial power in imagination? These are only some of the many ways that Gershman’s intriguing proposal of a “generative adversarial brain” may lead to innovative directions for trying to understand functional relationships within and between minds.

Acknowledgements

I would like to thank the IWAI 2020 organizing committee for arranging this workshop and providing me with an opportunity to share these ideas.

I would like to thank Karl Friston for his mentorship and for being an ongoing source of inspiration over many years.

I would like to thank Amelia Thomley and Victoria Klimaj for their help with illustrations and figures, and for their enduring friendship.

References

1. 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.
2. Chalmers, D.J.: Facing Up to the Problem of Consciousness. Journal of Consciousness Studies. 2, 200–19 (1995).
3. Friston, K.J.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138 (2010). https://doi.org/10.1038/nrn2787.
4. 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.
5. Conant, R., C., Ashby, W.R.: Every good regulator of a system must be a model of that system. International Journal of Systems Science. 1, 89–97 (1970). https://doi.org/10.1080/00207727008920220.
6. Hohwy, J., Seth, A.: Predictive processing as a systematic basis for identifying the neural correlates of consciousness. PsyArXiv (2020). https://doi.org/10.31234/osf.io/nd82g.
7. 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.
8. Safron, A.: Integrated World Modeling Theory (IWMT) Revisited. PsyArXiv (2019). https://doi.org/10.31234/osf.io/kjngh.
9. Clark, A.: Consciousness as Generative Entanglement, https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=jphil&id=jphil_2019_0116_0012_0645_0662, last accessed 2020/01/13. https://doi.org/10.5840/jphil20191161241.
10. Gazzaniga, M.S.: The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind. Farrar, Straus and Giroux (2018).
11. Chang, A.Y.C., Biehl, M., Yu, Y., Kanai, R.: Information Closure Theory of Consciousness. arXiv:1909.13045 [q-bio]. (2019).
12. Ziporyn, B.: Being and Ambiguity: Philosophical Experiments with Tiantai Buddhism. Open Court (2004).
13. Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K.J., Williford, K.: A mathematical model of embodied consciousness. J. Theor. Biol. 428, 106–131 (2017). https://doi.org/10.1016/j.jtbi.2017.05.032.
14. Williford, K., Bennequin, D., Friston, K., Rudrauf, D.: The Projective Consciousness Model and Phenomenal Selfhood. Front. Psychol. 9, (2018). https://doi.org/10.3389/fpsyg.2018.02571.
15. Metzinger, T.: The Ego Tunnel: The Science of the Mind and the Myth of the Self. Basic Books, New York (2009).

16. 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.
17. Corcoran, A.W., Pezzulo, G., Hohwy, J.: From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. (2019). https://doi.org/10.20944/preprints201911.0083.v1.
18. 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.
19. 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.
20. 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).
21. Klein, B., Hoel, E.: The Emergence of Informative Higher Scales in Complex Networks, https://www.hindawi.com/journals/complexity/2020/8932526/, last accessed 2020/04/05. https://doi.org/10.1155/2020/8932526.
22. Hohwy, J.: The Self-Evidencing Brain. Noûs. 50, 259–285 (2016). https://doi.org/10.1111/nous.12062.
23. Mashour, G.A., Roelfsema, P., Changeux, J.-P., Dehaene, S.: Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron. 105, 776–798 (2020). https://doi.org/10.1016/j.neuron.2020.01.026.
24. Whyte, C.J., Smith, R.: The Predictive Global Neuronal Workspace: A Formal Active Inference Model of Visual Consciousness. bioRxiv. 2020.02.11.944611 (2020). https://doi.org/10.1101/2020.02.11.944611.
25. Safron, A.: The radically embodied conscious cybernetic Bayesian brain: Towards explaining the emergence of agency. (2019). https://doi.org/10.31234/osf.io/udc42.
26. Fries, P.: Rhythms For Cognition: Communication Through Coherence. Neuron. 88, 220–235 (2015). https://doi.org/10.1016/j.neuron.2015.09.034.
27. Deco, G., Kringelbach, M.L.: Metastability and Coherence: Extending the Communication through Coherence Hypothesis Using A Whole-Brain Computational Perspective. Trends in Neurosciences. 39, 125–135 (2016). https://doi.org/10.1016/j.tins.2016.01.001.
28. Boly, M., Massimini, M., Tsuchiya, N., Postle, B.R., Koch, C., Tononi, G.: Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. J. Neurosci. 37, 9603–9613 (2017). https://doi.org/10.1523/JNEUROSCI.3218-16.2017.
29. Mumford, D.: On the computational architecture of the neocortex. Biol. Cybern. 65, 135–145 (1991). https://doi.org/10.1007/BF00202389.
30. 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.

31. 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.
32. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat]. (2014).
33. Khemakhem, I., Kingma, D.P., Monti, R.P., Hyvärinen, A.: Variational Autoencoders and Nonlinear ICA: A Unifying Framework. arXiv:1907.04809 [cs, stat]. (2020).
34. Marshel, J.H., Kim, Y.S., Machado, T.A., Quirin, S., Benson, B., Kadmon, J., Raja, C., Chibukhchyan, A., Ramakrishnan, C., Inoue, M., Shane, J.C., McKnight, D.J., Yoshizawa, S., Kato, H.E., Ganguli, S., Deisseroth, K.: Cortical layer–specific critical dynamics triggering perception. Science. 365, eaaw5202 (2019). https://doi.org/10.1126/science.aaw5202.
35. Redinbaugh, M.J., Phillips, J.M., Kambi, N.A., Mohanta, S., Andryk, S., Dooley, G.L., Afrasiabi, M., Raz, A., Saalmann, Y.B.: Thalamus Modulates Consciousness via Layer-Specific Control of Cortex. Neuron. 106, 66-75.e12 (2020). https://doi.org/10.1016/j.neuron.2020.01.005.
36. Rezaei, H., Aertsen, A., Kumar, A., Valizadeh, A.: Facilitating the propagation of spiking activity in feedforward networks by including feedback. PLOS Computational Biology. 16, e1008033 (2020). https://doi.org/10.1371/journal.pcbi.1008033.
37. Hesp, C.: Beyond connectionism: A neuronal dance of ephaptic and synaptic interactions: Commentary on “The growth of cognition: Free energy minimization and the embryogenesis of cortical computation” by Wright and Bourke (2020). Phys Life Rev. (2020). https://doi.org/10.1016/j.plrev.2020.08.002.
38. Northoff, G., Wainio-Theberge, S., Evers, K.: Is temporo-spatial dynamics the “common currency” of brain and mind? In Quest of “Spatiotemporal Neuroscience.” Physics of Life Reviews. 33, 34–54 (2020). https://doi.org/10.1016/j.plrev.2019.05.002.
39. Berrou, C., Glavieux, A., Thitimajshima, P.: Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1. In: Proceedings of ICC ’93 - IEEE International Conference on Communications. pp. 1064–1070 vol.2 (1993). https://doi.org/10.1109/ICC.1993.397441.
40. McEliece, R.J., MacKay, D.J.C., Jung-Fu Cheng: Turbo decoding as an instance of Pearl’s “belief propagation” algorithm. IEEE Journal on Selected Areas in Communications. 16, 140–152 (1998). https://doi.org/10.1109/49.661103.
41. 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).
42. Parr, T., Markovic, D., Kiebel, S.J., Friston, K.J.: Neuronal message passing using Mean-field, Bethe, and Marginal approximations. Scientific Reports. 9, 1889 (2019). https://doi.org/10.1038/s41598-018-38246-3.
43. Northoff, G., Mushiake, H.: Why context matters? Divisive normalization and canonical microcircuits in psychiatric disorders. Neurosci. Res. (2019). https://doi.org/10.1016/j.neures.2019.10.002.

44. 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.
45. 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.
46. Ahmad, S., Scheinkman, L.: How Can We Be So Dense? The Benefits of Using Highly Sparse Representations. arXiv preprint arXiv:1903.11257. (2019).
47. Da Costa, L., Parr, T., Sengupta, B., Friston, K.: Natural selection finds natural gradient. arXiv:2001.08028 [q-bio]. (2020).
48. 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.
49. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. J. Mach. Learn. Res. 11, 3371–3408 (2010).
50. Hawkins, J., Lewis, M., Klukas, M., Purdy, S., Ahmad, S.: A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex. Front. Neural Circuits. 12, (2019). https://doi.org/10.3389/fncir.2018.00121.
51. Kosiorek, A., Sabour, S., Teh, Y.W., Hinton, G.E.: Stacked Capsule Autoencoders. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d\textquotesingle, Fox, E., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 32. pp. 15512–15522. Curran Associates, Inc. (2019).
52. Hu, H.-Y., Li, S.-H., Wang, L., You, Y.-Z.: Machine Learning Holographic Mapping by Neural Network Renormalization Group. arXiv:1903.00804 [cond-mat, physics:hep-th]. (2019).
53. Li, S.-H., Wang, L.: Neural Network Renormalization Group. Phys. Rev. Lett. 121, 260601 (2018). https://doi.org/10.1103/PhysRevLett.121.260601.
54. Castro, S., El-Deredy, W., Battaglia, D., Orio, P.: Cortical ignition dynamics is tightly linked to the core organisation of the human connectome. PLOS Computational Biology. 16, e1007686 (2020). https://doi.org/10.1371/journal.pcbi.1007686.
55. Kingma, D.P., Salimans, T., Welling, M.: Variational Dropout and the Local Reparameterization Trick. arXiv:1506.02557 [cs, stat]. (2015).
56. Buzsáki, G., Tingley, D.: Space and Time: The Hippocampus as a Sequence Generator. Trends in Cognitive Sciences. 22, 853–869 (2018). https://doi.org/10.1016/j.tics.2018.07.006.
57. Hawkins, J., Blakeslee, S.: On Intelligence. Times Books (2004).
58. Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A.: Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron. 95, 709-721.e5 (2017). https://doi.org/10.1016/j.neuron.2017.06.041.
59. Whittington, J.C., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., Behrens, T.E.: The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalisation in the hippocampal formation. bioRxiv. 770495 (2019). https://doi.org/10.1101/770495.

60. Moser, E.I., Kropff, E., Moser, M.-B.: Place cells, grid cells, and the brain's spatial representation system. Annu. Rev. Neurosci. 31, 69–89 (2008). https://doi.org/10.1146/annurev.neuro.31.061307.090723.
61. Zhang, F., Li, S., Yuan, S., Sun, E., Zhao, L.: Algorithms analysis of mobile robot SLAM based on Kalman and particle filter. In: 2017 9th International Conference on Modelling, Identification and Control (ICMIC). pp. 1050–1055 (2017). https://doi.org/10.1109/ICMIC.2017.8321612.
62. 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.
63. McNamara, C.G., Dupret, D.: Two sources of dopamine for the hippocampus. Trends Neurosci. 40, 383–384 (2017). https://doi.org/10.1016/j.tins.2017.05.005.
64. 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.
65. Hasz, B.M., Redish, A.D.: Spatial encoding in dorsomedial prefrontal cortex and hippocampus is related during deliberation. Hippocampus. n/a,. https://doi.org/10.1002/hipo.23250.
66. Fraccaro, M., Kamronn, S., Paquet, U., Winther, O.: A disentangled recognition and nonlinear dynamics model for unsupervised learning. In: Advances in Neural Information Processing Systems. pp. 3601–3610 (2017).
67. 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.
68. 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.
69. Canolty, R.T., Knight, R.T.: The functional role of cross-frequency coupling. Trends Cogn. Sci. (Regul. Ed.). 14, 506–515 (2010). https://doi.org/10.1016/j.tics.2010.09.001.
70. Sarel, A., Finkelstein, A., Las, L., Ulanovsky, N.: Vectorial representation of spatial goals in the hippocampus of bats. Science. 355, 176–180 (2017). https://doi.org/10.1126/science.aak9589.
71. Hills, T.T.: Neurocognitive free will. Proceedings. Biological sciences. 286, 20190510 (2019). https://doi.org/10.1098/rspb.2019.0510.
72. MacKay, D.G.: Remembering: What 50 Years of Research with Famous Amnesia Patient H. M. Can Teach Us about Memory and How It Works. Prometheus Books (2019).
73. Kunz, L., Wang, L., Lachner-Piza, D., Zhang, H., Brandt, A., Dümpelmann, M., Reinacher, P.C., Coenen, V.A., Chen, D., Wang, W.-X., Zhou, W., Liang, S., Grewe, P., Bien, C.G., Bierbrauer, A., Schröder, T.N., Schulze-Bonhage, A., Axmacher, N.: Hippocampal theta phases organize the reactivation of large-scale

electrophysiological representations during goal-directed navigation. Science Advances. 5, eaav8192 (2019). https://doi.org/10.1126/sciadv.aav8192.

74. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Company, New York, NY (1983).
75. Prinz, J.: The Intermediate Level Theory of Consciousness. In: The Blackwell Companion to Consciousness. pp. 257–271. John Wiley & Sons, Ltd (2017). https://doi.org/10.1002/9781119132363.ch18.
76. Varela, F.J., Thompson, E.T., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. The MIT Press, Cambridge, Mass. (1992).
77. Barsalou, L.W.: Grounded cognition: past, present, and future. Top Cogn Sci. 2, 716–724 (2010). https://doi.org/10.1111/j.1756-8765.2010.01115.x.
78. Graziano, M.S.A.: Rethinking consciousness: a scientific theory of subjective experience. WWNorton & Company, New York (2019).
79. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph Neural Networks: A Review of Methods and Applications. arXiv:1812.08434 [cs, stat]. (2019).
80. 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).
81. Bapst, V., Keck, T., Grabska-Barwińska, A., Donner, C., Cubuk, E.D., Schoenholz, S.S., Obika, A., Nelson, A.W.R., Back, T., Hassabis, D., Kohli, P.: Unveiling the predictive power of static structure in glassy systems. Nature Physics. 16, 448–454 (2020). https://doi.org/10.1038/s41567-020-0842-8.
82. Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., Ho, S.: Discovering Symbolic Models from Deep Learning with Inductive Biases. arXiv:2006.11287 [astro-ph, physics:physics, stat]. (2020).
83. Xi, E., Bing, S., Jin, Y.: Capsule network performance on complex data. arXiv preprint arXiv:1712.03480. (2017).
84. 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.
85. Haun, A.: What is visible across the visual field? (2020). https://doi.org/10.31234/osf.io/wdpu7.
86. Kaplan, R., Friston, K.J.: Planning and navigation as active inference. Biol Cybern. 112, 323–343 (2018). https://doi.org/10.1007/s00422-018-0753-2.
87. Hills, T.T., Todd, P.M., Goldstone, R.L.: The Central Executive as a Search Process: Priming Exploration and Exploitation across Domains. J Exp Psychol Gen. 139, 590–609 (2010). https://doi.org/10.1037/a0020666.
88. Rochat, P.: Emerging Self-Concept. In: Bremner, J.G. and Wachs, T.D. (eds.) The Wiley-Blackwell Handbook of Infant Development. pp. 320–344. Wiley-Blackwell (2010). https://doi.org/10.1002/9781444327564.ch10.

89. Graziano, M.S.A.: The temporoparietal junction and awareness. Neurosci Conscious. 2018, (2018). https://doi.org/10.1093/nc/niy005.
90. Parr, T., Rikhye, R.V., Halassa, M.M., Friston, K.J.: Prefrontal computation as active inference. Cerebral Cortex. (2019).
91. Stephenson-Jones, M., Samuelsson, E., Ericsson, J., Robertson, B., Grillner, S.: Evolutionary conservation of the basal ganglia as a common vertebrate mechanism for action selection. Curr. Biol. 21, 1081–1091 (2011). https://doi.org/10.1016/j.cub.2011.05.001.
92. Houk, J.C., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P.J., Roy, S.A., Simo, L.S.: Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 362, 1573–1583 (2007). https://doi.org/10.1098/rstb.2007.2063.
93. Humphries, M.D., Prescott, T.J.: The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. Prog. Neurobiol. 90, 385–417 (2010). https://doi.org/10.1016/j.pneurobio.2009.11.003.
94. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., Botvinick, M.: A distributional code for value in dopamine-based reinforcement learning. Nature. 1–5 (2020). https://doi.org/10.1038/s41586-019-1924-6.
95. Morrens, J., Aydin, Ç., Rensburg, A.J. van, Rabell, J.E., Haesler, S.: Cue-Evoked Dopamine Promotes Conditioned Responding during Learning. Neuron. 0, (2020). https://doi.org/10.1016/j.neuron.2020.01.012.
96. Tani, J.: Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press (2016).
97. 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.
98. Honkanen, A., Adden, A., Freitas, J. da S., Heinze, S.: The insect central complex and the neural basis of navigational strategies. Journal of Experimental Biology. 222, (2019). https://doi.org/10.1242/jeb.188854.
99. Bengio, Y.: The Consciousness Prior. arXiv:1709.08568 [cs, stat]. (2017).
100. Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., Pineau, J., Precup, D., Bengio, Y.: Independently controllable factors. arXiv preprint arXiv:1708.01289. (2017).
101. Thomas, V., Bengio, E., Fedus, W., Pondard, J., Beaudoin, P., Larochelle, H., Pineau, J., Precup, D., Bengio, Y.: Disentangling the independently controllable factors of variation by interacting with the world. arXiv preprint arXiv:1802.09484. (2018).
102. Ding, Z., Shao, M., Fu, Y.: Robust Multi-view Representation: A Unified Perspective from Multi-view Learning to Domain Adaption. 5434–5440 (2018).
103. Friston, K.J., Breakspear, M., Deco, G.: Perception and self-organized instability. Front. Comput. Neurosci. 6, (2012). https://doi.org/10.3389/fncom.2012.00044.

104. 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.
105. Humphrey, N.: The Invention of Consciousness. Topoi. 39, 13–21 (2017). https://doi.org/10.1007/s11245-017-9498-0.
106. Barron, H.C., Auksztulewicz, R., Friston, K.: Prediction and memory: a predictive coding account. Progress in Neurobiology. 101821 (2020). https://doi.org/10.1016/j.pneurobio.2020.101821.
107. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books (2018).
108. Kay, K., Chung, J.E., Sosa, M., Schor, J.S., Karlsson, M.P., Larkin, M.C., Liu, D.F., Frank, L.M.: Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus. Cell. 180, 552-567.e25 (2020). https://doi.org/10.1016/j.cell.2020.01.014.
109. Dohmatob, E., Dumas, G., Bzdok, D.: Dark control: The default mode network as a reinforcement learning agent. Human Brain Mapping. 41, 3318–3341 (2020). https://doi.org/10.1002/hbm.25019.
110. Hill, F., Santoro, A., Barrett, D.G.T., Morcos, A.S., Lillicrap, T.: Learning to Make Analogies by Contrasting Abstract Relational Structure. arXiv:1902.00120 [cs]. (2019).
111. Crouse, M., Nakos, C., Abdelaziz, I., Forbus, K.: Neural Analogical Matching. arXiv:2004.03573 [cs]. (2020).
112. Safron, A.: Bayesian Analogical Cybernetics. arXiv:1911.02362 [q-bio]. (2019).
113. Baram, A.B., Muller, T.H., Nili, H., Garvert, M., Behrens, T.E.J.: Entorhinal and ventromedial prefrontal cortices abstract and generalise the structure of reinforcement learning problems. bioRxiv. 827253 (2019). https://doi.org/10.1101/827253.
114. Damasio, A.: Self Comes to Mind: Constructing the Conscious Brain. Vintage, New York (2012).
115. Livneh, Y., Sugden, A.U., Madara, J.C., Essner, R.A., Flores, V.I., Sugden, L.A., Resch, J.M., Lowell, B.B., Andermann, M.L.: Estimation of Current and Future Physiological States in Insular Cortex. Neuron. 0, (2020). https://doi.org/10.1016/j.neuron.2019.12.027.
116. Parascandolo, G., Buesing, L., Merel, J., Hasenclever, L., Aslanides, J., Hamrick, J.B., Heess, N., Neitz, A., Weber, T.: Divide-and-Conquer Monte Carlo Tree Search For Goal-Directed Planning. arXiv:2004.11410 [cs, stat]. (2020).
117. Rueter, A.R., Abram, S.V., MacDonald, A.W., Rustichini, A., DeYoung, C.G.: The goal priority network as a neural substrate of Conscientiousness. Human Brain Mapping. 39, 3574–3585 (2018). https://doi.org/10.1002/hbm.24195.
118. Verleger, R., Haake, M., Baur, A., Śmigasiewicz, K.: Time to Move Again: Does the Bereitschaftspotential Covary with Demands on Internal Timing? Front. Hum. Neurosci. 10, (2016). https://doi.org/10.3389/fnhum.2016.00642.
119. Park, H.-D., Barnoud, C., Trang, H., Kannape, O.A., Schaller, K., Blanke, O.: Breathing is coupled with voluntary action and the cortical readiness potential. Nature Communications. 11, 1–8 (2020). https://doi.org/10.1038/s41467-019-13967-9.

120. Travers, E., Friedemann, M., Haggard, P.: The Readiness Potential reflects expectation, not uncertainty, in the timing of action. bioRxiv. 2020.04.16.045344 (2020). https://doi.org/10.1101/2020.04.16.045344.
121. Adams, R., Shipp, S., Friston, K.J.: Predictions not commands: active inference in the motor system. Brain Struct Funct. 218, 611–643 (2013). https://doi.org/10.1007/s00429-012-0475-5.
122. Gershman, S.J.: The Generative Adversarial Brain. Front. Artif. Intell. 2, (2019). https://doi.org/10.3389/frai.2019.00018.
123. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv:1512.09300 [cs, stat]. (2016).
124. Ha, D., Schmidhuber, J.: World Models. arXiv:1803.10122 [cs, stat]. (2018). https://doi.org/10.5281/zenodo.1207631.
125. Magno, E., Foxe, J.J., Molholm, S., Robertson, I.H., Garavan, H.: The anterior cingulate and error avoidance. J. Neurosci. 26, 4769–4773 (2006). https://doi.org/10.1523/JNEUROSCI.0369-06.2006.
126. Garrison, J.R., Fernyhough, C., McCarthy-Jones, S., Haggard, M., Simons, J.S.: Paracingulate sulcus morphology is associated with hallucinations in the human brain. Nature Communications. 6, 8956 (2015). https://doi.org/10.1038/ncomms9956.
127. Stolyarova, A., Rakhshan, M., Hart, E.E., O'Dell, T.J., Peters, M. a. K., Lau, H., Soltani, A., Izquierdo, A.: Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. Nature Communications. 10, 1–14 (2019). https://doi.org/10.1038/s41467-019-12725-1.
128. Boroujeni, K.B., Tiesinga, P., Womelsdorf, T.: Interneuron Specific Gamma Synchronization Encodes Uncertain Cues and Prediction Errors in Lateral Prefrontal and Anterior Cingulate Cortex. bioRxiv. 2020.07.24.220319 (2020). https://doi.org/10.1101/2020.07.24.220319.
129. Lenhart, L., Steiger, R., Waibel, M., Mangesius, S., Grams, A.E., Singewald, N., Gizewski, E.R.: Cortical reorganization processes in meditation naïve participants induced by 7 weeks focused attention meditation training. Behavioural Brain Research. 395, 112828 (2020). https://doi.org/10.1016/j.bbr.2020.112828.
130. Vassena, E., Deraeve, J., Alexander, W.H.: Surprise, value and control in anterior cingulate cortex during speeded decision-making. Nat Hum Behav. 1–11 (2020). https://doi.org/10.1038/s41562-019-0801-5.
131. Bubb, E.J., Metzler-Baddeley, C., Aggleton, J.P.: The cingulum bundle: Anatomy, function, and dysfunction. Neuroscience & Biobehavioral Reviews. 92, 104–127 (2018). https://doi.org/10.1016/j.neubiorev.2018.05.008.
132. Robinson, R.: Stimulating the Cingulum Relieves Anxiety During Awake Neurosurgery: What Is the Therapeutic Potential? Neurology Today. 19, 27 (2019). https://doi.org/10.1097/01.NT.0000554700.13747.f2.
133. Craig, A.D.B.: Significance of the insula for the evolution of human awareness of feelings from the body. Ann. N. Y. Acad. Sci. 1225, 72–82 (2011). https://doi.org/10.1111/j.1749-6632.2011.05990.x.

134. Schmidhuber, J.: POWERPLAY: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem. arXiv:1112.5309 [cs]. (2012).
135. Izquierdo-Torres, E., Bührmann, T.: Analysis of a Dynamical Recurrent Neural Network Evolved for Two Qualitatively Different Tasks: Walking and Chemotaxis. In: ALIFE (2008).
136. Izquierdo, E., Aguilera, M., Beer, R.: Analysis of Ultrastability in Small Dynamical Recurrent Neural Networks. The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE). 25, 51–58 (2013). https://doi.org/10.1162/978-0-262-31709-2-ch008.
137. Pezzulo, G., Rigoli, F., Friston, K.J.: Hierarchical Active Inference: A Theory of Motivated Control. Trends in Cognitive Sciences. 22, 294–306 (2018). https://doi.org/10.1016/j.tics.2018.01.009.
138. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., Berker, A. de, Ganguli, S., Gillon, C.J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G.W., Miller, K.D., Naud, R., Pack, C.C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A.C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., Kording, K.P.: A deep learning framework for neuroscience. Nat Neurosci. 22, 1761–1770 (2019). https://doi.org/10.1038/s41593-019-0520-2.
139. 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.
140. 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).
141. Arese Lucini, F., Del Ferraro, G., Sigman, M., Makse, H.A.: How the Brain Transitions from Conscious to Subliminal Perception. Neuroscience. 411, 280–290 (2019). https://doi.org/10.1016/j.neuroscience.2019.03.047.