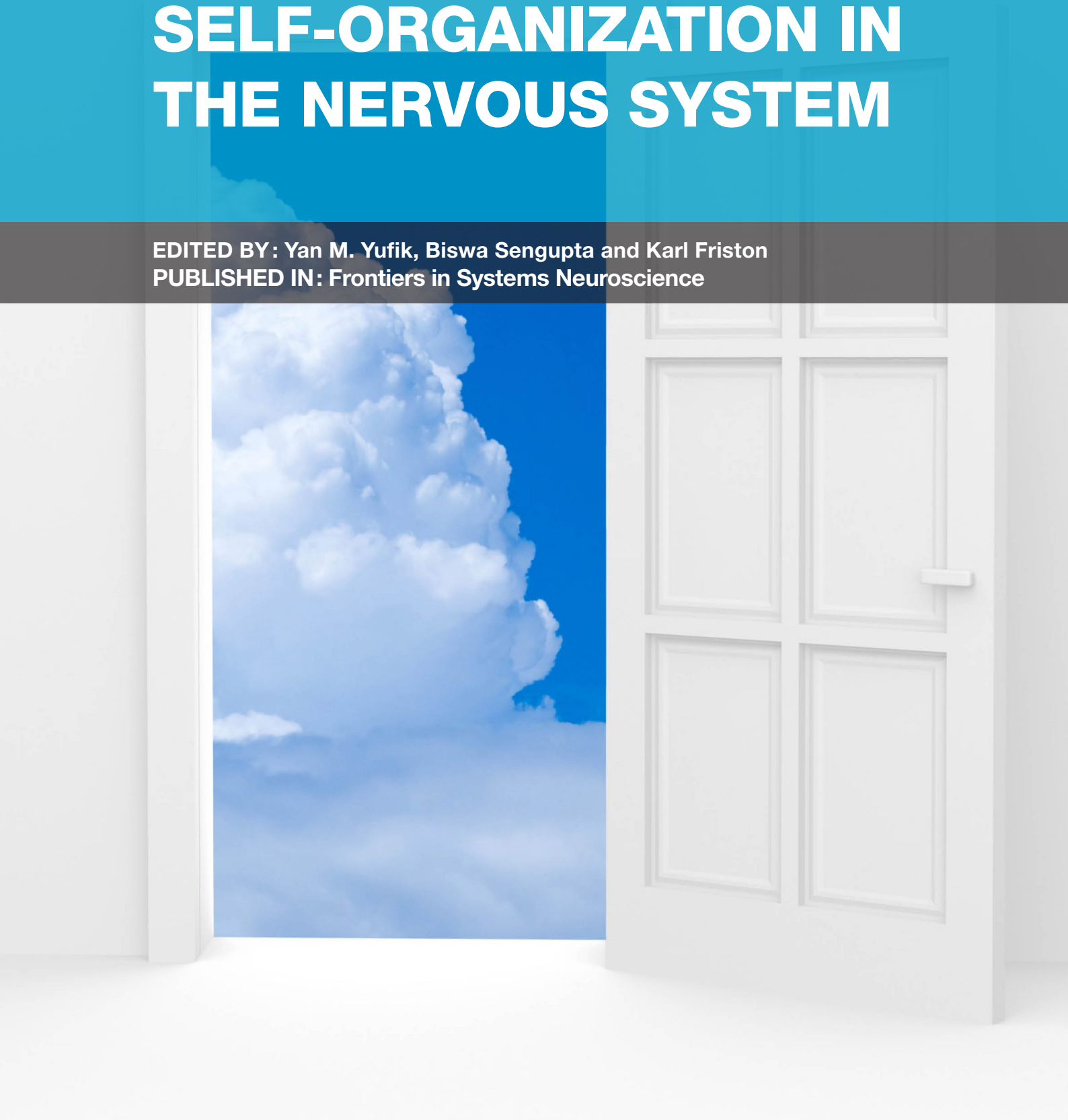
SELF-ORGANIZATION IN THE NERVOUS SYSTEM

EDITED BY: Yan M. Yufik, Biswa Sengupta and Karl Friston
PUBLISHED IN: Frontiers in Systems Neuroscience



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-340-5
DOI 10.3389/978-2-88945-340-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

SELF-ORGANIZATION IN THE NERVOUS SYSTEM

Topic Editors:
Yan M. Yufik, Virtual Structures Research, Inc., United States
Biswa Sengupta, University of Cambridge, United Kingdom
Karl Friston, Wellcome Trust Centre for Neuroimaging at UCL, United Kingdom

Cover image: Fedorov Oleksiy/Shutterstock.com

This special issue reviews state-of-the-art approaches to the biophysical roots of cognition. These approaches appeal to the notion that cognitive capacities serve to optimize responses to changing external conditions. Crucially, this optimisation rests on the ability to predict changes in the environment, thus allowing organisms to respond pre-emptively to changes before their onset. The biophysical mechanisms that underwrite these cognitive capacities remain largely

unknown; although a number of hypotheses has been advanced in systems neuroscience, biophysics and other disciplines. These hypotheses converge on the intersection of thermodynamic and information-theoretic formulations of self-organization in the brain. The latter perspective emerged when Shannon's theory of message transmission in communication systems was used to characterise message passing between neurons. In its subsequent incarnations, the information theory approach has been integrated into computational neuroscience and the Bayesian brain framework. The thermodynamic formulation rests on a view of the brain as an aggregation of stochastic microprocessors (neurons), with subsequent appeal to the constructs of statistical mechanics and thermodynamics. In particular, the use of ensemble dynamics to elucidate the relationship between micro-scale parameters and those of the macro-scale aggregation (the brain). In general, the thermodynamic approach treats the brain as a dissipative system and seeks to represent the development and functioning of cognitive mechanisms as collective capacities that emerge in the course of self-organization. Its explicanda include energy efficiency; enabling progressively more complex cognitive operations such as long-term prediction and anticipatory planning. A cardinal example of the Bayesian brain approach is the free energy principle that explains self-organizing dynamics in the brain in terms of its predictive capabilities – and selective sampling of sensory inputs that optimise variational free energy as a proxy for Bayesian model evidence. An example of thermodynamically grounded proposals, in this issue, associates self-organization with phase transitions in neuronal state-spaces; resulting in the formation of bounded neuronal assemblies (neuronal packets). This special issue seeks a discourse between thermodynamic and informational formulations of the self-organising and self-evidencing brain. For example, could minimization of thermodynamic free energy during the formation of neuronal packets underlie minimization of variational free energy?

Citation: Yufik, Y. M., Sengupta, B., Friston, K., eds. (2017). Self-Organization in the Nervous System. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-340-5

Table of Contents

frontiers in Systems Neuroscience

EDITORIAL
published: 26 September 2017
doi: 10.3389/fnsys.2017.00069

Editorial: Self-Organization in the Nervous System

Yan M. Yufik[1], *Biswa Sengupta*[2]* *and Karl Friston*[3]

[1] *Virtual Structures Research Inc., Potomac, MD, United States,* [2] *Department of Bioengineering, Imperial College London, London, United Kingdom,* [3] *Institute of Neurology, Wellcome Trust Centre for Neuroimaging, London, United Kingdom*

Keywords: self-organization, neural circuits, variational inference, bayesian inference, dynamical systems theory

Editorial on the Research Topic

Self-Organization in the Nervous System

"Self-organization is the spontaneous—often seemingly purposeful—formation of spatial, temporal, spatiotemporal structures, or functions in systems composed of few or many components. In physics, chemistry and biology self-organization occurs in open systems driven away from thermal equilibrium" (Haken, Scholarpedia). The contributions in this special issue aim to elucidate the role of self-organization in shaping the cognitive processes in the course of development and throughout evolution, or "from paramecia to Einstein" (Torday and Miller). The central question is: what self-organizing mechanisms in the human nervous system are common to all forms of life, and what mechanisms (if any) are unique to the human species?

Over the last several decades, the problem of self-organization has been at the forefront of research in biological and machine intelligence (Kohonen, 1989; Kauffman, 1993; Pribram, 1994, 1996, 1998; Kelso, 1997; Camazine et al., 2003; Zanette et al., 2004; Haken, 2010, 2012, and others). The articles collected in this issue present recent findings (and ideas) from diverse perspectives and address different facets of the problem. Two features of this collection might be of particular interest to the reader: (i) the scope of discussion is broad, stretching from general thermodynamic and information-theoretic principles to the expression of these principles in human cognition, consciousness and understanding and (ii) many of the ideas speak to a unifying perspective outlined below. In what follows, we will preview the collection of papers in this special issue and frame them in terms of a unified approach to self organization—leaving the reader to judge the degree to which subsequent articles are consistent with or contradict this framework.

Living organisms must regulate flows of energy and matter through their boundary surfaces to underwrite their survival. Cognitive development is the product of progressive fine-tuning (optimization) of regulatory mechanisms, under the dual criteria of minimizing surprise (Friston, 2010; Sengupta et al., 2013, 2016; Sengupta and Friston, 2017) and maximizing thermodynamic efficiency (Yufik, 2002, 2013). The former implies reducing the likelihood of encountering conditions impervious to regulation (e.g., inability to block inflows of destructive substances); the latter implies maintaining net energy intakes above some survival thresholds. Energy is expended in regulatory processes formed in the course of self-organization and predicated on lowering thermodynamic entropy "on the inside" and transporting excessive entropy (heat) "to the outside." Efficient regulation requires mechanisms that necessarily incorporate models of the system and its relation to environment (Conant and Ashby, 1970). Primitive animals possess small repertoires of genetically fixed, rigid models, while—in more advanced animals—the repertoires are larger and their models become more flexible; i.e., amenable to experience-driven modifications. Both the evolutionary and experience-driven modifications are forms of statistical learning: models are sculpted by external feedback conveying statistical properties of the environment. Human learning mechanisms, although built on the foundation of statistical learning, depart

OPEN ACCESS

Edited and reviewed by:
Maria V. Sanchez-Vives, Consorci Institut D'Investigacions Biomediques August Pi I Sunyer, Spain

***Correspondence:**
Biswa Sengupta
b.sengupta@imperial.ac.uk

Received: *21 June 2017*
Accepted: *11 September 2017*
Published: *26 September 2017*

Citation:
Yufik YM, Sengupta B and Friston K (2017) Editorial: Self-Organization in the Nervous System. Front. Syst. Neurosci. 11:69. doi: 10.3389/fnsys.2017.00069

radically from conventional (e.g., machine) learning: the implicit models become amenable to self-directed composition and modification based on interoceptive, as opposed (or in addition) to exteroceptive, feedback (Yufik, 1998). Interoceptive feedback underlies the feeling of grasp, or understanding that accompanies the organization of disparate "representations" into cohesive structures amenable to further operations (mental modeling). The work of mental modeling requires energy; consciousness is co-extensive with deliberate (attentive, focused) application of energy ("cognitive effort") in carrying out that work. Learning with understanding departs from statistical (machine) learning in three ways: (i) mental models anticipate experiences, as opposed to be shaped by them (e.g., the theory of relativity originated in gedanken experiments); (ii) feedback conveys properties of implicit models (coherence, simplicity, validation opportunities the models afford, etc.) and (iii) manipulating (executing or inverting) models enables efficient exchange with the environment, under conditions with no precedents (and thus no learnable statistical representation) (Yufik, 2013). Regulation of this sort—based on statistical learning—faces a challenging complexity. As the number of regulated variables grows; energy demands can quickly become unsustainable. Using self-organization to implement the process of "understanding" (i.e., composing more general models) has the triple benefit of minimizing surprise, while averting complexity and advancing thermodynamic efficiency of regulatory processes into the vicinity of theoretical limits.

Annila argues that the most fundamental function performed by the nervous system is shared by all open systems and entails a generation of entropy, by extracting high-grade free energy from the environment and returning low-grade energy. As dictated by the second law of thermodynamics, cognitive processes seek out opportunities (paths) for consuming free energy in the least time. Evolution obtains progressively more efficient mechanisms for detecting and exploiting free energy deposits, culminating in consciousness that emerges in systems pertaining to the ability to "integrate various neural networks for coherent consumption of free energy..." (Annila, this issue).

Street reviews discussions in the literature that examine the tension between—and synthesis of—information-theoretic and thermodynamics-motivated conceptualizations of brain processes. Tensions are rooted in the theory of information, designed to allow analysis of information transfer, irrespective of the physical processes that mediate transfer. Synthesis is necessitated by considerations of energy costs incurred in neuronal signaling. A consensus is anticipated, within a theoretical framework that views cognitive development as self-organization in the nervous system—seeking to minimize surprise, while incurring minimum energy costs.

Torday and Miller discuss the conceptual framework needed for tracing evolution of the mammalian brain "from paramecia to Einstein." The framework encompasses three key notions: (i) complex multicellular organisms share fundamental organizational properties, with precursors in unicellular forms of life, (ii) the most basic property is the ability to extract energy from the environment and dissipate heat in a manner enabling homeostasis and processing of information and (iii) evolutionary improvements in homeostasis, self-maintenance and information processing derive from increased cellular collaboration (coherence). Within this framework, "life is cognition at every scope and scale" and "any cognitive action as a form of cellular coherence can be better understood as both an information exchange and reciprocally then, as energy conversion and transfer" (Torday and Miller).

Campbell argues that Darwinian evolution can be expressed as a process of Bayesian updating. Conventionally, the ability to draw inferences and update Bayesian models has been attributed exclusively to (human) reasoning. The range of attribution can be expanded to include all organisms, by assuming that genotypes carry latent models of the environment receiving varying expressions in the phenotype. On that view, genetically transmitted models are the source of hypotheses (phenotype variations) subjected to confirmation (survival) or rejection (extinction) by the environment. Changes in the phenotype over somatic time and the genotype over evolutionary time minimize surprise thus increasing the likelihood of survival of individuals and the species.

Kozma and Freeman analyze alternations between highly organized (low entropy) and disorganized (high entropy) neuronal activities induced by visual stimuli. In rabbits implanted with ECoG arrays of electrodes fixed over the visual cortex, presentations of stimuli were accompanied by metastable patterns of synchronized activity—collapsing quickly into the background activity upon cessation of the stimuli. The authors define alternations between metastable patterns and disorganized firings as phase transitions and propose a "cinematic" theory of perception; treating alternations that spread across the cortex as successions of "frames" combined into perceptual units (percepts). Synchronized neuronal populations are identified with Hebbian assemblies, acting in a self-catalytic fashion: Interactions between assemblies maintain the cortex in the critical state, conducive to the emergence of organized (low entropy) structures, such as Hebbian assemblies.

Stankovski et al. present novel findings concerning the coherence of neuronal assemblies. Assemblies oscillate within characteristic frequency intervals, with cross-frequency coupling serving to integrate assemblies into functional networks that span distant regions in the brain. In this study, cross-frequency coupling functions were reconstructed from EEG recordings from human subjects in the state of rest, with the eyes either open or closed. They review early evidence that closing the eyes triggers an increase in coupling strength. A novel method of analysis then allows them to determine variations in coupling strength across frequency ranges: crucially, they find that increases in the strength of inter-assembly coupling are accompanied by narrowing variation envelopes.

Tang et al. recorded experience-induced changes in the connectivity of large-scale brain networks. Subjects were resting in a state of "mindfulness," under minimal exposure to external stimuli. A comprehensive array of mathematical analyses was applied to the fMRI data. The analyses reveal statistically significant increases in connectivity between different brain areas. Many earlier studies have demonstrated increased connectivity in brain networks under external stimuli; however,

according to this study, similar increases can be produced in the course of internally-induced, restful states.

Werbos and Davis review progress to date in modeling cognitive functions, focusing on the neural net model of learning employing back-propagation algorithms. Neural nets represent learning as the acquisition of desired mappings between input vectors (environmental conditions) and output vectors (desired responses), via iterative reduction of mapping errors. The model posits successions of calculations propagating forward and backward in the neuronal system, orchestrated by some global clock. Empirical substantiations of this model have been scarce—but new experimental findings and analysis are presented that speak to its biological plausibility.

Perlovsky's "physics in the mind" research program tries to define the principles of cognition in a rigorous way (a la Newtonian mechanics). Some principles are suggested including *mental modeling*, *vague representations*, *knowledge instinct, dynamic logic* and *dual hierarchy*. A *mental model* is the basic functional unit of cognition, *models* are *vague* (lacking detail), while sensory inputs are *crisp* (rich in detail). Acquiring knowledge involves reconciling models and inputs in a process driven by *knowledge instinct* and employing mechanisms of *dynamic logic*. Model hierarchy has a counterpart in linguistic hierarchy (hence, the *dual hierarchy*).

Newton analyzes composition of understanding and identifies three constituents: (i) imagery, (ii) the state of mental tension (surprise) caused by a novel situation and (iii) the state of tension resolution, provided by having worked out responses afforded by the situation. The feeling of having reached understanding (Aha!) precedes response execution and thus depends on factors other than external feedback (although failures can restore tension). Execution involves some forms of bodily activities; so "understanding" is anchored in the mechanisms that control such activities. Understanding can then expand via mapping new situations onto those that are already understood.

Yufik and Friston suggest that the same self-organization principle manifests in both the emergence of life and evolution of regulatory mechanisms sustaining life: Regions (subnets) in networks of interacting units (molecules, neurons) fold into bounded structures stabilized by boundary processes. Evolution expanded regulation mechanisms from conditioning to anticipatory planning—that is accomplished via self-directed composition and execution of *mental models*. Hebbian assemblies stabilized by boundary energy barriers (*neuronal packets*) are produced by folding and phase transition in neuronal networks and represent (model) persistent constellations of stimuli (*objects*). Variations in packet responses (changes in the composition of responding groups and the order of their firing inside the packet) represent *behavior*. "Understanding" accompanies the composition of *models* representing behavior coordination (*inter-object relations)*, as bi-directional (reversible) mapping between packets. Such reversible mapping underlies behavior prediction and explanation (retrodiction). Coordination establishes thermodynamic equilibrium in the volume of a model thus minimizing dissipation (costs) and enabling reversible execution. Expanding models and exploring new inputs necessary moves the system away from equilibrium. Regulation via anticipation and explanation is a uniquely human form of surprise minimization. The regulatory process is supported by verbalization and imagery but is driven by *modeling*. Arguably, mental modeling, i.e., coordination of packets (mental objects) in the mental space builds on the neuronal machinery engaged in coordinating limbs and objects in the physical space.

This concludes our brief survey of the articles offered in the special issue. To an outside observer, cars might appear to have the purpose of seeking out gas stations and converting fuel into heat and exhaust. A closer inspection will reveal intelligent regulators inside the cars (i.e., you and me) concerned with having enough fuel to reach the next station—and averting the "surprise" of finding the fuel tank empty. Other concerns—that contextualize this regulation—are the cost of fuel and the desire to keep the car running for the greatest distance possible. In the process, cars must maneuver in coordination with other cars, traffic rules and terrain. Such self-motivated, self-evidencing and self-regulated cars might be a plausible metaphor for minds embedded in a self-organizing nervous system.

AUTHOR CONTRIBUTIONS

YY, BS, and KF contributed equally to this editorial.

REFERENCES

Camazine, S., Deneubourg, J.-L., Franks, N. R., Sneyd, J., Theraula, J., and Bonabeau, E. (2003). *Self-Organization in Biological Systems*. Princeton, NJ: Princeton University Press.

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Systems Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Friston, K. (2010). The free energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Haken, H. (2010). *Information and Self-Organization: A Macroscopic Approach to Complex Systems*. Berlin; Heidelberg: Springer-Verlag.

Haken, H. (2012). *Principles of brain functioning: A synergetic approach to brain activity, Behavior and Cognition*. Berlin; Heidelberg: Springer-Verlag.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection In Evolution*. New York, NY: Oxford University Press.

Kelso, J. A. S. (1997). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: The MIT Press.

Kohonen, T. (1989). *Self-Organization and Associative Memory*. Berlin; Heidelberg: Springer-Verlag.

Pribram, K. H. (ed.). (1994). *Origins: Brain and Self-Organization*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Pribram, K. H. (ed.). (1996). *Learning As Self-Organization*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Pribram, K. H. (ed.). (1998). *Brain and Values*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Sengupta, B., and Friston, K. J. (2017). Sentient self-organization: minimal dynamics and circular causality *arXiv* 1705.08265.

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and efficiency in the nervous system—A synthesis. *PLoS Comput. Biol.* 9:e1003157. doi: 10.1371/journal.pcbi.1003157

Sengupta, B., Tozzi, A., Cooray, G. K., Douglas, P. K., and Friston, K. J. (2016). Towards a neuronal gauge theory. *PLoS Biol.* 14:e1002400. doi: 10.1371/journal.pbio.1002400

Yufik, Y. M. (1998). "Virtual associative networks: a framework for cognitive modeling," in *Brain and Values*, ed K. H. Pribram (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 109–177.

Yufik, Y. M. (2002). "How the mind works," in *Proceedings IEEE World Congress on Computational Intelligence* (Honolulu, HI).

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Zanette, D. H., Manrubia, S. C., and Mikhailov, A. S. (2004). *Emergence of Dynamical Order: Synchronization Phenomena in Complex Systems*. Singapore: World Scientific Publishing.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 30 March 2016
doi: 10.3389/fnsys.2016.00027

On the Character of Consciousness

Arto Annila [1, 2]*

[1] Department of Physics, University of Helsinki, Helsinki, Finland, [2] Department of Biosciences, University of Helsinki, Helsinki, Finland

The human brain is a particularly demanding system to infer its nature from observations. Thus, there is on one hand plenty of room for theorizing and on the other hand a pressing need for a rigorous theory. We apply statistical mechanics of open systems to describe the brain as a hierarchical system in consuming free energy in least time. This holistic tenet accounts for cellular metabolism, neuronal signaling, cognitive processes all together, or any other process by a formal equation of motion that extends down to the ultimate precision of one quantum of action. According to this general thermodynamic theory cognitive processes are no different by their operational and organizational principle from other natural processes. Cognition too will emerge and evolve along path-dependent and non-determinate trajectories by consuming free energy in least time to attain thermodynamic balance within the nervous system itself and with its surrounding systems. Specifically, consciousness can be ascribed to a natural process that integrates various neural networks for coherent consumption of free energy, i.e., for meaningful deeds. The whole hierarchy of integrated systems can be formally summed up to thermodynamic entropy. The holistic tenet provides insight to the character of consciousness also by acknowledging awareness in other systems at other levels of nature's hierarchy.

Keywords: causality, cognition, free energy, non-determinism, the principle of least action, the second law of thermodynamics

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virual Structures Research, Inc., USA

Reviewed by:
Andrew A. Fingelkurts,
BM-Science - Brain & Mind Technologies Research Centre, Finland
Todd L. Hylton,
Brain Corporation, USA

***Correspondence:**
Arto Annila
arto.annila@helsinki.fi

Received: *04 January 2016*
Accepted: *11 March 2016*
Published: *30 March 2016*

Citation:
Annila A (2016) On the Character of Consciousness. Front. Syst. Neurosci. 10:27. doi: 10.3389/fnsys.2016.00027

INTRODUCTION

Cognition is an ability that one has inherited from the evolutionary course of human species and its ancestors as well as accumulated in the course of one's own life from numerous experiences and incidences during diverse developmental and maturation processes. To perceive cognition in this way as a product of various processes raises a profound question: What is a change? Namely, an event, development or evolution as a whole ultimately consists of changes from one state to another. The decimation of any process to a series is familiar from physics but the conceptualization is not remote to neuroscience either (Fingelkurts and Fingelkurts, 2001; John, 2002; Perlovsky and Kozma, 2007; Freeman and Vitiello, 2009; Fingelkurts et al., 2010a, 2013). Moreover, cognition is not only one's arsenal from the past, but a present process for one to target toward future. Consequently we think that the concept of change is pivotal in comprehending cognition in general and its consciousness character in particular.

We are further motivated to make sense of cognition using the universal notion of change because the human brain, as the primary premise of cognition, displays in its structures and functions the same patterns as numerous other systems throughout nature (Linkenkaer-Hansen et al., 2001; Eguíluz et al., 2005; Mäkelä and Annila, 2010; He et al., 2013). For example, neural activity is no different from seismic activity, both comply with power

laws (Touboul and Destexhe, 2010). A neuronal network, just as the World Wide Web, has a skewed distribution of nodes' degrees (van den Heuvel et al., 2008). Neural activity exhibits waves, oscillations, spiraling sequences and at times chaotic behavior just like economic activity displays cycles, trends and occasionally tumultuous conducts (Schroeder, 1991; Huang et al., 2010; Friedman and Landsberg, 2013).No question, the ubiquitous patterns have been recognized in diverse disciplines including neuroscience (Chialvo, 2010), but the main point remains unappreciated: The common characteristics result from natural processes, that is, from series of changes.

Evolution of any kind, when broken down to a succession of changes, can be given by an equation of motion. In this way the thermodynamic theory explains the recurrent patterns to result from least-time free energy consumption (Sharma and Annila, 2007). In other words, the skewed distributions are energetically optimal, and hence their cumulative sigmoid growth and decline curves are also optimal in energetic terms. The power laws, in turn, are ubiquitous by being central approximations of the sigmoid curves. The evolutionary equation asserts that natural systems evolve in non-deterministic and path-dependent manner (Annila and Salthe, 2010a). Also cognitive processes, unmistakably learning and decision making, share these universal attributes (Arthur, 1994; Bassanini and Dosi, 2001; Anttila and Annila, 2011). For these reasons we are motivated to employ the general theory to make sense of cognition and especially of its seemingly elusive conscious character.

Disciplines have branched far from their common stem in natural philosophy, and hence holism is today an unconventional tenet. Thus, our assertion that the human brain is no different by its operational and organizational principle from any other system in nature may appear odd and groundless at first sight. To justify our reasoning we will begin by outlining the thermodynamic theory (Chapter 2) and thereafter work insight to consciousness by relating the holistic perspective to various puzzles, phenomena, and well-known stances (Chapter 3). Finally, we summarize conclusions of the thermodynamic tenet to further debate and discourse (Chapter 4). As it will become apparent, our study does not yield groundbreaking resolutions, rather it substantiates common sense by a firm formalism.

THERMODYNAMICS OF OPEN SYSTEMS

We reason that the human brain is no different from other systems in nature because its structures and functions display the ubiquitous patterns, i.e., distributions that sum up along sigmoid curves which, in turn, mostly follow power laws. Hence, the brain ought to be described and comprehended in the same way as any other system.

To this end the general principle of nature is known, in fact by many names, most notably as the second law of thermodynamics, the principle of least action and Newton's second law of motion. These three laws appear as if they were distinct from one and other when erroneously expressed in their determinate, i.e., calculable forms. For example, textbooks tend to present Newton's second law of motion so that force $\mathrm{F} = m\mathrm{a}$ equals mass m times acceleration $\mathrm{a} = d_t\mathrm{v}$, i.e., the change in velocity v. However, Newton himself wrote that the force $\mathrm{F} = d_t\mathrm{p}$ equals a change in momentum p, which yields by the definition $\mathrm{p} = m\mathrm{v}$ not one but two terms $\mathrm{F} = md_t\mathrm{v} + \mathrm{v}d_t m$. The change in mass relates via $dm = dE/c^2$ to dissipation of photons ultimately to the cold space. Dissipation is inherent in any change, and hence it is also integral to cognition.

Likewise, the principle of least action in its original form due to Maupertuis includes dissipation in contrast to the familiar constant-energy, hence deterministic Lagrangian (De Maupertuis, 1746; Tuisku et al., 2009). Furthermore, statistical mechanics, as the probabilistic many-body theory underlying thermodynamics, can be formulated for open dissipative systems. However, when imposing the constant-energy condition, statistical mechanics limits to stationary systems (Kondepudi and Prigogine, 1998).

Dissipation, despite being an integral component of any change, may still appear as a downright secondary byproduct of neural activity. Yet, when a systems theory misses even a single and seemingly insignificant photon, such a theory does obviously not account for everything and leaves room for unaccounted effects, surmise and speculation. Of course, when probing neural activity in practice, knowledge of numerous factors will remain imperfect, but all the more the theory's bookkeeping of causes and effects, i.e., forces and ensuing motions, ought to be perfect.

The Physical Basis

Today, when complex systems are more often modeled and simulated than described and explained, our ambition to account for everything with accuracy and precision extending down to a single photon might seem as an exceptional, perhaps even as an unattainable and abstract attempt. Therefore, it is worth stressing that for us an explanation is genuine only when it relates to everyday experience. For instance, the well-known conjecture that quantum mechanics could underlie consciousness (Bohm, 2002; Pylkkänen, 2014) does not qualify for us as an explanation, because entangled and superposed states do not make sense to us. The legendary illustration of a microscopic system being in two states at the same time by a cat being alive and dead at the same time simply does not seem sensible to us. The observed indeterminism implies to us that we just do not know the state of cat that goes missing. Likewise, we refute the idea in statistical mechanics that an observable state would sum up from a probability distribution of microscopic configurations, because the microstate (Mandl, 1971), in contrast to the state, is a theoretical concept without a discernable counterpart. In practice one microstate cannot be distinguished from another.

Surely, our stance can be argued against by claiming that not everything is necessarily tangible to the human being, but then again no observation is either free from some interpretation. Mere numbers mean nothing. Thus, mere agreement with recordings is no guarantee that non-determinism and purported non-localism as well as emergence could not be explained without conceptual conundrums (Annila and Kallio-Tamminen, 2012). It is worth noting that Schrödinger equation is devoid

of dissipation (Griffiths, 2004), and hence it does not comply with observations that all changes are dissipative. Likewise, the textbook statistical mechanics accounts for the system when at thermodynamic equilibrium, not when in dissipative evolution from one state to another (Gibbs, 1902).

We think that the theory of cognition ought to be given in the form of an equation because mathematical notation leaves less room for ambiguity than natural language. By the same token, Darwin's theory, as the corner stone of biology, is not a theory by standards of physics but a narrative, albeit a conceivable one. Then again, an equation alone is no theory. Namely, when variables of a mathematical model fail to correspond to causes and effects, there is no enlightenment.

Traditionally rules and regularities have been deduced from meticulous measurements. Kepler's laws are examples of formalized observations. In neuroscience this approach is hardly an option. Recordings do not reproduce precisely enough to infer an equation of motion. Instead mathematical models, such as Markov chains that mimic data, more or less, are fashionable in providing predictions, at least trends (Laing and Lord, 2009). However, the model parameters do not map one-to-one with causes and effects. The introduced statistical indeterminism, i.e., randomness without reason, is not a substitute for non-determinism. It follows from the path-dependence of natural processes.

To obtain an equation by starting from an axiom is yet another possibility. For example, the axiom that inertia is distinguishable from gravity, known as equivalence principle, underlies general relativity (Misner et al., 1973). In neuroscience this approach for finding axioms does not appear amenable either. Recordings hardly display invariants to get hold of the foundation. Nonetheless, one may construct the theory by inferring or postulating self-evident axioms and challenge only ensuing conclusions (Tononi, 2008; Tononi and Koch, 2015). However, we would prefer axioms that are directly verifiable in terms of physics, but then neuroscience cannot stand out as a distinct discipline, its concepts cannot be chosen self-sufficiently, and its objects of study cannot be singled out as unique phenomena.

We find the ancient atomism (Berryman, 2011) as a sound and solid stance. It claims that everything comprises indivisible basic building blocks. Since the atom, as a chemical element, turned out to be divisible, the most elementary constituent was renamed as the quantum of action. The quantum of light is its most familiar embodiment. The human eye can register even a single photon and our skin is sensitive to photon influxes and effluxes that are sensed as hot and cold. Thus, the photons are real by everyday experience, and hence the quantum of action qualifies for us as a tangible entity. The quantum-embodied atomism is further motivated because every chemical reaction will either emit or absorb at least one photon. Also annihilation of matter with antimatter yields only photons. Also other observations substantiate the axiom that everything, and hence also cognition, is ultimately embodied by the quantized actions (Annila, 2010, 2012; Varpula et al., 2013). The atomism is not new to neurosciences either. It has been formulated in at least in neurophysiological context (Fingelkurts et al., 2009, 2010a).

The quantum of action has energy E and time t as its attributes, or equivalently momentum p and wavelength x, so that their product is invariant known as Planck's constant

$$h = Et = \mathrm{p} \cdot \mathrm{x}. \tag{1}$$

In other words, energy and time do not exist as such. They are characteristics of the quanta (Annila, 2016). Surely, Equation (1) is mathematically equivalent to the textbook form $E = hf$, where frequency $f = 1/t$, but then it is not evident that h is the quantum's measure. The invariance means, for example, that the wavelength will change along with changing momentum but the photon itself remains intact. Consequently, we find virtual photons as an abstract theoretical construct without correspondence to reality (Peskin and Schroeder, 1995).

A system changes from one state to another by acquiring quanta from its surroundings or by losing quanta to its surroundings. Thus, the change in energy is, according to Equation (1), invariably accompanied with the change of time. This is common sense. For example, a chemical reaction will progress in the course of time by acquiring or expelling quanta that carry energy as heat until a stationary state has been attained. Many a biological system is recurrently subject to changes due to its changing surroundings. Therefore, animate will hardly ever attain and reside in thermodynamic steady states. Specifically, the central nervous system is incessantly receiving and sending impulses to its surroundings comprising the body and beyond.

In practice there is hardly a way to keep track of all quanta embodying even a microscopic system, but formally the system can be described with the precision of one quantum. This is not only a remarkable but consequential resolution. Not only is the neural network no different from any other energy transduction system, but the atomistic axiom excludes other factors. Put differently, if one were to argue that consciousness is not embodied by quanta, the stance would violate causality by introducing some other constituents from nothing. That is to say, a cause of any kind is ultimately nothing but an energy difference, i.e., some form of free energy. Its ensuing effect is nothing but a quantized flow of energy. Thus, causal power, as a characteristic of consciousness (Kim, 1992), is inherent in the thermodynamic description.

Our approach to account for the entirety in terms of quanta undoubtedly resembles reductionism. The idea that the system is nothing but the sum of its parts has been refuted, for instance, by referring to emergent characteristics of consciousness. Likewise, properties of a molecule cannot be inferred from properties of its constituent atoms. However, the molecule does not form only from atoms, but also from the photons that couple from surroundings to the synthesis (Pernu and Annila, 2012). If these quanta are not included in the description, obviously the molecular characteristics remain unaccounted. Conversely, no new property will appear from mere permutations of systemic constituents. Instead a novel characteristic will appear along with the flux of quanta from the surroundings to the system or *vice versa*. In other words, the monistic account (Stoljar, 2015) is in fact complete when every quantum of action is included. This essential role of surroundings in emergence has been pointed

also in neuroscience (Rudrauf et al., 2003; Revonsuo, 2006; Fingelkurts et al., 2010b).

We realize that our physicalism does not immediately enlighten, for instance, subjective conscious experience, i.e., qualia, which is the contested concept about *the ways things seem to us* (Dennett, 1988; Chalmers, 1995). True enough, one does adhere meanings beyond mere perception. For instance, the sensation of red color is not only about registering corresponding energy of the photons at the retina, but the influx will trigger processes that involve more. What exactly is implicated may not be easily exposed in practice, but in any case we maintain that the supervening processes can be formally described with the exactness of one quantum.

The Systems Description

The above preliminaries pave the way for the formal description of a system. Since all entities are understood to comprise of the basic building blocks, any entity can be related to any other in energetic terms. So, all those entities that one chooses to refer to as the system can be placed on an energy level diagram (**Figure 1**). This description can be formalized mathematically irrespective of complexity (Mäkelä and Annila, 2010).

According to the general theory of many-body systems the state can be expressed concisely and completely in terms of probability P. It is the measure of what it takes to have, for example, a pool of certain neurotransmitter molecules in a synaptic vesicle. Undoubtedly, it will take a lot of things. Precursors are needed for syntheses of transmitters as well as energy-rich chemicals are required to power the production. Moreover, machinery for the syntheses and molecular transport is necessary. In practice we do not know all factors that are involved in attaining the particular state of synaptic vesicle. Nonetheless, we may formally denote the probability P_j for the pool of neurotransmitters, in numbers N_j, by accounting for all those vital ingredients, each in numbers N_k, using the product form P_k. It ensures that if any one of the vital ingredients is missing altogether, not a single neurotransmitter molecule will be found in the vesicle. Of course, it is not the mere number N_k of substrates that matters but also the substrate's energy attribute G_k. Specifically, P_j depends on the difference between energy $N_k \exp(G_k/k_BT)$ that is bound in the substrates and energy that is bound in the product $\exp(G_j/k_BT)$ as well as on the difference in energy that couples from surroundings via flux of photons to the synthesis of the j-entities from k-entities, i.e., $\exp(\Delta Q_{jk}/k_BT)$. Formally this dependence of P_j on energetics is given by

$$P_j = \left[\prod_{k=1} N_k \mathrm{e}^{-\Delta G_{jk}/k_BT} \, \mathrm{e}^{+i\Delta Q_{jk}/k_BT} \right]^{N_j} / N_j! \qquad (2)$$

for the population of N_j products. All energy terms are relative to the average energy of the system per particle, denoted by k_BT for historical reasons. The division by factorial $N_j!$ takes into account energetically equivalent permutations. It is worth emphasizing that these configurations, that the system cannot distinguish energetically, populate the same state. This is of course common sense. If one cannot distinguish one entity from another, one claims that they are identical. One's ability or any other system's ability to make a distinction requires ultimately recognition of some difference in energy. For the sake of clarity, imaginary part i in Equation (2) distinguishes energy in radiation, known as the vector potential, from quantized material forms of energy, known as the scalar potential (**Figure 1**). The probability of any other population can be denoted in the same way as P_j. Then the total

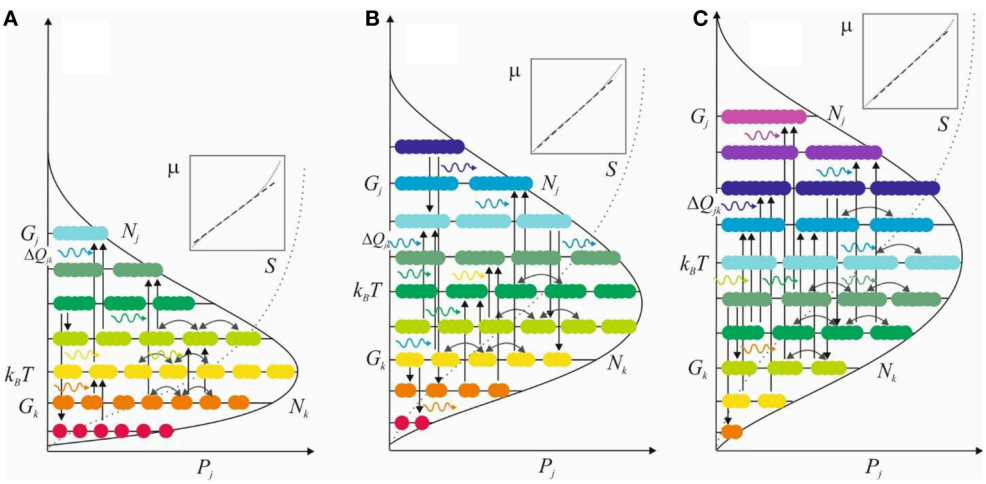


FIGURE 1 | System is portrayed in terms of an energy level diagram along its evolutionary path at three states (A–C). Each diagram pictures various populations N_k of entities, each with energy attribute G_k. Vertical arrows indicate paths of transformations, i.e., changes from k-entities in the population N_k to j-entities in the population N_j. Horizontal wavy arrows denote influx and efflux of photons that invariably couple to these transformations. Horizontal bow arrows, in turn, mean inconsequential exchange of indistinguishable entities. The system evolves, step-by-step, via absorptive and emissive jk-transformations from one state to another toward ever more probable partitions, denoted by $P = \prod P_j$, eventually arriving at a stationary-state balance where its average energy k_BT equals energy density in the system's surroundings. The outlined skewed partition accumulates along a sigmoid curve (dotted) which follows mostly a straight line on a log-log scale (insert) for entropy $S = k_B \ln P$ vs. [chemical] potential energy μ.

probability P of the whole system is simply the product

$$P = \prod_{j=1} P_j. \quad (3)$$

Although the status of a system is accurately and precisely given by the product form (Equation 3), one would prefer an additive measure to make comparisons. In statistical mechanics entropy S is the additive measure of a system's status. It is obtained from the logarithm of P

$$S = k_B \ln P = \frac{1}{T} \sum_{j=1} N_j \left[k_B T + \sum_{k=1} \left(\mu_k - \mu_j + i\Delta Q_{jk} \right) \right] \quad (4)$$

by multiplying with k_B for historical reasons. The shorthand notation $\mu_k = k_B T \ln N_k + G_k$, known as chemical potential for the logarithm of density in energy $N_k \exp(G_k/k_B T)$ is expedient (Gibbs, 1902). Also Stirling's approximation $\ln N_j! \approx N_j \ln N_j - N_j$ is convenient. It holds the better the larger N_j. For instance, when $N_j > 100$, the error relative to $\ln N_j! < 1\%$. In the Equation 4 the first term sums all energy that is bound in the system's entities, including, for example, the pool of neurotransmitter molecules. The second term sums all energy differences, i.e., free energy terms that reside within the system as well as between the system and its surrounding systems, including, for instance, differences in electrochemical potentials across the vesicle's membrane. All these forces drive the system and its surroundings toward thermodynamic balance. The resulting change in entropy is obtained from the time differential

$$\frac{dS}{dt} = \frac{1}{T} \sum_{j=1} \frac{dS}{dN_j} \frac{dN_j}{dt} = \frac{1}{T} \sum_{j=1} \frac{dN_j}{dt} \sum_{k=1} \left(\mu_k - \mu_j + i\Delta Q_{jk} \right) \quad (5)$$

Where dN_j/dt denotes the change in N_j that result from consumption of free energy $\Sigma\mu_{jk} - \mu_j + i\Delta Q_{jk}$. For example, the accumulation rate of transmitter molecules in the synaptic vesicle

$$\frac{dN_j}{dt} = \frac{1}{k_B T} \sum_{k=1} \sigma_{jk} \left(\mu_k - \mu_j + i\Delta Q_{jk} \right) \quad (6)$$

is proportional to free energy by rate parameters $\sigma_{jk} > 0$. Each parameter is associated with a transformation mechanism, such as an enzyme, that consumes free energy in the form of chemical energy. An energy transducer of any kind is according to the scale-free theory a system of its own. Hence, it is subject to changes too. For example, a mutation in a gene may lead to an altered catalytic activity, and hence affecting the flow rate from the substrates to the products and *vice versa*.

Although we do not know the details of how the system evolves from one state to another, the formal scale-free expressions (Equations 1–6) include every detail down to the precision of one quantum. In other words, the numerous flows of energy in a complex system are all formally included in Equation (5). Since there is no option to create the quanta from nothing or to destroy the quanta for nothing, the flows will have to direct along the least-time paths of free energy consumption. In biological terms evolution from one state to another will naturally select those means and mechanisms that facilitate survival. The scale-free patterns are consequences of this least-time imperative (Mäkelä and Annila, 2010).

When inserting Equation (6) to Equation (5), the quadratic form proves the second law of thermodynamics, i.e., $dS \geq 0$ both for $\Sigma\mu_k - \mu_j + i\Delta Q_{jk} > 0$ and < 0. The thermodynamic entropy cannot ever decrease. For example, the neurotransmitter population will increase $dN_j > 0$ when there are resources $\Sigma\mu_k - \mu_j + i\Delta Q_{jk} > 0$ for its production. Conversely, the population will decline $dN_j < 0$ when $\Sigma\mu_k - \mu_j + i\Delta Q_{jk} < 0$. Thus, their product in Equation (5) is always non-negative.

It is worth stressing that entropy by Equation (4) is a measure of bound and free energy, not of disorder or the number of microstates. Although the definition of P by Equation (3) differs from the one referred to by the free-energy principle (Nicolis and Prigogine, 1977; Haken, 1983; Friston et al., 2006; Friston, 2010), the idea is the same: Evolution of any kind directs toward a free energy minimum state. The least-time principle parallels also of the principle of least effort (Zipf, 1949).

According to the holistic tenet any system is at the mercy of its surroundings. Therefore, changes in surroundings will manifest themselves as activity that will move the system in quest of regaining balance. For instance, when a neuron reverts its polarity, the synaptic vesicle will respond by releasing neurotransmitters. Conversely, during repolarization the transmitter population will recover, provided that there is chemical potential available for the restoration. Note, irrespective of which way the energy gradient lies between the system and its surroundings, free energy can only decrease, and hence entropy can only increase. At the maximum entropy state all forms of free energy have been consumed, and accordingly all energy is bound in the stationary populations. Then there are no net forces that would drive the system away from the thermodynamic balance. Many a living system hardly ever resides in thermodynamic balance because its surroundings keeps changing, but formally Equations (2–6) do express the path-dependent, and hence intractable evolution toward balance as well as complex dynamics at the balance.

It is worth underscoring that entropy by Equation (4) does not convey any additional information about the system than what is given in energetic terms, i.e., by multiplying S with T. Thus, the least-time free energy consumption means that entropy will not only increase but it will increase at the maximal rate. Importantly, S does not relate to disorder, i.e., incoherence. Order or disorder is no end in itself but a mere consequence of free energy consumption. Organization, just like disorder, follows from the quest of consuming free energy. The widespread but unwarranted association of entropy with disorder dates back to the derivation of entropy for closed systems by Boltzmann. Obviously, when the system is defined as invariant in energy, nothing can change per definition. However, life is all about changes, and for that matter, in the expanding Universe no stationary motion will last forever either.

A Neural Network as a Thermodynamic System

Thermodynamic terms are commonly used in metabolism, but seldom applied in the context of cognition. However, there is

no principal difference. Electromagnetic potentials of nerve cells arise from chemical potentials, and hence neuronal signaling can be expressed alike, in terms of the scalar potential $U = \int \mu dN$ due to bound quanta and in terms of the vector potential $Q = \int \Delta Q dN$ due to absorbed or emitted quanta. Since Equations (1–6) apply also for electromagnetism (Tuisku et al., 2009), a network of neurons engaged in cognition is by thermodynamic principle no different from a reaction network of chemical compounds involved in metabolism. The neural network also evolves, just as the chemical reaction mixture, by consuming free energy in least time (Hartonen and Annila, 2012). For example, evolution of a neural network from one state to another is about accumulating products, e.g., physically embodied representations of experiences and memories. Likewise, lapses and larger losses of memories or mental skills invariably involve changes in the neural network. However, we make no attempt to specify what these changes are in detail, say in diagnostic terms, we only claim that whatever they are, they all are formally contained in the systems theory.

In concord with naturalistic consent we reason that all cognitive processes, for example, learning is ultimately embodied in neuronal systems or in some other systems. The chosen definition of a system is inconsequential because all systems amidst surrounding systems are perceived to evolve, develop and mature, that is, to change from one state to another in one way or another by consuming free energy in least time. Therefore, irrespective of how one chooses to demark the system from its surrounding, the bookkeeping of quanta in the system and of the quantized influxes and effluxes across the interface is perfect.

The freedom for one to define a system does not mean that the classification would be meaningless. Namely, a natural interface is there where strengths of interactions change significantly. For example, neurons in the central nervous system (CNS) are more strongly connected to each other than to rest of the body. Accordingly, the brain and spinal cord are recognized as subsystems of CNS, and, in turn, medulla, pons, thalamus, hypothalamus, cerebellum, hippocampus, basal ganglia, etc., can be recognized as subsystems of brain by their high internal connectivity. The natural interfaces are not impermeable, only fluxes across them are less intense than fluxes within the system.

Connectivity as the natural determinant of a system manifest itself, for instance, when connections across callosum are progressively reduced. The split-brain condition, where the two lobes behave as distinct systems, does not emerge gradually but abruptly (Tononi and Koch, 2015). We claim that the threshold is reached when the free energy consumption via inter-hemisphere connectivity falls significantly below the free energy consumption via the intra-hemisphere connectivity. The underlying principle is the same when two persons grow apart, they will at some stage speak out the split. Likewise, when two populations in a country grow more and more apart from each other, they will at some point declare themselves as two independent nations.

It is worth stressing that the least-time imperative does not specify any particular outcome, e.g., a split or union. This means, for instance, that a memory circuit has evolved to consume free energy by making an "appropriate" recollection, not by recollecting an event exactly as it actually took place. This energetically optimal conduct is customarily referred to as survival. One may easily imagine circumstances where a frank, yet unfaithful recollection will be vital, and scenarios where an exact recollection would be fatal. The same conclusion has been expressed in terms of utility in the context of vision (Purves et al., 2015). Indeed, it is no new thought to think of cognition as a means of survival, but still some might find it unusual to speak about the fittest in thermodynamic terms without making any distinction between animate and inanimate. It is this universality of thermodynamics from which we draw insight to consciousness.

CONSCIOUSNESS BY THE THERMODYNAMIC TENET

According to thermodynamics there is nothing extraordinary about consciousness; why it exists, what it does and how it arises. On the contrary, its existence, functions and arousal follow from the universal imperative. The Equations (1–6) express in quantitative forms the general biological position that consciousness is a result of evolution, among all other characteristics. Thermodynamically speaking flows of energy will naturally select those characteristic paths that will level off energy differences in least time. According to this perspective, consciousness integrates sensory and other inputs with recollections and representations from the past for coherent responses to consume energy gradients more effectively than by unconscious deeds. In concord with common sense a conscious person acts in a more meaningful way than an unconscious one. The augmented consumption of free energy means enhanced survival. In the following we will examine by the least-time free energy perspective some well-known questions and established stances about consciousness to enlighten its character.

On the Definition

One hand definitions serve to organize diversity of nature. On the other hand a dividing line creates a problem because everything depends on everything else. The border between one category and another is practical but in the end ambiguous when things change from one to the other. Ultimately one quantum of action is enough to make a change from one category to the other. Most notably it is hard to make a clear-cut distinction between living and non-living, although the notions of animate and inanimate themselves are practical. Similarly, it is unclear what exactly is meant by an economy. For example, is a bee hive part of an economic or an ecological system? Similarly, distinction between consciousness and unconsciousness is useful but ambiguous. The scope of awareness, wakefulness and sentience is wide and vague. Also the range of subjectivity and the sense of selfhood are broad and obscure. Capacity to experience and feel varies from one individual to another as well as from one moment to another in an individual.

Consciousness defies categorization precisely because it is functional. The change is the very characteristic of a conscious system. By the same token, a steady state does not display causal relationships, i.e., irreversibility. A mere exchange of

quanta without a net flow of energy between the system and its surroundings does not drive the system from one state to another.

Despite these arguments one could perhaps imagine of defining consciousness exactly by taking a snapshot of it. The still frame, however, would not represent any changes, so it would be devoid of the principle characteristic of consciousness. One could eventually think of enclosing the conscious system by a fictitious border, but only in a stationary system quantized trajectories are closed, i.e., bounded. Put differently, evolving and invariant, just as indefinable and definite, are mutually exclusive attributes.

It is no wonder that philosophers since Descartes and Locke have struggled to pin down essential properties of consciousness, because the definition depends on both the content and context of what is deemed as essential, in fact, functional. For example, search for neural, psychological and behavioral correlates is not free from a preset idea of what consciousness is. Physically speaking, the free energy consumption, i.e., functioning is proportional to the changes in energy, not to some absolute and invariant values of energy, i.e., stationarity. Therefore, the search for a set of neural events and structures implies as if consciousness was bounded by a definition rather than being an open operational notion.

It worth emphasizing that not only consciousness but also many other definitions are ambiguous by depending on the subjective choice of key characteristics. For example, the definition of an ecological community depends on what will be listed as its characteristic organisms. Likewise, the definition of a multi-cellular organism is dictated by the list of its cells. The cell, in turn, is defined by its molecules, and so on. The thermodynamic theory claims that definitions are ambiguous when the change is the principal characteristic.

Obviously our account of consciousness by the universal notation of physics encompassing everything reminds of panpsychism, the philosophy that the mind is not only present in humans but in all things (Seager and Allen-Hermanson, 2015). We see this thought to emerge from the correct comprehension that it is impossible to single out anyone evolving system, specifically consciousness, from its surroundings as well as from the accurate observations that all systems behave in the same way, that is, consume free energy in least time. Then again, there is hardly a point in equating the specific notion of mind with the general notion of an evolving system. Thus, we refer to mind merely as a practical term for what the brain does. Likewise, we choose to speak about consciousness merely as an attribute for an integrated system that is consuming free energy coherently.

Still, one might regard consciousness as an umbrella term, for example, in analogy to furniture which, as a term, includes tables, chairs, beds, etc. Since furniture refers to movable objects that support various human activities such as seating and sleeping, one should ask: What functions does consciousness support? Only to realize that the list will remain open. Thus, there is no closed definition for consciousness.

All in all, the trouble in defining consciousness appears to us as contrived. Problems stem from attempts either to single out or to separate consciousness from its surroundings or to attribute consciousness with some unique rather than universal characteristic. The renowned Cartesian dualism appears to us an unfortunate misinterpretation that *res cogitans*, i.e., the realm of thought would mean an immaterial domain and that *res extensa*, i.e., the realm of extension, would mean the domain of material things. Isn't Descartes only naming the system capable of interoception and exteroception as consciousness and referring to the rest as its surroundings so that their interactions convene in the brain? In our mind the purported qualitative distinction between material and immaterial is not his message. Thus, the mind–body problem of how non-physically labeled beliefs, actions and thinking, etc., relate to the physically embodied human being, appears to us utterly artificial.

On the Quantification

Although consciousness defies a closed definition, it is still quantifiable in terms of entropy (Equation 4). The irrevocable increase in entropy $d_t S \geq 0$ (Equation 5) implies somewhat paradoxically the state of consciousness, measured by S, can only increase. This is true when consciousness is understood as the attribute of an integrated system that consumes free energy relative to its surroundings, not relative to some absolute invariant reference.

Despite the relativeness of entropy, one may easily imagine in some absolute terms that the degree of consciousness has been increasing over eons when humans have been consuming energy differences relative to their energy-rich surroundings. Consciousness will flourish when supplies are rich and versatile. Conversely, when the surrounding resources narrow down so that the subject faces hunger, sleep deprivation, stress, etc., consciousness will decrease relative to the arbitrary absolute reference. However, the absolute value of entropy, high, or low is only imaginary because entropy is in relation to resources, i.e., a function of free energy (Equation 4). In biological terms the cognitive capacity will adapt to circumstances. In thermodynamic terms the cognitive system will regain balance with its surroundings either by acquiring or abandoning some subsystems and paths of energy transduction. Thus, a high level of consciousness is no end in itself but consciousness, as any other attribute of a system, develops and evolves to attain the entropy maximum, i.e., the free energy minimum state in a given circumstances. This is, of course, common sense. In poverty a high level of awareness is simply unaffordable.

In practice there is hardly a way to sum up numerous bound and free forms of energy to quantify consciousness. Above all it is difficult to gauge all forms of free energy that are represented in one's neural network. These energy differences reside between the system, known as the conscious self, and its surroundings. For one thing, one's perception of its surroundings is dynamic. For the other thing, one's identity, i.e., the system itself is an ambiguous and dynamic notion that prevents from defining it as distinct from its surrounding systems. For example, the problems of altruism and tragedy of commons resolve by identifying one's identity (Annila and Salthe, 2009; Anttila and Annila, 2011).

Although exact quantification of consciousness remains illusory its characteristics can be recognized from the determinants of entropy and its change (Equations 4 and 5). Entropy, as the measure of state, increases with increasing connectivity, not only by an increasing number of nodes,

such as neurons, but also by an increasing capacity and rates of mutual interactions (**Figure 1**). Thus, it is no coincidence that the brain with the fastest processing capacity and highest connectivity among all organs is the primary premise of consciousness. Conversely, the conscious capacity will degrade when connections and central nodes disintegrate but remains largely untroubled by solitary losses. Still, it is worth emphasizing that the comparison in absolute terms of entropy has no real meaning because any state of consciousness is in relation to its resources. High holism remains only imaginary when there are no resources and no means to attain it.

On the Subjectivity

The subjective nature of consciousness is inherent in the thermodynamic account (**Figure 1** and Equations 1–6). Namely, the system is the subject. The system is unique via its interactions with its surroundings. A flow of quanta from the surroundings to the system is not shared by any other system. For example, the photon that one's retina happens to absorb cannot be absorbed by anyone else. Accordingly, there is no objective way of defining or measuring any system because any observation will ultimately embody a unique flow of energy from the target to the specific observer. Thermodynamics of open systems acknowledges this uniqueness, i.e., the subjective character of nature. The theory works even when the system is defined at will, because it keeps track of all quanta that move between the system and its surroundings.

All meanings presented in various forms of free energy are subjective. The way things appear to one depend on who one is, that is to say, on evolutionary courses of human species and its ancestors as well as on one's own developmental processes and experiences. Common sensations imply the same origin and ordinary experiences where singular sensations indicate diversification. Since no objective account can be given, it is best to realize consequences of subjectivity. For example, one may begin by defining gamma waves as a necessary, yet insufficient characteristic of consciousness (Aru et al., 2002), and proceed by including other characteristics. When completed with one's list, one may label consciousness as impaired or disrupted when anyone of the predefined characteristics is missing or misplaced.

Neuronal and behavioral correlates of consciousness are undoubtedly needed for medical diagnoses and other purposes, but they are neither comprehensive nor objective. For instance, alcohol and other drugs, or spiritual and meditative techniques will alter the state of consciousness. This is sensed by the subject itself and other subjects, but differently since flows of energy are different. In turn, denial of impairment is a striking example where the subject's view of consciousness is deemed by others as disturbed (Hirstein, 2005). The subjective character of consciousness manifest itself pronouncedly when a patient, who has become blind, claims to see normally and continues to maintain the view despite all evidence to the contrary. This is perplexing, yet ordinary in another context. Isn't it only common that despite all evidence to the contrary, many an individual retains unrealistic thoughts about himself? Also, it is not unusual that one assures of recalling an event which never happened. Consciousness is not and it does not even aim to be a faithful, say objective or inter-subjective representation of reality. It is one's response to reality.

According to thermodynamics a conscious system forms from its constituent systems, like any other integrated hierarchy, The conscious system will consume free energy along the least-time paths, irrespective of how irrational these paths are judged "objectively" by other systems. For example, the changed meaning of a percept demonstrates how a tapered connection will redirect signals, i.e., flows of energy, from a sensory system to an "incorrect" locus at the cortical system. It is odd but still understandable that one may sense the sound of trumpet as "scarlet" (Krohn, 1892). The erroneous outcome is no different from a train arriving on a wrong platform because of a misplaced switch along the track. Put differently, the curious complications are not normal but natural according to the scale-free thermodynamic imperative. Our viewpoint of subjectivity as a natural characteristic complies with monistic consent that consciousness is a real subjective experience embodied by physical processes in the brain. This view is compatible with so-called biological realism at the interface between neural and mental phenomena (Revonsuo, 2006; Freeman, 2007; Fingelkurts et al., 2009, 2010a, 2013).

On the Hierarchy

The scale-free thermodynamic theory pictures the conscious system as comprising of systems (Salthe, 1985; Chialvo et al., 2007; Fingelkurts et al., 2013; Werner, 2013). Consciousness supervenes via least-time energy transduction from lower-level systems, say neuronal networks that represent sensations, coordination, memories, etc. In other words, knowing with oneself integrates existing systems with inputs from surroundings. This is to say, consciousness emerges in a form that best serves the least-time imperative rather than being a comprehensive report of either the state of mind or the state of surroundings. This conclusion about consciousness, as an integrated hierarchal construct, agrees with the impression that consciousness is the opinion or internal feeling that one has from what one does. Also that consciousness is deemed as unitary we understand as the coherent outcome of integration, not that it would mean a monolithic entity.

The view of consciousness supervening lower-level processes parallels the proposition of various narrative fragments, "drafts," coming together the way a coherent behavior of an individual calls for (Dennett, 1991; Chafe, 1994; Varela, 1999; Freeman, 2007; Fingelkurts et al., 2010a, 2013). The need in thermodynamic terms is a force that will expire by the least-time free energy consumption. In view of that it is natural that new aspects about oneself will surface to one's mind first when one senses corresponding driving forces. As long as one has no mechanisms to sense such forces, it makes no difference if someone else is aware of them. The blind is unaware of her beautiful face, but when learning about it from others, may make all the difference. In general, when sensory outputs from the surroundings are deprived by and large, it will become difficult to maintain a focused state of consciousness. The loss of external energy gradients results in a peculiar state of consciousness where theta waves prevail (Ballard, 1986). These low-frequency

oscillations disperse farther away than gamma waves. The extended scale of coherence underlines that consciousness is at its brightest as a focused construct. Yet, consciousness does not reside at any distinct locus in the neuronal network but integrates functional loci to an attentive response (Baars, 1988; Seth et al., 2005; Revonsuo, 2006; Tononi, 2008; Fingelkurts et al., 2010a, 2013; Marchetti, 2012; De Sousa, 2013). Then again, holism is emphasized when an optimal reaction recruits a broad range of processes, including also unconscious functions.

Moreover, consciousness embodying to-and-fro flows of energy (**Figure 1**) is consistent with observations that activity in primary sensory areas alone is insufficient for consciousness (Koch, 2004). Higher brain areas, especially the prefrontal cortex is involved in a range of cognitive functions, so that executive functions sum up from frontal cortex inputs and also so that neural activity propagates down to sensory areas (Crick and Koch, 2003). These up-and-down flows, so to speak upward and downward causation (Kim, 1984; Meyering, 2003), are consistent with a conscious system resulting from integration systems for the least-time free energy consumption (**Figure 1**).

On the Hard Problem

The so-called hard problem of consciousness is about how a physical process in the brain gives rise to subjective experience (Chalmers, 1995). The eminent claim is that even complete knowledge of the brain would not yield complete knowledge of conscious experience. The assertion means, for example, that even if one knew everything about how the brain processes colors, one would not know what it is like to see them.

According to thermodynamics subjectivity is the characteristic of any system. It is the only option. Subjectivity is not only associated with experience, but equally so with information processing such as reasoning, reporting, focusing attention, etc. Since many immediate stages of information processing at sensory organs are known in quite some detail, it may seem as if there were nothing subjective in the elementary processes, e.g., following the photon absorption at retina. But there are subtle differences among the involved entities. One retinal molecule, as a system of atoms, may seem identical to another one, but each setting is unique. The energy differences, say electromagnetic fields, about the molecule are dictated by everything else, e.g., by other molecules, whose coordination is not identical, i.e., symmetrical for anyone molecule. When the surroundings is unique, also the system is unique, which manifests itself, for instance, as a unique molecular conformation. Undoubtedly it would be very difficult to resolve these subtle differences, e.g., fine structure of electronic orbitals imposed by the surrounding fields. This degree of subjectivity, i.e., energy differences between various retinal molecules, is much smaller than that higher up in hierarchy. Using a powerful microscope there is no difficulty to resolve differences in cells that house those seemingly similar retinal molecules. Unmistakably the cells are subjects. Further up along the line of information processing there are more and more diversity, i.e., energy differences among representations. Therefore, we claim that there is no qualitative difference between the elementary percept of a color that is defined by the photon wavelength and the color-induced subjective experience that is represented uniquely by numerous energy attributes of a neuronal network. The degree of subjectivity is ultimately gauged in energetic terms, and hence the subjective experience does not single out from other phenomena.

The specific experience, i.e., the particular series of changes in one's neural system, depends on one's history. The past processes dictate what paths are available as well as what forms of free energy are at disposal to open up new paths or to close down existing paths to represent the experience. Therefore, what exactly one will experience beyond mere perception of light depends on these diverse assets that one has accumulated during life and inherited from ancestors as well as on forces that are imposed by the surroundings. For example, the experience will be moderated when the visual stimulus is accompanied with sound or sense of touch (Witten and Knudsen, 2005).

No question, the mere perception of color is a simpler and more predictable process than the full experience, simply because changes in energy are smaller and less dispersed at the retinal molecules than those associated with the experience of color at cortical levels. Still, we see no evidence that the two processes would be qualitatively different from each other. Put differently, we cannot see that introspection, as knowing about one's mental life, and phenomenality, as having experience about *something it is like*, would be qualitatively distinct from each other. For the same reason, not all of that what is conscious can be categorized simply as introspective or phenomenal. A finer classification beyond introspection and phenomenality is conceivable (Lycan, 1996), but to us the reductionist approach when missing the integrated character of consciousness, does not seem particularly insightful.

Undoubtedly certain aspects of cognition are more accessible for one to report verbally, reason and to control than others such as experiences of sounds, sensations, emotions, feelings, and others coined as qualia. Nevertheless, we see no line of demarcation between introspection and phenomenality. Isn't it exactly about a fine line between introspection and phenomenality why one admires an artist who is able to portray something one cannot quite picture and spokesman who is able voice something what one cannot quite express?

We do not deny that there are various aspects about consciousness, but their categorization is ambiguous, subjective and circumstantial. For example, there are numerous reports from battlefields when pain is not experienced (Morrison and Bennett, 2006). The subject recognizes the loss of a leg and even reasons its immediate consequences pretty much the same way as his comrades, by shouting "bring a stretcher". The tense circumstances call for vital activities that suppress experiencing the loss thoroughly. Cognition focuses for survival, that is, for the least-time free energy consumption. Only later, when circumstances allow, the meaning of loss, *something it is like*, will be sensed beyond a verbal account. No words will say it all, because walking is distinct from speaking. No images will expose it all either, because walking is distinct from seeing. For one thing, pain is experienced because touch with the leg is lost. For the other thing, agony is experienced because one's identity is at stake. The leg is an integral part of oneself. Sorrow gauges the loss of one's compromised future possibilities as a disabled. One is in for

a major restructuring of neuronal representations of oneself to match the new state of affairs. Yet, all what is experienced due to the loss of a leg is ultimately commensurable in terms of free energy. Loss of a toe as a less devastating experience would entail a smaller revision in one's free energy spectrum.

It is common that music, say a certain melody, will trigger a strong subjective experience, when one associates a whole lot with the piece. Similarly, a familiar scene or a memorable scent might move one from one state of consciousness to another. Curiously, many a scientist has described the moment of a discovery as an elation (Birney, 2013). Apparently mere introspection is not enough to construct the full meaning of a sensational discovery. Only the experience by integrating a whole lot more does do the full justice.

Consciousness as an integrative process is best comprehended in holistic terms. In a sense consciousness amplifies, or more accurately inflates, an elementary sensory signal to an experience by integrating various assets from the past. Also Locke's portrayal of consciousness as *the perception of what passes in a man's own mind* we like to read so that consciousness is an inflated perception. Likewise, the idea that consciousness is about broadcasting information around the brain from one's memory bank (Baars, 1988), we like take to mean that consciousness emerges principally from the existing assets whereas the primary trigger makes only a minor component in the final product. The Latin phrase *conscius sibi*, literally as knowing with oneself, provides yet a complementary perspective on the system of systems by emphasizing that consciousness is about sharing the present impulse with representations of the past.

On the Binding Problem

The least-time imperative provides perspective also on how brain creates from sensory inputs coherent perceptual experience. This binding problem (Revonsuo and Newman, 1999; Singer, 2001) comprises both the problem of how the brain segregates elements in an input pattern to discrete entities and the problem of how the brain constructs a phenomenological object from the entities. This formulation parallels the metastability concept (Kelso, 1995, 2012; Bressler and Kelso, 2001; Fingelkurts and Fingelkurts, 2004; Fingelkurts et al., 2009).

It is a common experience when listening to a foreign language, that one has a hard time to distinguish individual words. The separation of words is impaired because one's analyzer is not yet tuned to recognize contrasts, i.e., energy differences among sounds. Undoubtedly one's ear is capable of consuming energy differences in the changes of pitch, but in successive stages of neuronal processing the input fails to recruit one's memory to amplify the unacquainted input to meanings. The unfamiliar input does not trigger further consumption of free energy to produce meanings. Likewise, when facing an unusual view, perhaps after being knocked down all of a sudden, one struggles to discern objects in sight because the familiar reference, as the source of meanings, is tilted. Therefore, we argue that elements in sensory inputs are segregated by consuming free energy in least time. This involves mechanisms that have been established in the course of one's life as well as inherited from the course of evolution. Thus, the outcome of segregation is subjective and context-dependent. One looks for meanings every day, yet practically every textbook of biology shuns the idea that there are meanings and purposes, physically speaking forces, in nature.

Consistently, the construction of an object from the segregated elements is guided by the least-time free energy consumption. The brain is equipped with mechanisms from the past to assemble an object from the segregated ingredients as well as from those ingredients that are available in memory for the integration process. It is not unusual to jump into conclusions which demonstrates that the construction is not and does not even aim to be faithful and consistent. It is biased by prior expectations, physically speaking, by energy gradients. One's conclusion is motivated by gradients in the evolving energy landscape that one's neuronal network is representing.

Admittedly, the thermodynamic tenet clarifies only the principle of how experiences are constructed, not mechanistic details of the processes, for instance, in terms of neuronal correlates. Yet, thermodynamics indicates that familiar stimuli will invoke more rapid, intense and wide-spread responses than unfamiliar stimuli, simply because "familiar" associates with what is already present. Put differently, when a lot of free energy is consumed, things make a lot of sense, and conversely nothing is consumed by non-sense. This may well appear as differentiation of brain states during meaningful stimuli vs. non-meaningful stimuli (Boly et al., 2015).

On the Intractability

It is in place to make few remarks on the integration process itself. The evolutionary equation (Equation 5) reveals that motions consume their driving forces which, in turn, affect the motions, and so on. Mathematically speaking the equation is inseparable, and hence it cannot be solved. Thus, there is no algorithm for consciousness. This point is familiar from the Chinese room argument (Searle, 1980). Consciousness emerges in a non-deterministic manner. Yet, supervenience is not a random, i.e., indeterminate, but a path-dependent process. In other words, one does not know exactly what one will think before thinking and one does not know exactly what one will experience before experiencing. We reason that due to intractability certain neural and behavior responses correlate with consciousness at times while at other times they appear as uncorrelated. Therefore, inability to make precise predictions about cognition are ultimately not due to complexity of the process but due to its path-dependent character.

From the thermodynamic perspective probabilistic inference models, most notably Bayesian models (Knill and Pouget, 2004; Bielza and Larrañaga, 2014) including prior knowledge, mimic natural processes, but the modeled probabilities are not faithful representations of energetics (Equation 3), only parameters. Moreover, the original Markov chain does not carry memory of past events, but only the current state determines the probability distribution of the next state. Even when the future state is modeled to depend on a sequence of the past states (Camproux et al., 1996; Seidemann et al., 1996), the ensuing projection, i.e., a trend does not parallel the actual energy gradient because in reality the force is affected by the motion itself.

On the Intentionality

To be directed toward a goal or thing is an apparent characteristic of consciousness (Evans, 1970). By the free energy perspective an intention means a force, that is, an energy gradient. A conscious mind gazes for various forms of free energy and exploits opportunities to consume them. The intention is fulfilled when associated free energy is fully consumed.

Since these driving forces are sensed by oneself, intentions are subjective. Ambivalent intensions imply that one has difficulties in constructing the resultant force. Also ambiguity about oneself may trouble the process. For example, one might commit a crime intentionally, only to realize later that the act in fact hurts oneself. In other words, one's identity was at the critical moment so narrow that only immediate forces manifested themselves.

In addition to conscious intentions there are also other forms of free energy that one is unconscious about, that is, are not integrated for coherence responses. Subliminal stimuli, e.g., presented as flashes, are too short substrates for the construction of consciousness. Nevertheless, these flows of energy will suffice to prime or bias one for an intended action (Loftus and Klinger, 1992). Thermodynamically speaking, the subliminal stimuli shape the energy landscape of one's mind to channel more readily a more comprehensive flow of energy later. In terms of neuroscience the subliminal stimuli lead to construction of some connections but apparently not enough to pave the full way to consciousness.

The thermodynamic tenet gives also a practical meaning to the philosophical concept of free will (O'Connor, 2014). Free will equates with free energy in one's disposal (Annila and Salthe, 2010b). One may execute at most as much as one has free energy in command. Non-determinism means that when one invests free energy to pursue along a path then some other paths are no longer affordable. This is to say, one is responsible as much as one is in capacity to consume free energy. For example, when in captivity, one is limited to act or even to express oneself, and hence free energy only in metabolic form powers free thinking. When deprived from free energy altogether, one has no choices whatsoever to respond by moving from one state to another, i.e., one is no longer responsive.

On the Rate

Why does the experience of oneself reel at a particular rate, at about one "frame" per 100 ms (Potter et al., 2014)? The coherence calls for synchrony (Singer and Gray, 1995). However, consciousness cannot cohere at the maximal firing rate of individual neurons because inputs must exist before they can integrate to the high-level construct. The common experience of escaping from danger by a reflex reaction demonstrates that more time is consumed in integrating awareness than that it takes for subsystems to act. This emphasizes that the notion of self is not a monolith but a composed union.

The reflex reaction demonstrates also that synchrony, e.g., gamma waves of reticular activating system, alone is insufficient indicator of consciousness. For example, visual information can control behavior without producing a conscious sensation. Fluent functions, i.e., automated sequences remain unconsciously generated until a change away from the ordinary happens. This implies that consciousness is a response, physically speaking a reaction to consume free energy, not in an algorithmic fashion, but in a non-determinate way.

Sleep by displaying a wide range of frequencies gives insight to wakefulness. After a daily dose of high-frequency stimulation the sleep, as a natural process, serves to revise the brain's connectivity spectrum toward a free energy minimum partition. In particular, long wavelengths of deep sleep amend long-range connectivity. Sleeping on a problem, exemplifies the value of balancing neuronal network by adjusting connections, i.e., making new thoughts (Bos et al., 2011). In turn, short wavelengths of sleep, characterized by rapid eye movement (REM), tweak the short-range connectivity, but without the objective of conscious free energy consumption. Therefore, vivid dreams do not necessarily make sense, i.e., cohere. The fact that most muscles are paralyzed during sleep, also implies that sleep, by its broadband iterative choreography, consumes in a path-dependent manner a whole range of imbalances in the brain's connectivity spectrum. In contrast, consciousness, by its high-band coherent activity, consumes various forms of free energy in the subject's spectrum of surroundings.

On the Problem of Other Minds

The question whether animals and organisms in general are conscious or not, is according to the thermodynamic tenet, like many other queries, troubled by ambiguity in defining consciousness. Nonetheless, it is possible to assess the *degree* of consciousness by the free energy consumption. Of course, one can still imagine a system that is conscious but not responsive, as in a lock-in syndrome (Nordgren et al., 1971), but it is inconceivable that a high-degree consciousness as the integrated response would emerge in the first place via minimal interactions with its surroundings.

The thermodynamic tenet asserts that consciousness is subjective in the spirit of the influential essay (Nagel, 1974) *What is it like to be a bat?* The writing argues that an organism is conscious *if and only if there is something that it is like to be that organism—something it is like for the organism*. A dog can be conscious about those forms of free energy that it can access, say, in the form of food and shelter. Its consciousness manifests itself as behavioral correlates, for example, defending its master. A bacterium is likewise consciousness of its free energy sources, say, in the form of sugar, by displaying chemotaxis as its coherent behavioral correlate. An ion is consciousness about its driving forces, say, in the form of electromagnetic fields by responding to them by motion and deformation.

So, what is it like to be a dog, bat or a bacterium? One may relate to another system as much as one shares the same means and mechanisms to consume the same forms of free energy. Apparently the human being shares with the dog some means of companionship, and hence one experiences collaborative behavior associated consumption of free energy somewhat similar to the dog. For one to know, what it is like to be piloting as a bat does, is possible only as much as one is able to navigate solely by hearing echoes. Still, if one were blind, one would value this skill as much as one is able to benefit from it. For one to know, what it is like to be a bacterium, is possible

up to the degree that one shares the same metabolic machinery. Of course for one it means hardly anything to digest few sugar molecules. In this respect it is not much of anything to be a bacterium. Though, something when the intake of sugar leads to the integrated function of chemotaxis. Obviously the question about other minds is not only about how much one shares with diverse animate and inanimate the same machinery of free energy consumption but more relevantly about how much shares with other human beings. Indeed, peer support is highly valuable.

The scale-free thermodynamics recognizes consciousness on other levels of natural hierarchy. For example, awareness of a nation accumulates from numerous activities, such as surveys, polls, and compilation of statistics on various things as well as from foreign sources by diverse means. To compare these activities with those that construct consciousness in a human being is, of course, nothing new. Already Hobbes wrote that *Where two, or more men, know of one and the same fact, they are said to be conscious of it one to another* (Hobbes, 1651). Also the Latin word *conscius* by literally meaning *knowing together* implies the scale-free character of consciousness. Specifically, when we claim that a society is consciousness too, we do not exactly mean Durkheim's collective conscience of beliefs and sentiments among members of a society (Durkheim, 1893) but refer to the natural processes that integrate the society for the coherent free energy consumption. These integrated actions, i.e., culture as a whole (Annila and Salthe, 2010b), can be regarded as meaningful or responsible, i.e., conscious. It is not about *analogy* between a consciousness society and a consciousness individual, it is about *equality* because the theory describes both systems exactly the same way.

The scale-free stance is of a practical value. Namely, it is much easier to observe how the society acquires and integrates information to act in an orchestrated manner and how the society cultivates its identity than it is to obtain data and relate recordings from the human brain to responses and development of the self. For example, certain structures, say, claustrum in the brain and a central hub in the computer network are alike critical for the construction of consciousness and situational picture of the nation (Crick and Koch, 2005). Consciousness is unable to manifest itself when claustrum is disturbed, and similarly the government cannot command when the central hub is out of power. However, not any one vital mechanism is the locus of consciousness. Instead consciousness integrates subsystems to a hierarchal construct in a path-dependent manner. This portrayal resembles the model for spotting meanings in percepts that integrates sensory information to virtual associative networks (Yufik, 1998). The thermodynamic theory bears also a clear resemblance to the integrated information theory of consciousness (Tononi, 2008). The universality in the laws of physics has been recognized earlier to underlie characteristics of consciousness such as criticality, self-organization and emergence (Fingelkurts et al., 2013). All in all, we have not put forward a more acute account on consciousness but merely related the prior comprehension to the profound and universal physical basis.

The universality of thermodynamics implicates also artificial consciousness (Russell and Norvig, 2009). Machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from that of a human being is not an issue, because thermodynamics does not make such a classification. Accordingly, a functionally equivalent but non-conscious organism, i.e., a philosophical zombie (Kirk, 2009) cannot possibly have the same survival advantage, i.e., capacity to consume free energy, as the conscious organism. Consciousness is not an epiphenomenon, but a reaction to forces. From this perspective realization of cognitive robotics is not an algorithmic problem, i.e., not a task of automating versatile and fine motor skills. Consciousness entails embedding evolutionary history and life experience, i.e., a long series of changes, to the machine. Without extensive free energy perspective the machine, just like a human being, will be small-minded. In other words such a creature will not have many processes to integrate for a coherent response. According to Equation (1) it will take time to acquire comprehensive cognitive capacity.

CONCLUSIONS

The least-time free energy consumption is hardly a new perspective on consciousness. Our interpretations and conclusions about consciousness are not original either. The main point is holistic. We regard consciousness, like any other phenomenon, as a manifestation of the natural law. Therefore, the proposed percept is falsifiable, not only by measurements of consciousness, but also by observations on anything else that will disprove the axiomatic basis, namely that quantized actions embody everything. Then again, due to our narrow knowledge and lack of expertize we might have reasoned incorrectly or imprecisely how the mind displays the universal principle in some cases, but such lapses do not jeopardize the theory itself, but call for a revision. Unquestionably our account of consciousness is far from being exhaustive, but hopefully it would be exemplary enough to motivate counterarguments and to provoke discourse.

The notion of information, so central in neuroscience, is conspicuous by its absence. To correct for this shortage we maintain that quanta embody also information in one form or another. Accordingly, information is subject to the universal imperative and its manifestation (Karnani et al., 2009). Specifically, information equates with free energy that is consumed by its receiver. In other words, the meaning of a message is subjective. This definition of information in the tangible terms of physics differs from that given by the abstract information entropy (Shannon, 1948).

Obviously there are other consciousness-associated notions besides information which we have not addressed. Just for curiosity we demonstrate the power of considering everything in terms of quanta by inspecting the association of mass with consciousness that was proposed in the popular thriller *The Lost Symbol* (Brown, 2009). The suggestion makes sense, because any change of state, say from conscious to unconscious, invariably involves either emission of quanta from the system to the surroundings or *vice versa*. The dissipated quanta carry energy E ultimately to the vacuum characterized by the squared speed of

light c^2. The ensuing change in energy dE relates to the change in mass by $dE = dmc^2$. The familiar relationship is pronounced in nuclear reactions, but discernable in chemical reactions, and inferable from gravitational changes.

Finally, one might ask: What a conclusion drawn from the least-time imperative stands out as the most insightful? Reminding of the subjective character of consciousness, it may be that only we find it somewhat surprising that consciousness is mostly generated from archives of mind and comparatively little from momentary inputs. On second thought in this way one will generate integrated responses in least time. This revelation allows us to understand also that it is only natural, not belligerent that an unconventional but profound perception hardly finds any place to take root in an established mind.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Annila, A. (2010). All in action. *Entropy* 12, 2333–2358. doi: 10.3390/e12112333

Annila, A. (2012). The meaning of mass. *Int. J. Theor. Math. Phys.* 2, 67–78. doi: 10.5923/j.ijtmp.20120204.03

Annila, A. (2016). Natural thermodynamics. *Physica A* 444, 843–852. doi: 10.1016/j.physa.2015.10.105

Annila, A., and Kallio-Tamminen, T. (2012). Tangled in entanglement. *Phys. Essays* 25, 495–499. doi: 10.4006/0836-1398-25.4.495

Annila, A., and Salthe, S. (2009). Economies evolve by energy dispersal. *Entropy* 11, 606–633. doi: 10.3390/e11040606

Annila, A., and Salthe, S. (2010a). Physical foundations of evolutionary theory. *J. Non-Equilb. Thermodyn.* 35, 301–321. doi: 10.1515/JNETDY.2010.19

Annila, A., and Salthe, S. (2010b). Cultural naturalism. *Entropy* 12, 1325–1343. doi: 10.3390/e12061325

Anttila, J., and Annila, A. (2011). Natural games. *Phys. Lett. A* 375, 3755–3761. doi: 10.1016/j.physleta.2011.08.056

Arthur, B. W. (1994). *Increasing Returns and Path Dependency in the Economy*. Ann Arbor, MI: University of Michigan Press.

Aru, J., Axmacher, N., Do Lam, A. T., Fell, J., Elger, C. E., Singer, W., et al. (2002). Local category-specific gamma band responses in the visual cortex do not reflect conscious perception. *J. Neurosci.* 32, 14909–14914. doi: 10.1523/JNEUROSCI.2051-12.2012

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge, MA: Cambridge University Press.

Ballard, E. (1986). Flow of consciousness in restricted environmental stimulation. *Imagin. Cogn. Pers.* 5, 219–230.

Bassanini, A. P., and Dosi, G. (2001). "When and How Chance and Human Will Can Twist the Arms of Clio," in *Path-Dependence and Creation,* eds R. Garud and P. Karnøe (Mahwah, NJ: Lawrence Erlbaum Associates), 41–68.

Berryman, S. (2011). "Ancient Atomism," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/win2011/entries/atomism-ancient

Bielza, C., and Larrañaga, P. (2014). Bayesian networks in neuroscience: a survey. *Front. Comput. Neurosci.* 8:131. doi: 10.3389/fncom.2014.00131

Birney, E. (2013, February 10). There is a mistaken view that scientists are unemotional people. *Science*.

Bohm, D. (2002). *Wholeness and the Implicate Order*. Hoboken, NJ: Routledge.

Boly, M., Sasai, S., Gosseries, O., Oizumi, M., Casali, A., Massimini, M., et al. (2015). Stimulus set meaningfulness and neurophysiological differentiation: a functional magnetic resonance imaging study. *PLoS ONE* 10:e0125337. doi: 10.1371/journal.pone.0125337

Bos, M. W., Dijksterhuis, A., and van Baaren, R. B. (2011). The benefits of "sleeping on things": unconscious thought leads to automatic weighting. *J. Consum. Psychol.* 21, 4–8. doi: 10.1016/j.jcps.2010.09.002

Bressler, S. L., and Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition. *Trends Cogn. Sci.* 5, 26–36. doi: 10.1016/S1364-6613(00)01564-3

Brown, D. (2009). *The Lost Symbol.* New York, NY: Doubleday.

Camproux, A. C., Saunier, F., Chouvet, G., Thalabard, J. C., and Thomas, G. (1996). A hidden Markov model approach to neuron firing patterns. *Biophys. J.* 71, 2404–2412. doi: 10.1016/s0006-3495(96)79434-1

Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.* Chicago, IL: University of Chicago Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 200–219.

Chialvo, D. R. (2010). Emergent complex neural dynamics. *Nat. Phys.* 6, 744–750. doi: 10.1038/nphys1803

Chialvo, D. R., Balenzuela, P., and Fraiman, D. (2007). The brain: what is critical about it? *AIP Conf. Proc.* 1028, 28–45. doi: 10.1063/1.2965095

Crick, F. C., and Koch, C. (2005). What is the function of the claustrum? *Phil. Trans. R. Soc. B* 360, 1271–1279. doi: 10.1098/rstb.2005.1661

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

De Maupertuis, P.-L. M. (1746). Les loix du mouvement et du repos déduites d'un principe metaphysique. *Histoire de l'Académie Royale des Sciences et des Belles-Lettres de Berlin* 267–294.

Dennett, D. C. (1988). "*Quining qualia,*" in *Consciousness in Contemporary Science*, eds A. J. Marcel and E. Bisiach (Oxford: Oxford University Press), 42–77.

Dennett, D. C. (1991). *Consciousness Explained.* Boston, MA:Little, Brown and Company.

De Sousa, A. (2013). Towards an integrative theory of consciousness: part 1 (Neurobiological and Cognitive Models). *Mens Sana Monogr.* 11, 100–150. doi: 10.4103/0973-1229.109335

Durkheim, E. (1893). *The Division of Labor in Society*. New York, NY: Free Press.

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-free brain functional networks. *Phys. Rev. Lett.* 94:018102. doi: 10.1103/PhysRevLett.94.018102

Evans, C. O. (1970). *The Subject of Consciousness.* London; New York, NY: George Allen and Unwin Ltd; Humanities Press Inc.

Fingelkurts, A. A., and Fingelkurts, A. A. (2001). Operational architectonics of the human brain biopotential field: towards solving the mind-brain problem. *Brain Mind* 2, 261–296. doi: 10.1023/A:1014427822738

Fingelkurts, A. A., and Fingelkurts, A. A. (2004). Making complexity simpler: multivariability and metastability in the brain. *Int. J. Neurosci.* 114, 843–862. doi: 10.1080/00207450490450046

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2009). Phenomenological architecture of a mind and operational architectonics of the brain: the unified metastable continuum. *New Math. Nat. Comput.* 5, 221–244. doi: 10.1142/S1793005709001258

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2010a). Natural world physical, brain operational, and mind phenomenal space–time. *Phys. Life Rev.* 7, 195–249. doi: 10.1016/j.plrev.2010.04.001

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2010b). Emergentist monism, biological realism, operations and brain–mind problem. *Phys. Life Rev.* 7, 264–268. doi: 10.1016/j.plrev.2010.05.005

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2013). Consciousness as a phenomenon in the operational architectonics of brain organization: criticality and self-organization considerations. *Chaos Solitons Fractals* 55, 13–31. doi: 10.1016/j.chaos.2013.02.007

Freeman, W. J. (2007). Indirect biological measures of consciousness from field studies of brains as dynamical systems. *Neural Netw.* 20, 1021–1031. doi: 10.1016/j.neunet.2007.09.004

Freeman, W. J., and Vitiello, G. (2009). Dissipative neurodynamics in perception forms cortical patterns that are stabilized by vortices. *J. Physics Conf. Ser.* 174:012011. doi: 10.1088/1742-6596/174/1/012011

Friedman, E. J., and Landsberg, A. S. (2013). Hierarchical networks, power laws, and neuronal avalanches. *Chaos* 23:013135. doi: 10.1063/1.4793782

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics.* New York, NY:Charles Scribner's Sons.

Griffiths, D. J. (2004). *Introduction to Quantum Mechanics.* Upper Saddle River, NJ: Prentice Hall.

Haken, H. (1983). *Synergistics: An Introduction. Non-Equilibrium Phase Transition and Self-Organisation in Physics, Chemistry and Biology.* New York, NY: Springer.

Hartonen, T., and Annila, A. (2012). Natural networks as thermodynamic systems. *Complexity* 18, 53–62. doi: 10.1002/cplx.21428

He, B. J., Daffertshofer, A., and Boonstra, T. W. (eds.). (2013). *Scale-Free Dynamics and Critical Phenomena in Cortical Activity.* Frontiers Media S.A.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation.* Cambridge, MA:MIT Press.

Hobbes, T. (1651). *Leviathan. Revised Edition.* Peterborough, ON: Broadview Press.

Huang, X., Xu, W., Liang, J., Takagaki, K., Gao, X., and Wu, J. Y. (2010). Spiral wave dynamics in neocortex. *Neuron* 68, 978–990. doi: 10.1016/j.neuron.2010.11.007

John, E. R. (2002). The neurophysics of consciousness. *Brain Res. Rev.* 39, 1–28. doi: 10.1016/S0165-0173(02)00142-X

Karnani, M., Pääkkönen, K., and Annila, A. (2009). The physical character of information. *Proc. R. Soc. A* 465, 2155–2175. doi: 10.1098/rspa.2009.0063

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge, MA: MIT Press.

Kelso, J. A. S. (2012). Multistability and metastability: understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 906–918. doi: 10.1098/rstb.2011.0351

Kim, J. (1984). Epiphenomenal and supervenient causation. *Midwest Stud. Philos.* 9, 257–270. doi: 10.1111/j.1475-4975.1984.tb00063.x

Kim, J. (1992). "Downward causation in emergentism and nonreductive physicalism," in *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism,* eds A. Beckermann, H. Flohr, and J. Kim (Berlin: de Gruyter), 119–138.

Kirk, R. (2009). "Zombies," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/sum2009/entries/zombies/.

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

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

Kondepudi, D., and Prigogine, I. (1998). *Modern Thermodynamics.* New York, NY:Wiley.

Krohn, W. O. (1892). Pseudo-chromaesthesia, or The Association of Color with Words, Letters, and Sounds. *Am. J. Psychol.* 5, 20–41. doi: 10.2307/1410812

Laing, C., and Lord, G. J. (2009). *Stochastic Methods in Neuroscience.* Oxford: Oxford University Press.

Linkenkaer-Hansen, K., Nikouline, V. V., Palva, J. M., and Ilmoniemi, R. J. (2001). Long-range temporal correlations and scaling behavior in human brain oscillations. *J. Neurosci.* 21, 1370–1377.

Loftus, E. F., and Klinger, M. R. (1992). Is the unconscious smart or dumb? *Am. Psychol.* 47, 761–765. doi: 10.1037/0003-066X.47.6.761

Lycan, W. (1996). *Consciousness and Experience.* Cambridge, MA:MIT Press.

Mäkelä, T., and Annila, A. (2010). Natural patterns of energy dispersal. *Phys. Life Rev.* 7, 477–498. doi: 10.1016/j.plrev.2010.10.001

Mandl, F. (1971). *Statistical Physics.* Chichester: Wiley.

Marchetti, G. (2012). Against the view that consciousness and attention are fully dissociable. *Front. Psychol.* 3:36. doi: 10.3389/fpsyg.2012.00036

Meyering, T. C. (2003). "Upward causation," in *Encyclopedia of Science and Religion.* Available online at: http://www.encyclopedia.com.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation.* San Francisco, CA: Freeman, W. H.

Morrison, V., and Bennett, P. (2006). *An Introduction to Health Psychology.* New York, NY:Pearson.

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914

Nicolis, G., and Prigogine, I. (1977). *Self-Organisation in Non-Equilibrium Systems.* New York, NY:Wiley.

Nordgren, R. E., Markesbery, W. R., Fukuda, K., and Reeves, A. G. (1971). Seven cases of cerebromedullospinal disconnection: the "locked-in" syndrome. *Neurology* 21, 1140–1148. doi: 10.1212/WNL.21.11.1140

O'Connor, T. (2014). "Free will," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/fall2014/entries/freewill/.

Perlovsky, L., and Kozma, R. (2007). *Neurodynamics of Higher-Level Cognition and Consciousness.* Heidelberg: Springer.

Pernu, T. K., and Annila, A. (2012). Natural emergence. *Complexity* 17, 44–47. doi: 10.1002/cplx.21388

Peskin, M. E., and Schroeder, D. V. (1995). *An Introduction to Quantum Field Theory.* Reading, MA:Addison-Wesley.

Potter, M. C., Wyble, B., Hagmann, C. E., and McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Atten. Percept. Psychophys.* 76, 270–279. doi: 10.3758/s13414-013-0605-z

Purves, D., Morgenstern, Y., and Wojtach, W. T. (2015). Perception and reality: why a wholly empirical paradigm is needed to understand vision. *Front. Syst. Neurosci.* 9:156. doi: 10.3389/fnsys.2015.00156

Pylkkänen, P. (2014). Can quantum analogies help us to understand the process of thought? *Mind Matter* 12, 61–91.

Revonsuo, A. (2006). *Inner Presence: Consciousness as a Biological Phenomenon.* Cambridge, MA:MIT Press.

Revonsuo, A., and Newman, J. (1999). Binding and consciousness. *Conscious. Cogn.* 8, 123–127. doi: 10.1006/ccog.1999.0393

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-P., and Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: francisco varela's exploration of the biophysics of being. *Biol. Res.* 36, 21–59. doi: 10.4067/S0716-97602003000100005

Russell, S. J., and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach. Upper Saddle* River, NJ:Pearson Education, Inc.

Salthe, S. N. (1985). *Evolving Hierarchical Systems: Their Structure and Representation.* New York, NY:Columbia University Press.

Schroeder, M. (1991). *Fractals, Chaos, Power Laws.* New York, NY:Freeman.

Seager, W., and Allen-Hermanson, S. (2015). "Panpsychism," in *The Stanford Encyclopedia of Philosophy,* ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/fall2015/entries/panpsychism/.

Searle, J. (1980). Minds, brains and programs. *Behav. Brain Sci.* 3, 417–457. doi: 10.1017/S0140525X00005756

Seidemann, E., Meilijson, I., Abeles, M., Bergman, H., and Vaadia, E. (1996). Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J. Neurosci.* 16, 752–768.

Seth, A. K., Baars, B. J., and Edelman, D. B. (2005). Criteria for consciousness in humans and other mammals. *Conscious. Cogn.* 14, 119–139. doi: 10.1016/j.concog.2004.08.006

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sharma, V., and Annila, A. (2007). Natural process – natural selection. *Biophys. Chem.* 127, 123–128. doi: 10.1016/j.bpc.2007.01.005

Singer, W. (2001). Consciousness and the binding problem. *Ann. NY Acad. Sci.* 929, 123–146. doi: 10.1111/j.1749-6632.2001.tb05712.x

Singer, W., and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586. doi: 10.1146/annurev.ne.18.030195.003011

Stoljar, D. (2015). "Physicalism," in *The Stanford Encyclopedia of Philosophy,* ed E. N. Zalta. Available online at: http://plato.stanford.edu/archives/spr2015/entries/physicalism/.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi: 10.2307/25470707

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Phil. Trans. R. Soc. B* 370:20140167. doi: 10.1098/rstb.2014.0167

Touboul, J., and Destexhe, A. (2010). Can power-law scaling and neuronal avalanches arise from stochastic dynamics? *PLoS ONE* 5:e8982. doi: 10.1371/journal.pone.0008982

Tuisku, P., Pernu, T. K., and Annila, A. (2009). In the light of time. *Proc. R. Soc. A.* 465, 1173–1198. doi: 10.1098/rspa.2008.0494

van den Heuvel, M. P., Stam, C. J., Boersma, M., and Hulshoff Pol, H. E. (2008). Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *Neuroimage* 43, 528–539. doi: 10.1016/j.neuroimage.2008.08.010

Varela, F. (1999). "The specious present: a neurophenomenology of time consciousness," in *Naturalizing Phenomenology,* eds J. Petitot, F. J. Varela, B. Pacoud, and J.-M. Roy (Stanford, CA: Stanford University Press), 266–314.

Varpula, S., Annila, A., and Beck, C. (2013). Thoughts about thinking. *Adv. Stud. Biol.* 5, 135–149.

Werner, G. (2013). Consciousness viewed in the framework of brain phase space dynamics, criticality, and the Renormalization Group. *Chaos Solitons Fractals* 55, 3–12. doi: 10.1016/j.chaos.2012.03.014

Witten, I. B., and Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron* 48, 489–496. doi: 10.1016/j.neuron.2005.10.020

Yufik, Y. M. (1998). "Virtual associative networks: a framework for cognitive modeling," in *Brain and Values,* ed K. H. Pribram (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.,), 109–177.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA:Addison-Wesley Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

REVIEW
published: 15 November 2016
doi: 10.3389/fnsys.2016.00090

Neurobiology as Information Physics

*Sterling Street**

Department of Cellular Biology, Franklin College of Arts and Sciences, University of Georgia, Athens, GA, USA

This article reviews thermodynamic relationships in the brain in an attempt to consolidate current research in systems neuroscience. The present synthesis supports proposals that thermodynamic information in the brain can be quantified to an appreciable degree of objectivity, that many qualitative properties of information in systems of the brain can be inferred by observing changes in thermodynamic quantities, and that many features of the brain's anatomy and architecture illustrate relatively simple information-energy relationships. The brain may provide a unique window into the relationship between energy and information.

Keywords: information thermodynamics, Landauer limit, free energy principle, optimization, Bekenstein bound

INTRODUCTION

That information is physical has been suggested by evidence since the founding of classical thermodynamics (Lloyd, 2006; Gleick, 2011). In recent years, Landauer's principle (Landauer, 1996; Bennett, 2003), which relates information-theoretic entropy to thermodynamic information, has been confirmed (Parrondo et al., 2015), and the experimental demonstration of a form of information-energy equivalence (Alfonso-Faus, 2013) has verified that Maxwell's demon cannot violate any known laws of thermodynamics (Maruyama et al., 2009). The theoretical finding that entropy is conserved as event horizon area is leading to the resolution of the black hole information paradox (Davies, 2010; Moskowitz, 2015), and there is a fundamental relationship between information and the geometry of spacetime itself (Bousso, 2002; Eling et al., 2006). Current formulations of quantum theory are revealing properties of physical information (Wheeler, 1986; Brukner and Zeilinger, 2003; Lloyd, 2006; Vedral, 2010), and information-interpretive attempts to show that gravity is quantized (Smolin, 2001; Lee et al., 2013) could even lead to the unification of quantum mechanics and the theories of relativity. Although similar approaches are increasingly influential in biology (Schneider and Sagan, 2005; England, 2013; Flack, 2014), "a formalization of the relationship between information and energy is currently lacking in neuroscience" (Collell and Fauquet, 2015). The purpose of this article is to explore a few different sides of this relationship and, along the way, to suggest that many hypotheses and theories in neuroscience can be unified by the physics of information.

INFORMATION BOUNDS

> "How can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" – (Schrödinger, 1944, from Friston, 2013).

As a fundamental physical entity (Lloyd, 2015), information is not fully understood, and there is currently a significant amount of disagreement over different definitions of information and entropy in the literature (Poirier, 2014; Ben-Naim, 2015). In thermodynamics, however,

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research, Inc., USA

Reviewed by:
Guillem Collell,
Katholieke Universiteit Leuven, Belgium
Arto Annila,
University of Helsinki, Finland

***Correspondence:**
Sterling Street
sterling.street@uga.edu

Received: *10 August 2016*
Accepted: *28 October 2016*
Published: *15 November 2016*

Citation:
Street S (2016) Neurobiology as Information Physics. Front. Syst. Neurosci. 10:90. doi: 10.3389/fnsys.2016.00090

information can be defined as a negation of thermodynamic entropy (Beck, 2009):

$$I \equiv -S$$

A bit of thermodynamic entropy represents the distinction between two alternative states in a physical system (Stone, 2015). As a result, the total thermodynamic entropy of a system is proportional to the total number of distinguishable states contained in the system (Bekenstein, 2001, 2007). Because thermodynamic entropy is potential information relative to an observer (Lloyd, 2006), and an observer in a physical system is a component of the system itself, the total thermodynamic entropy of a system includes the portion of entropy that is accessible to the observer as relative thermodynamic information (Wheeler, 1989; Collell and Fauquet, 2015):

$$I_{\text{relative}} = S_{\text{total}} - S_{\text{relative}}$$

Since entropy in any physical system is finite (Lloyd, 2006; Rovelli, 2015), the total thermodynamic entropy of any system of the brain can be quantified by applying the traditional form of the universal (Bekenstein, 1981, 1984, 2001, 2004, 2007) information-entropy bound:

$$S_{\text{sys}} = \zeta \frac{AEk}{\hbar c}$$

where A is area, E is energy including matter, $\hbar$ is the reduced Planck constant, c is the speed of light, k is Boltzmann's constant, and ζ is a factor such that $0 \leq \zeta \leq 1$.

Setting this factor to 1 in order to quantify the total thermodynamic entropy of a system at a certain level of structure now allows us to quantify thermodynamic information by partitioning the factor into a relative information component ($\zeta_{I=1-\zeta_s}$) and a relative entropy component ($\zeta_{s=1-\zeta_I}$),

$$I_{\text{sys}} = \zeta_I \frac{AEk}{\hbar c} = (1 - \zeta_s) \frac{AEk}{\hbar c}$$

Because a maximal level of energy corresponds to a maximal level of thermodynamic information, and a minimal level of energy corresponds to a minimal level of thermodynamic information (Duncan and Semura, 2004), any transitions between energy levels occur as transitions between informational extrema. So, in the event that information enters a system of the brain,

$$\Delta I_{\text{sys}} = \frac{\Delta E_{\text{sys}}}{kT} = \Delta \zeta_I$$

where T is temperature. And, in the case that information exits a system,

$$-\Delta I_{\text{sys}} = \frac{\Delta E_{\text{surr}}}{kT} = \Delta \zeta_s$$

Various forms of these relationships, including information-entropy bounds, have been applied in neuroscience (Friston, 2010; Sengupta et al., 2013a,c, 2016; Collell and Fauquet, 2015; Sterling and Laughlin, 2015). The contribution of this review is simply to show that these relationships can be united into a common theoretical framework.

NEUROBIOLOGY

> *"... classical thermodynamics... is the only physical theory of universal content which I am convinced, that within the framework of applicability of its basic concepts, will never be overthrown."* – (Einstein, 1949, from Bekenstein, 2001).

This section reviews thermodynamic relationships in systems neuroscience with a focus on information and energy. Beginning with neurons, moving to neural networks, and concluding at the level of the brain as a whole, I discuss the energetics of processes such as learning and memory, excitation and inhibition, and the production of noise in neurobiological systems.

The central role of energy in determining the activity of neurons exposes the close connection between information and thermodynamics at the level of the cell. For instance, the process of depolarization, which occurs as a transition to E_{max} from a resting state E_{min}, clearly shows that cellular information content is correlated with energy levels. In this respect, the resemblance between ion concentration gradients in neurons and temperature gradients in thermodynamic demons (i.e., agents that use information from their surroundings to decrease their thermodynamic entropy) is not a coincidence – in order to acquire information, neurons must expend energy to establish proper membrane potentials. Recall that Landauer's principle (Plenio and Vitelli, 2001; Parrondo et al., 2015) places a lower bound on the quantity of energy released into the surroundings with the removal of information from a system. Thus, reestablishing membrane potentials after depolarization – the neuronal equivalent of resetting a demon's memory – dissipates energy. Because Landauer's principle applies to all levels of structure, and cells process large quantities of information, neurons use energy efficiently despite operating at several orders of magnitude above the nominal limit. Parameters including membrane area, spiking frequency, and axon length have all been optimized over the course of evolution to allow neurons to process information efficiently (Sterling and Laughlin, 2015). Examining the energetics of information processing in neurons reinforces the notion that, while it is often convenient to imagine the neuron to be a simple binary element, these cells are intricate computational structures that process more than one bit of information.

Relationships between information and energy can also be seen at the level of neural networks. Attractor networks naturally stabilize by seeking energy minima, and the relative positions of basins of attraction define the geometry of an energy landscape (Amit, 1992). As a result, the transition into an active attractor state occurs as a transition into an information-energy maximum. These transitions correspond to the generation of informational entities such as memories, decisions, and perceptual events (Rolls, 2012). In this way, the energy basins of attractor networks may be analogous to lower-level cellular and molecular energy gradients; a transition between any number of distinguishable energy levels follows the passage of a finite quantity of information. Since processing information requires the expenditure of energy, competitive network features also underscore the need to minimize unnecessary information processing. Lateral inhibition

at this level may optimize thermodynamic efficiency by reducing metabolic expenses associated with networks responding less robustly to entering signals. Another interesting thermodynamic property of networks concerns macrostates: the functional states of large-scale neural networks rest emergently on the states of neuronal assemblies (Yuste, 2015). As a result, new computational properties may arise with the addition of new layers of network structure. Finally, the energetic cost of information has influenced network connectivity by imposing selective pressures to save energy by minimizing path length between network nodes (Bullmore and Sporns, 2009).

Again, in accordance with Landauer's principle, the displacement of information from any system releases energy into the surroundings (Plenio and Vitelli, 2001; Duncan and Semura, 2004). This principle can be understood by imagining an idealized memory device, such as the brain of a thermodynamic demon. Since information is conserved (Susskind and Hrabovsky, 2014), and clearing a memory erases information, the thermodynamic entropy of the surroundings must increase when a demon refreshes its memory to update information. This fundamental connection between information, entropy, and energy appears in many areas of the neurobiology of learning. For example, adjusting a firing threshold in order to change the probability that a system will respond to a conditioned stimulus (Takeuchi et al., 2014; Choe, 2015) optimizes engram fitness by minimizing the quantity of energy needed for its activation (Still et al., 2012). Recurrent collateral connections further increase engram efficiency by enabling a minimal nodal stimulus to elicit its full energetic activation (Rolls, 2012). Experimental evidence also shows that restricting synaptic energy supply impairs the formation of stable engrams (Harris et al., 2012). Because the formation and disassembly of engrams during learning and forgetting optimizes the growth and pruning of networks in response to external conditions, the process of learning is itself a mechanism for minimizing entropy in the brain (Friston, 2003).

As another example of a multiscale process integrated across many levels by thermodynamics, consider the active balance between excitation and inhibition in neurobiological systems. Maintaining proper membrane potentials and adequate concentrations of signaling molecules requires the expenditure of energy, so it is advantageous for systems of the brain to minimize the processing of unnecessary information – to "send only what is needed" (Sterling and Laughlin, 2015). Balancing excitation and inhibition is therefore a crucial mechanism for saving energy. Theoretical evidence that this balancing maximizes the thermodynamic efficiency of processing Shannon information (Sengupta et al., 2013b) is consistent with experimental findings in several areas of research on inhibition. For instance, constant inhibitory modulation is needed to stabilize internal states, and hyperexcitation (e.g., in epilepsy, intoxication syndromes, or trauma) can decrease relative information by reducing levels of consciousness (Haider et al., 2006; Lehmann et al., 2012). Likewise, selective attention is mediated by the activation of inhibitory interneurons (Houghton and Tipper, 1996), and sensory inhibition appears to sharpen internal perceptual states (Isaacson and Scanziani, 2011). The need to balance excitation and inhibition at all levels of structure highlights the energetic cost of information.

A final example worth discussing is the relationship between thermodynamics and the production of noise in neurobiological systems. Noise is present in every system of the brain, and influences all aspects of the organ's function (Faisal et al., 2008; Rolls and Deco, 2010; Destexhe and Rudolph-Lilith, 2012). Even in the absence of any potential forms of classical stochastic resonance, the noise-driven exploration of different states may optimize thermodynamic efficiency by allowing a system to randomly sample different accessible configurations. Theoretical arguments suggest indeed that noise enables neural networks to respond more quickly to detected signals (Rolls, 2012), and empirical evidence implicates noise as a beneficial means of optimizing the performance of diverse neurobiological processes (McDonnell and Ward, 2011). For example, noise in the form of neuronal DNA breaking (Guo et al., 2011; Herrup et al., 2013; Tognini et al., 2015) could enhance plasticity, since any stochastically optimized configuration would be more likely to survive over time as, in this case, a strengthened connection in a modifiable network. Because noise is a form of relative entropy, optimizing the signal-to-noise ratio in any neurobiological system promotes the efficient use of energy.

At the level of the brain as a whole, the connection between information and thermodynamics is readily apparent in the organ's functional reliance on energy (Magistretti and Allaman, 2015), its seemingly disproportionate consumption of oxygen and energy substrates (e.g., ATP, glucose, ketones, etc.; Raichle and Gusnard, 2002; Herculano-Houzel, 2011), its vulnerability to hypoxic-ischemic damage (Lutz et al., 2003; Dreier et al., 2013) and in the reduction of consciousness often conferred by the onset of energy restrictions (Shulman et al., 2009; Stender et al., 2016). All fMRI, PET, and EEG interpretation rests on the foundational assumption that changes in the information content of neurobiological systems can be inferred by observing energy changes (Attwell and Iadecola, 2002; Collell and Fauquet, 2015), and it is well known that the information processing capacities of neurobiological systems are limited by energy supply (Howarth et al., 2012; Fox, 2015). Overall, these relationships are consistent with the form of information-energy equivalence predicted by Landauer's principle and information-entropy bounds. The living brain appears to maintain a state of thermodynamic optimization.

CONSCIOUSNESS AND FREE WILL

> *"... science appears completely to lose from sight the large and general questions; but all the more splendid is the success when, groping in the thicket of special questions, we suddenly find a small opening that allows a hitherto undreamt of outlook on the whole."* – (Boltzmann, 1892, from Von Baeyer, 1999).

Although neuroscience has yet to explain consciousness or free will at any satisfactory level of detail, relationships between information and energy seem to be recognizable even at this level

of analysis. This section reviews attempts to conceptualize major properties of consciousness (unity, continuity, complexity, and self-awareness) as features of information processing in the brain, and concludes with a discussion of free will.

At any given moment, awareness is experienced as a unified whole. Physical information is the substrate of consciousness (Annila, 2016), and the law of conservation of information requires any minimal unit of information to be transferred into a thermodynamic system as a temporally unitary quantity. As a result, it is possible that the passage of perceptual time itself occurs secondarily to the transfer of information, and that the information present in any integrated system of the brain at any observed time is necessarily cohesive and temporally unified. In this framework, the passage of time would vary in proportion to a system's rate of energy dissipation. Although it is possible that physical systems in general exchange information in temporally unitary quantities, it is likely that many of the familiar features of the perceptual unity of consciousness require the structure and activity of neural networks in the brain. The biological basis of this unity may be the active temporal consolidation of observed events by integrated higher-order networks (Revonsuo, 1999; Varela et al., 2001; Greenfield and Collins, 2005; Dehaene and Changeux, 2011). An informational structure generated by the claustrum has been speculated to contribute to this experiential unity (Crick and Koch, 2005; Koubeissi et al., 2014), but it has also been reported that complete unilateral resection of the system performed in patients with neoplastic lesions of the region produces no externally observable changes in subjective awareness (Duffau et al., 2007). Overall, it appears unlikely that the presence of information in any isolated or compartmentalized network of the brain is responsible for generating the unified nature of conscious experience.

While perceptual time is likely the product of a collection of related informational processes rather than a single, globalized function mediated by any one specific system of the brain, some of the perceptual continuity of consciousness may result from the effectively continuous flow of thermodynamic information into and out of integrated systems of the brain. In this framework, the quantum (Prokopenko and Lizier, 2014) of perceptual time would be the minimal acquisition of information, and the entrance of information into neurobiological systems would occur alongside the entrance of energy. This relationship is implicit in the simple observation that the transition of a large-scale attractor network is progressively less discrete and smoother in time than the activation of a small-scale engram, the propagation of a cellular potential, the docking of a vesicle, the release of an ion, and so forth. Likewise, electroencephalography shows that the summation of a large number of discrete cellular potentials can accumulate into an effectively continuous wave as a network field potential (Nunez and Srinivasan, 2006), disruptions of which are often correlated with decreases in levels of consciousness (Blumenfeld and Taylor, 2003). It is also well known that higher frequency network oscillations tend to indicate states of wakefulness and active awareness, while lower frequency oscillations tend to be associated with internal states of lesser passage of perceptual time, such as dreamless sleep or unconsciousness. The possibility that the experiential arrow of time and the thermodynamic arrow of time share a common origin in the flow of information is supported both by general models of time in neuroscience and the physical interpretation of time as an entropy gradient (Stoica, 2008; Mlodinow and Brun, 2014).

The subjective complexity of consciousness may show that extensive network integration is needed for maximizing the mutual thermodynamic information and internal energy content of systems of the brain (Torday and Miller, 2016). An exemplary structure enabling such experience, likely one of many that together account for the subjective complexity of consciousness, is the thalamocortical complex (Calabrò et al., 2015; Hannawi et al., 2015). The functional architecture of such a network may show that, at any given moment in the internal model of a living brain, a wide range of integrated systems are sharing mutual sources of thermodynamic information. This pattern of structure may reveal that the perceptual depth and complexity of conscious experience is a direct product of recognizable features of the physical brain. However, it also seems that extensive local cortical processing of information is necessary for producing a refined and coherent sensorium within a system, and that both the thalamocortical complex and the brain stem are involved in generating the subjective complexity of consciousness (Edelman et al., 2011; Ward, 2011). The dynamics of attractor networks at higher levels of network structure may show that quantities of complex internal information can be observed as changes in cortical energy landscapes (Rolls, 2012), with a transition between attractor states following the transfer of information. The degree of subjective complexity of information enclosed by such a transition would be proportional to the degree of structural integration of underlying networks.

Self-awareness likely arose as a survival necessity rather than as an accident of evolution (Fabbro et al., 2015), and rudimentary forms of self-awareness likely began to appear early in the course of brain evolution as various forms of perceptual self-environment separation. As a simple example, consider the tickle response (Linden, 2007), which requires the ability to differentiate self-produced tactile sensations from those produced by external systems. The early need to distinguish between self-produced tactile states and those produced by more threatening non-self sources may be reflected by the observation that this recognition process is mediated to a great extent by the cerebellum (Blakemore et al., 2000). While it is possible that other similar developments began occurring very early on, the evolutionary acquisition of the refined syntactical and conceptual self present in the modern brain likely required the merging of pre-existing self networks with higher-level cortical systems. The eventual integration of language and self-awareness would have been advantageous for coordinating social groups (Graziano, 2013), since experiencing self-referential thought as inner speech facilitates verbal communication. Likewise, the coupling of self-awareness to internal sensory, cognitive, and

motor states (Metzinger, 2004; Northoff et al., 2006) may be advantageous for maximizing information between systems within an individual brain. Neuropsychological conditions involving different forms of agnosia, neglect, and self-awareness deficits do show that a reduced awareness of self-ownership of motor skills, body parts, or perceptual states can result in significant disability (Parton et al., 2004; Morin, 2006; Orfei et al., 2007; Prigatano, 2009; Tsakiris, 2010; Overgaard, 2011; Fabbro et al., 2015; Chokron et al., 2016). Since experiencing self-awareness optimizes levels of mutual information between the external world and the brain's internal model (Apps and Tsakiris, 2014), and this activity decreases thermodynamic entropy (Torday and Miller, 2016), self-awareness may be a mechanism for optimizing the brain's consumption of energy.

Thermodynamic information is also interesting to consider in the context of free will. The brain is predictable within reason, and the performance of an action can be predicted before a decision is reported to have been made (Haggard, 2008). Entities such as ideas, feelings, and beliefs seem to exist as effectively deterministic evaluations of information processed in the brain. Whether or not the flow of information is subject to the brain's volitional alteration, neuroscience also shows that information can be internally real to a system of the brain, even if this information is inconsistent with an external reality. That the brain can generate an externally inconsistent internal reality is demonstrated by phenomena such as confabulation, agnosia, blindsight, neglect, commissurotomy and hemispherectomy effects, placebo and nocebo effects, reality monitoring deficits, hallucinations, prediction errors, the suspension of disbelief during dreaming, the function of communication in minimizing divergence between internal realities, the quality of many kinds of realistic drug-induced experiences, and the effects of many neuropsychological conditions. The apparent fact that subjective reality is an active construction of the physical brain has even led to the proposal of model-dependent realism (Hawking and Mlodinow, 2011) as a philosophical paradigm in the search for a unified theory of physics. In any case, it is likely that beliefs, including those in free will, exist as information, and that their internal reality is a restatement of its frequently observer-dependent nature.

EMPIRICAL OUTLOOK

Before concluding, it is worth reviewing a few notable experiments in greater detail. While considerable advances have been made in discovering how neurobiological systems operate according to principles of thermodynamic efficiency (Sterling and Laughlin, 2015), relationships between information and energy in the brain are only beginning to be understood. The following studies are examples of elegant and insightful experiments that should inspire future research.

Several recent brain imaging studies support the proposal (Annila, 2016) that thermodynamics is able to explain a number of mysteries involving consciousness. For example, Stender et al. (2016) used PET to measure global resting state energy consumption in 131 brain injury patients with impairments of consciousness as defined by the revised Coma Recovery Scale (CRS-R). The preservation of consciousness was found to require a minimal global metabolic rate of $\approx$ 40% of the average rate of controls; global energy consumption above this level was reported to predict the presence or recovery of consciousness with over 90% sensitivity. These results must be replicated and studied in closer detail before their specific theoretical implications are clear, but it is now established that levels of consciousness are correlated with energetic metrics of brain activity. To what extent there exists a well-defined "minimal energetic requirement for the presence of conscious awareness" (Stender et al., 2016) remains an open question. However, the empirical confirmation of a connection between consciousness and thermodynamics introduces the possibility of developing new experimental methods in consciousness research.

Neurobiological systems, and biological systems in general (Von Baeyer, 1999; Schneider and Sagan, 2005), can be considered thermodynamic demons in the sense that they are agents using information to decrease their thermodynamic entropy. Landauer's principle requires that, in order not to violate any known laws of thermodynamics, such agents dissipate heat when erasing information from their memory storage devices. In an experimental test of this principle, reviewed along with similar experiments in Parrondo et al. (2015) and Bérut et al. (2012) studied heat dissipation in a simple memory device created by placing a glass bead in an optical double-well potential. Intuitively, this memory stored a bit of information by retaining the bead on one side of the potential rather than on the alternative. By manipulating the height of the optical barrier between wells, researchers moved the bead to one side of the memory without determining its previous location in the potential. This process was therefore logically irreversible, requiring the erasure of prior information from the memory device. Landauer's principle predicts that, since information is conserved, the entropy of the memory's surroundings must increase when this occurs. Bérut et al. (2012) have verified that energy is emitted when a memory is cleared. As noted by the authors, "this limit is independent of the actual device, circuit or material used to implement the irreversible operation." It would be interesting to study the erasure principle in the context of neuroscience.

Experimental applications of information theory in cell biology have already led to the discovery of general principles of brain organization related to thermodynamics (Sterling and Laughlin, 2015). In one particularly interesting study, Niven et al. (2007) measured the energetic efficiency of information coding in retinal neurons. Intracellular recordings of membrane potential and input resistance were used to calculate rates of ATP consumption in response to different background light intensities. These rates of energy consumption were then compared with rates of Shannon information transmission in order to determine metabolic performance. It was found that metabolic demands increase non-linearly with respect to increases in information processing rate: thermodynamics

appears to impose a "law of diminishing returns" on systems of the brain. The authors interpret these results as evidence that nature has selected for neurons that minimize unnecessary information processing. Studying how thermodynamics has influenced cellular parameters over the course of evolution is likely to raise many new empirically addressable questions.

CONCLUSION

This article has reviewed information-energy relationships in the hope that they may eventually provide a general framework for uniting theory and experiment in neuroscience. The physical nature of information and its status as a finite, measurable resource are emphasized to connect neurobiology and thermodynamics. As a scientific paradigm, the information movement currently underway in physics promises profound advances in our understanding of the relationship between energy, information, and the physical brain.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

I am grateful to Baroness Susan Greenfield, Dr. Francesco Fermani, Dr. Karl Friston, Dr. Biswa Sengupta, Dr. Roy Frieden, Dr. Bernard Baars, Dr. Brett Clementz, Dr. Cristi Stoica, Dr. Satoru Suzuki, Dr. Paul King, Guillem Collell, Dr. Jordi Fauquet, and others who have helped me improve these ideas. I am also grateful to Dr. Shanta Dhar and her team for introducing me to academic research, to Jim Reid for introducing me to biology, to Alex Tisch for introducing me to physics, and to those affiliated with the Department of Neurosurgery at the University of Virginia Medical Center for introducing me to neuroscience.

REFERENCES

Alfonso-Faus, A. (2013). "Fundamental principle of information-to-energy conversion," in *Proceedings of the 7th European Computing Conference*, Dubrovnik.

Amit, D. J. (1992). *Modeling Brain Function*. Cambridge: Cambridge University Press.

Annila, A. (2016). On the character of consciousness. *Front. Syst. Neurosci.* 10:27. doi: 10.3389/fnsys.2016.00027

Apps, M. A., and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* 41, 85–97. doi: 10.1016/j.neubiorev.2013.01.029

Attwell, D., and Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends Neurosci.* 25, 621–625. doi: 10.1016/S0166-2236(02)02264-6

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemp. Phys.* 50, 495–510. doi: 10.1080/00107510902823517

Bekenstein, J. D. (1981). Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D* 23, 287. doi: 10.1103/PhysRevD.23.287

Bekenstein, J. D. (1984). Entropy content and information flow in systems with limited energy. *Phys. Rev. D* 30, 1669. doi: 10.1103/PhysRevD.30.1669

Bekenstein, J. D. (2001). The limits of information. *Stud. Hist. Philos. Mod. Phys.* 32, 511–524. doi: 10.1016/S1355-2198(01)00020-X

Bekenstein, J. D. (2004). Black holes and information theory. *Contemp. Phys.* 45, 31–43. doi: 10.1080/00107510310001632523

Bekenstein, J. D. (2007). Information in the holographic universe. *Sci. Am.* 17, 66–73. doi: 10.1038/scientificamerican0407-66sp

Ben-Naim, A. (2015). *Information, Entropy, Life and the Universe*. Singapore: World Scientific, 4–5.

Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Stud. Hist. Philos. Sci. B* 34, 501–510. doi: 10.1016/S1355-2198(03)00039-X

Bérut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., and Lutz, E. (2012). Experimental verification of Landauer/'s principle linking information and thermodynamics. *Nature* 483, 187–189. doi: 10.1038/nature10872

Blakemore, S. J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport* 11, R11–R16. doi: 10.1097/00001756-200008030-00002

Blumenfeld, H., and Taylor, J. (2003). Why do seizures cause loss of consciousness? *Neuroscientist* 9, 301–310. doi: 10.1177/1073858403255624

Boltzmann, L. (1892). "On the methods of theoretical physics," in *Theoretical Physics and Philosophical Problems*, ed. B. McGuinness (Dordrecht: Springer Netherlands), 5–12.

Bousso, R. (2002). The holographic principle. *Rev. Mod. Phys.* 74, 825. doi: 10.1103/RevModPhys.74.825

Brukner, Č, and Zeilinger, A. (2003). "Information and fundamental elements of the structure of quantum theory," in *Time, Quantum and Information*, eds L. Castell and O. Ischebeck (Heidelberg: Springer), 323–354.

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575

Calabrò, R. S., Cacciola, A., Bramanti, P., and Milardi, D. (2015). Neural correlates of consciousness: what we know and what we have to learn! *Neurol. Sci.* 36, 505–513. doi: 10.1007/s10072-015-2072-x

Choe, Y. (2015). "Hebbian learning," in *Encyclopedia of Computational Neuroscience*, eds D. Jaeger and R. Jung (New York, NY: Springer), 1305–1309. doi: 10.1007/978-1-4614-7320-6_672-1

Chokron, S., Perez, C., and Peyrin, C. (2016). Behavioral consequences and cortical reorganization in homonymous hemianopia. *Front. Syst. Neurosci.* 10:57. doi: 10.3389/fnsys.2016.00057

Collell, G., and Fauquet, J. (2015). Brain activity and cognition: a connection from thermodynamics and information theory. *Front. Psychol.* 6:818. doi: 10.3389/fpsyg.2015.00818

Crick, F. C., and Koch, C. (2005). What is the function of the claustrum? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1271–1279. doi: 10.1098/rstb.2005.1661

Davies, P. (2010). "Universe from bit," in *Information and the Nature of Reality*, eds P. Davies and N. H. Gregersen (Cambridge: Cambridge University Press), 83–117.

Dehaene, S., and Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Destexhe, A., and Rudolph-Lilith, M. (2012). *Neuronal Noise*. New York, NY: Springer.

Dreier, J. P., Isele, T., Reiffurth, C., Offenhauser, N., Kirov, S. A., Dahlem, M. A., et al. (2013). Is spreading depolarization characterized by an abrupt, massive release of Gibbs free energy from the human brain cortex? *Neuroscientist* 19, 25–42. doi: 10.1177/1073858412453340

Duffau, H., Mandonnet, E., Gatignol, P., and Capelle, L. (2007). Functional compensation of the claustrum: lessons from low-grade glioma surgery. *J. Neurooncol.* 81, 327–329. doi: 10.1007/s11060-006-9236-8

Duncan, T. L., and Semura, J. S. (2004). The deep physics behind the second law: information and energy as independent forms of bookkeeping. *Entropy* 6, 21–29. doi: 10.3390/e6010021

Edelman, G. M., Gally, J. A., and Baars, B. J. (2011). Biology of consciousness. *Front. Psychol.* 2:4. doi: 10.3389/fpsyg.2011.00004

Einstein, A. (1949). "Autobiographical notes," in *Albert Einstein*, ed. P. A. Schilpp (La Salle: Open Court), 33.

Eling, C., Guedens, R., and Jacobson, T. (2006). Nonequilibrium thermodynamics of spacetime. *Phys. Rev. Lett.* 96, 121301. doi: 10.1103/PhysRevLett.96.121301

England, J. L. (2013). Statistical physics of self-replication. *J. Chem. Phys.* 139, 121923. doi: 10.1063/1.4818538

Fabbro, F., Aglioti, S. M., Bergamasco, M., Clarici, A., and Panksepp, J. (2015). Evolutionary aspects of self-and world consciousness in vertebrates. *Front. Hum. Neurosci.* 9:157. doi: 10.3389/fnhum.2015.00157

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303. doi: 10.1038/nrn2258

Flack, J. C. (2014). Life's information hierarchy. *Santa Fe Inst. Bull.* 28, 13–24.

Fox, D. (2015). The limits of intelligence. *Sci. Am.* 305, 36–43.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Netw.* 16, 1325–1352. doi: 10.1016/j.neunet.2003.06.005

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Gleick, J. (2011). *The Information*. New York, NY: Random House, 269–270.

Graziano, M. S. (2013). *Consciousness and the Social Brain*. Oxford: Oxford University Press.

Greenfield, S. A., and Collins, T. F. T. (2005). A neuroscientific approach to consciousness. *Prog. Brain Res.* 150, 11–23. doi: 10.1016/S0079-6123(05)50002-5

Guo, J. U., Ma, D. K., Mo, H., Ball, M. P., Jang, M. H., Bonaguidi, M. A., et al. (2011). Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nat. Neurosci.* 14, 1345–1351. doi: 10.1038/nn.2900

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* 9, 934–946. doi: 10.1038/nrn2497

Haider, B., Duque, A., Hasenstaub, A. R., and McCormick, D. A. (2006). Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* 26, 4535–4545. doi: 10.1523/JNEUROSCI.5297-05.2006

Hannawi, Y., Lindquist, M. A., Caffo, B. S., Sair, H. I., and Stevens, R. D. (2015). Resting brain activity in disorders of consciousness. *Neurology* 84, 1272–1280. doi: 10.1212/WNL.0000000000001404

Harris, J. J., Jolivet, R., and Attwell, D. (2012). Synaptic energy use and supply. *Neuron* 75, 762–777. doi: 10.1016/j.neuron.2012.08.019

Hawking, S. W., and Mlodinow, L. (2011). *The Grand Design*. New York, NY: Random House. 7, 46.

Herculano-Houzel, S. (2011). Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution. *PLoS ONE* 6:e17514. doi: 10.1371/journal.pone.0017514

Herrup, K., Chen, J., and Li, J. (2013). Breaking news: thinking may be bad for DNA. *Nat. Neurosci.* 16, 518–519. doi: 10.1038/nn.3384

Houghton, G., and Tipper, S. P. (1996). Inhibitory mechanisms of neural and cognitive control: applications to selective attention and sequential action. *Brain Cogn.* 30, 20–43. doi: 10.1006/brcg.1996.0003

Howarth, C., Gleeson, P., and Attwell, D. (2012). Updated energy budgets for neural computation in the neocortex and cerebellum. *J. Cereb. Blood Flow Metab.* 32, 1222–1232. doi: 10.1038/jcbfm.2012.35

Isaacson, J. S., and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron* 72, 231–243. doi: 10.1016/j.neuron.2011.09.027

Koubeissi, M. Z., Bartolomei, F., Beltagy, A., and Picard, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy Behav.* 37, 32–35. doi: 10.1016/j.yebeh.2014.05.027

Landauer, R. (1996). The physical nature of information. *Phys. Lett. A* 217, 188–193. doi: 10.1016/0375-9601(96)00453-7

Lee, J. W., Kim, H. C., and Lee, J. (2013). Gravity from quantum information. *J. Korean Phys. Soc.* 63, 1094–1098. doi: 10.3938/jkps.63.1094

Lehmann, K., Steinecke, A., and Bolz, J. (2012). GABA through the ages: regulation of cortical function and plasticity by inhibitory interneurons. *Neural Plast* 2012:892784. doi: 10.1155/2012/892784

Linden, D. J. (2007). *The Accidental Mind*. Cambridge: Harvard University Press, 9–12.

Lloyd, S. (2006). *Programming the Universe*. New York, NY: Random House.

Lloyd, S. (2015). *Interview in Closer to Truth: Is Information Fundamental?*. Available at: https://www.closertotruth.com/series/information-fundamental#video-2621

Lutz, P. L., Nilsson, G. E., and Prentice, H. M. (2003). *The Brain Without Oxygen*. New York, NY: Kluwer.

Magistretti, P. J., and Allaman, I. (2015). A cellular perspective on brain energy metabolism and functional imaging. *Neuron* 86, 883–901. doi: 10.1016/j.neuron.2015.03.035

Maruyama, K., Nori, F., and Vedral, V. (2009). Colloquium: the physics of Maxwell's demon and information. *Rev. Mod. Phys.* 81, 1. doi: 10.1103/RevModPhys.81.1

McDonnell, M. D., and Ward, L. M. (2011). The benefits of noise in neural systems: bridging theory and experiment. *Nat. Rev. Neurosci.* 12, 415–426. doi: 10.1038/nrn3061

Metzinger, T. (2004). *Being No One*. Cambridge: MIT Press.

Mlodinow, L., and Brun, T. A. (2014). Relation between the psychological and thermodynamic arrows of time. *Phys. Rev. E* 89:052102. doi: 10.1103/PhysRevE.89.052102

Morin, A. (2006). Levels of consciousness and self-awareness: a comparison and integration of various neurocognitive views. *Conscious. Cogn.* 15, 358–371. doi: 10.1016/j.concog.2005.09.006

Moskowitz, C. (2015). Stephen hawking Hasn't solved the black hole paradox just yet. *Sci. Am.* 27.

Niven, J. E., Anderson, J. C., and Laughlin, S. B. (2007). Fly photoreceptors demonstrate energy-information trade-offs in neural coding. *PLoS Biol.* 5:e116. doi: 10.1371/journal.pbio.0050116

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., and Panksepp, J. (2006). Self-referential processing in our brain–a meta-analysis of imaging studies on the self. *Neuroimage* 31, 440–457. doi: 10.1016/j.neuroimage.2005.12.002

Nunez, P. L., and Srinivasan, R. (2006). *Electric Fields of the Brain*. New York, NY: Oxford University Press.

Orfei, M. D., Robinson, R. G., Prigatano, G. P., Starkstein, S., Rüsch, N., Bria, P., et al. (2007). Anosognosia for hemiplegia after stroke is a multifaceted phenomenon: a systematic review of the literature. *Brain* 130, 3075–3090. doi: 10.1093/brain/awm106

Overgaard, M. (2011). Visual experience and blindsight: a methodological review. *Exp. Brain Res.* 209, 473–479. doi: 10.1007/s00221-011-2578-2

Parrondo, J. M. R., Horowitz, J. M., and Sagawa, T. (2015). Thermodynamics of information. *Nat. Phys.* 11, 131–139. doi: 10.1038/nphys3230

Parton, A., Malhotra, P., and Husain, M. (2004). Hemispatial neglect. *J. Neurol. Neurosurg. Psychiatry* 75, 13–21.

Plenio, M. B., and Vitelli, V. (2001). The physics of forgetting: Landauer's erasure principle and information theory. *Contemp. Phys.* 42, 25–60. doi: 10.1080/00107510010018916

Poirier, B. (2014). *A Conceptual Guide to Thermodynamics*. Chichester: Wiley, 77–78.

Prigatano, G. P. (2009). Anosognosia: clinical and ethical considerations. *Curr. Opin. Neurol.* 22, 606–611. doi: 10.1097/WCO.0b013e328332a1e7

Prokopenko, M., and Lizier, J. T. (2014). Transfer entropy and transient limits of computation. *Sci. Rep.* 4:5394. doi: 10.1038/srep05394

Raichle, M. E., and Gusnard, D. A. (2002). Appraising the brain's energy budget. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10237–10239. doi: 10.1073/pnas.172399499

Revonsuo, A. (1999). Binding and the phenomenal unity of consciousness. *Consci. Cogn.* 8, 173–185. doi: 10.1006/ccog.1999.0384

Rolls, E. T. (2012). *Neuroculture*. Oxford: Oxford University Press.

Rolls, E. T., and Deco, G. (2010). *The Noisy Brain*. Oxford: Oxford University Press.

Rovelli, C. (2015). "Relative information at the foundation of physics," in *It From Bit or Bit From It?*, eds A. Aguirre, B. Foster, and Z. Merali (Cham: Springer), doi: 10.1007/978-3-319-12946-4_7

Schneider, E. D., and Sagan, D. (2005). *Into the Cool*. Chicago: University of Chicago Press.

Schrödinger, E. (1944). *What is Life*? Cambridge: Cambridge University Press, 3.

Sengupta, B., Faisal, A. A., Laughlin, S. B., and Niven, J. E. (2013a). The effect of cell size and channel density on neuronal information encoding and energy efficiency. *J. Cereb. Blood Flow Metab.* 33, 1465–1473. doi: 10.1038/jcbfm.2013.103

Sengupta, B., Laughlin, S. B., and Niven, J. E. (2013b). Balanced excitatory and inhibitory synaptic currents promote efficient coding and metabolic efficiency. *PLoS Comput. Biol.* 9:e1003263. doi: 10.1371/journal.pcbi.1003263

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013c). Information and efficiency in the nervous system–a synthesis. *PLoS Comput. Biol.* 9:e1003157. doi: 10.1371/journal.pcbi.1003157

Sengupta, B., Tozzi, A., Cooray, G. K., Douglas, P. K., and Friston, K. J. (2016). Towards a neuronal gauge theory. *PLoS Biol.* 14:e1002400. doi: 10.1371/journal.pbio.1002400

Shulman, R. G., Hyder, F., and Rothman, D. L. (2009). Baseline brain energy supports the state of consciousness. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11096–11101. doi: 10.1073/pnas.0903941106

Smolin, L. (2001). *Three Roads to Quantum Gravity*. New York, NY: Basic Books. 103, 169–178.

Stender, J., Mortensen, K. N., Thibaut, A., Darkner, S., Laureys, S., Gjedde, A., et al. (2016). The minimal energetic requirement of sustained awareness after brain injury. *Curr. Biol.* 26, 1494–1499. doi: 10.1016/j.cub.2016.04.024

Sterling, P., and Laughlin, S. (2015). *Principles of Neural Design*. Cambridge: MIT Press.

Still, S., Sivak, D. A., Bell, A. J., and Crooks, G. E. (2012). Thermodynamics of prediction. *Phys. Rev. Lett.* 109:120604. doi: 10.1103/PhysRevLett.109.120604

Stoica, O. C. (2008). *Flowing with a Frozen River. FQXi, The Nature of Time essay contest*. Available at: http://fqxi.org/community/essay/winners/2008.1#Stoica

Stone, J. V. (2015). *Information Theory*. Sheffield: Sebtel Press, 171.

Susskind, L., and Hrabovsky, G. (2014). *The Theoretical Minimum*, Vol. 9. New York, NY: Basic Books, 170.

Takeuchi, T., Duszkiewicz, A. J., and Morris, R. G. (2014). The synaptic plasticity and memory hypothesis: encoding, storage and persistence. *Philos. Trans. R. Soc. B* 369:20130288. doi: 10.1098/rstb.2013.0288

Tognini, P., Napoli, D., and Pizzorusso, T. (2015). Dynamic DNA methylation in the brain: a new epigenetic mark for experience-dependent plasticity. *Front. Cell. Neurosci.* 9:331. doi: 10.3389/fncel.2015.00331

Torday, J. S., and Miller, W. B. Jr. (2016). On the evolution of the mammalian brain. *Front. Syst. Neurosci.* 10:31. doi: 10.3389/fnsys.2016.00031

Tsakiris, M. (2010). My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia* 48, 703–712. doi: 10.1016/j.neuropsychologia.2009.09.034

Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

Vedral, V. (2010). *Decoding Reality*. Oxford: Oxford University Press.

Von Baeyer, H. C. (1999). *Maxwell's Demon*. New York, NY: Random House, 100–101.

Ward, L. M. (2011). The thalamic dynamic core theory of conscious experience. *Conscious. Cogn.* 20, 464–486. doi: 10.1016/j.concog.2011.01.007

Wheeler, J. (1986). "John Wheeler," in *The Ghost in the Atom* eds P. Davies and J. Brown (Cambridge: Cambridge University Press), 62.

Wheeler, J. A. (1989). "Information, physics, quantum: the search for links," in *Proceedings of the Third International Symposium on Foundations of Quantum Mechanics*, Tokyo, 354–368.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* 16, 487–497. doi: 10.1038/nrn3962

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 19 April 2016
doi: 10.3389/fnsys.2016.00031

On the Evolution of the Mammalian Brain

John S. Torday[1]* *and William B. Miller Jr.*[2]

[1] *Evolutionary Medicine Program, University of California- Los Angeles, Los Angeles, CA, USA,* [2] *Independent Researcher, Paradise Valley, AZ, USA*

Hobson and Friston have hypothesized that the brain must actively dissipate heat in order to process information (Hobson et al., 2014). This physiologic trait is functionally homologous with the first instantation of life formed by lipids suspended in water forming micelles- allowing the reduction in entropy (heat dissipation). This circumvents the Second Law of Thermodynamics permitting the transfer of information between living entities, enabling them to perpetually glean information from the environment, that is felt by many to correspond to evolution *per se*. The next evolutionary milestone was the advent of cholesterol, embedded in the cell membranes of primordial eukaryotes, facilitating metabolism, oxygenation and locomotion, the triadic basis for vertebrate evolution. Lipids were key to homeostatic regulation of calcium, forming calcium channels. Cell membrane cholesterol also fostered metazoan evolution by forming lipid rafts for receptor-mediated cell-cell signaling, the origin of the endocrine system. The eukaryotic cell membrane exapted to all complex physiologic traits, including the lung and brain, which are molecularly homologous through the function of neuregulin, mediating both lung development and myelinization of neurons. That cooption later exapted as endothermy during the water-land transition (Torday, 2015a), perhaps being the functional homolog for brain heat dissipation and conscious/mindful information processing. The skin and brain similarly share molecular homologies through the "skin-brain" hypothesis, giving insight to the cellular-molecular "arc" of consciousness from its unicellular origins to integrated physiology. This perspective on the evolution of the central nervous system clarifies self-organization, reconciling thermodynamic and informational definitions of the underlying biophysical mechanisms, thereby elucidating relations between the predictive capabilities of the brain and self-organizational processes.

Keywords: evolution, brain, entropy, lipids, endothermy, skin-brain, exaptation, self-organization

OPEN ACCESS

Edited by:
Biswa Sengupta,
University College London, UK

Reviewed by:
Karl Friston,
University College London, UK
Stuart Hameroff,
University of Arizona Health Sciences Center, USA

***Correspondence:**
John S. Torday
jtorday@ucla.edu

Received: *09 December 2015*
Accepted: *22 March 2016*
Published: *19 April 2016*

Citation:
Torday JS and Miller WB Jr. (2016) On the Evolution of the Mammalian Brain. Front. Syst. Neurosci. 10:31. doi: 10.3389/fnsys.2016.00031

INTRODUCTION

The origins of consciousness and the evolution of physiologic pathways in mammalian brain are arguably among the most challenging of all evolutionary puzzles. It is becoming increasingly evident that the proper point of initiation of any understanding of these phenomena channels through a fuller understanding of the capacities of individual and networked cells and even more particularly, a reassessment of the significance of the unicellular zygotic phase of all eukaryotes (Minami et al., 2007; Ikeda et al., 2010). If considered from within that cellular frame, the development of any higher level of consciousness must relate to an enabling continuum

of physiologic evolution that begins from that unicellular form through which all sentient eukaryotes must recapitulate. Furthermore, such a path would necessarily extend beyond any prior assumptions of multicellular physiological development as a simple progression forward from unicellular life. Instead, eukaryotic physiology must be evaluated through the unfamiliar perspective that all the consequential processes that have antecedents from within the unicellular form retain a consistent and inherent anchor within that origin throughout development. Further too, those same initiating factors remain foundational throughout organic development not only at the level of any individual organism, but also throughout any evolutionary narrative.

This perspective is underscored by two significant biological principles. First, life is cognition at every scope and scale (Baluška and Mancuso, 2009; Shapiro, 2011; Miller, 2013; Lyon, 2015). And second, and conditional upon the first, all complex physiologic traits have evolved from the unicellular state as derivative exaptations of the complex cellular cytoskeletal elements and cell membrane as well as the crucial genetic material that is generally accredited as the centrality of that process (Torday, 2013, 2015a). All of these cellular constituents consistently inter-react to create any functioning cell. And since all cells have cognitive capacity, then, physiology is best understood as both a continuous enactment of cellular "self" and the biologic means by which that cellular "self" is maintained and advanced within cellular boundaries throughout all eukaryotic biology.

If the reality that life depends upon cognition is embraced, then in turn, any such cognition is naturally a property that must have devolved from a pre-existing physical state and the conditions that preceded that faculty. Therefore, life must adhere to basic physics, and then too, cognition must as well. Accumulating evidence supports that an appropriate frame for understanding the evolution of cognition lies within an inferential understanding of physics in a quantum informational framework as expressed in biological terms. Indeed, recent research is demonstrating that quantum processes are essential to life (Aerts et al., 2011; Wang et al., 2013). Awareness is both content and the awareness of that content which then, through quantum inference, participates in the settling of cognitive ambiguities (Conte et al., 2009). It has been proposed that many elements of the cytoskelton are crucial to the process of consciousness, particularly microtubules that demonstrate coordinated vibrational beat frequencies that may produce quantum coherences that permit the collapse of the superimposition of possibilities inherent to quantum phenomena (Hameroff and Penrose, 2014). Additional research has implicated other intracellular participants in information transfer and cognition in partnership with microtubules, such as actin filaments and collagen (Friesen et al., 2015). Similar types of quantum signal propagation have been observed within tubulin subunit proteins that comprise microtubules, particularly in the chromophores in light harvesting photosynthetic complexes (Craddock et al., 2014). Similar quantum phenomena have also been ascribed to non-polar protein interiors and membrane lipid peroxidation processes that interact either directly with microtubules or indirectly through serotonin production (Tonello et al., 2015).

With that as precursor, it can be maintained that in direct terms, any system of cognition that might eventually be embodied in the mammalian brain must be based upon quantum processes that are similar to those that produce an exchange of information between elemental receptive entities (Nurse, 2008; Walker and Davies, 2013). It is a necessary conclusion then, that physiology supports this preconditioning state in eukaryotes (Torday, 2015a). However, on a thermodynamic basis, this necessary exchange of information is also a transfer of energy linked to heat production. Therefore, any cognitive action as a form of cellular coherence can be better understood as both an information exchange and reciprocally then, as energy conversion and transfer (Adolphs and Renger, 2006; Dahlberg et al., 2015).

When this is our consideration, then "self" is best considered as a function of both energy and information transfer whose targets need not be identical. Since these latter two faculties are amply demonstrated within unicellular life, then "elf" is also invested at that scale. Even within that scope, when communication is accomplished, and information is transmitted as communication between a sender and a receiver, then both independent entities are now linked through that process (De Loof, 2015). It would seem evident then that self-awareness arises as a derivative of physical processes, based upon coordinate and reicprocating functions that understand its context within the larger organism. In multicellular organisms then, self-awareness becomes a function of cellular constituencies that both compete and collaborate.

Therefore, in a cellular frame, self-awareness as a cognitive function must be dependent upon the discrimination of cellular status through active cell homeostasis as the basis upon which physiology is constructed (Takada and Jameson, 2009). That faculty of discrimination of biologic status as opposed to the external environment as assessed through differential cellular physiological function thereby becomes the basis for self-awareness. "Self" is therefore cognitive awareness of homeostatic flux as maintained within cellular boundaries. Since homeostatic flux is itself dependent upon continuous physiologic activity, then "self" even as its own property must then be interpreted through physiological mechanisms.

Within any quantum biologic frame, dissipation of energy as generated heat can be viewed in terms of disappearance and emergence of coherence within and between cells (Engel et al., 2007; Larson, 2014). This parallel process is best understood as based upon cellular properties effectuated in support of the imperative of cellular homeostasis to maintain self. Such coherence can properly be construed as the ability of biologic organisms to resolve ambiguities toward survival in preferred states. Yet, in biologically active systems, all such actions have a thermodynamic (energy) cost, producing the need for active heat exchange with the environment.

In biological terms, this thermodynamic gradient is enacted as a series of downhill thermodynamic paths balancing energy usage and output with energy (heat) dissipation (Aledo and del Valle, 2004). It can be presumed then that this is best achieved

in the multicellular form. Therefore, multicellular physiology is a mechanistic solution for the utilization of energy and heat dissipation in order to control local homeostasis upon which "self" is dependent. It is this process in series that yields multicellular entities directed toward that same end that are appraised by us in biologic terms as evidence of "self-organization."

Any such reiterative process must always represent a continuum from basic thermodynamic principles brought forward as varied biologic manifestations based upon energy utilization and information transfer. Although there are crucial transitions within biologic phenomena by which free energy in thermodynamic terms must be differentiated from variational minimal free energy in biologic information space, the dissipation of heat inherent to all life forms can be considered as active suppression of free energy toward its minimum, a process that has been directly linked to how the brain acts to limit prediction errors (Friston et al., 2006). In this manner, any agent tends toward self-organization by minimizing free energy and thereby lowering any probablility of surprise. It proceeds in that direction through reciprocating interaction with its environment that rests upon Bayesian inferences about its context (Friston, 2009). In a cellular-based dynamic for cognition such as is being proposed, each cell is recognized as a discrete cognitive entity that acts both individually and collectively. Therefore, the statistical power of Markov blankets is pertinent as a descriptor of the manner in which cellular membranes uphold their intracellular matrix and separate from extracellular influences. Within the general conditions of that frame, Bayesian inferences are based upon random dynamical systems as features of variational free energy and local coupling (Friston et al., 2014). However, in a system based upon self-referential cognition as a thermodynamically-derived state function, there are direct biological limits placed upon the bounded dispersion of sensed states by which cells experience epiphenomena and the outward environment (Friston, 2013). Therefore, in quantum biological circumstances, these critical models can be subject to modification within the proscriptions of self-awareness and the constraining physiologic processes that are enacted to sustain it. The inevitability of self-organization as a form of active Bayesian inference within the spatial boundary conditions of any living organism therefore remains valid but becomes empowered. Biological uncertainties are resolved through reciprocal biological signaling in which inputs are not necessarily random, and via coupling in quantum systems that are subject to both local and non-local correlations (Al-Khalili and McFadden, 2014). Therefore, the ability of biological organisms to settle ambiguities within expanded inferential terms extends beyond typical statistical informational matrices, and thereby becomes a definitional crux of biological action. In multicellular entities then, any instrinsic drive toward a mandate to minimize variational free energy places each cell not merely as "in" its environment, but as a reciprocating organic entity that is "of" its external milieu at the same time. In this way, it becomes a specific embodiment of the "good regulator system" (Conant and Ross Ashby, 1970). It is continually isomorphic through self-assessment of its internal milieu as opposed to its external environment, of which it is a both quantum observer and participant. This consistent reciprocation becomes the specific basis for its homeostatic regulatory mechanism, both minimizing variational free energy, and underscoring its self-referential appraisal of conditional status.

Importantly, the "self" that exists within all living entities is itself a state function. Consciousness as awareness of an external environment is present in every form of life that is differentiated from the inanimate (Giuditta, 2010; Trewavas and Baluška, 2011). As such, it must be considered as a basic property of biology as a quantum system. Similar to energy and enthalpy, that state function may have differing values but is independent of the exact number of steps that were required to create its exact moment. Therefore, entropy, enthalpy, and self are linked, yet separable variables as state functions that are all directed toward maintaining homeostatic balance against environmental stresses. Although homeostatic status is dependent upon physiologic processes to efficiently utilize energy and dissipate heat, this process still must be co-aligned with those major state functions. Therefore, it can properly be imputed that physiologic pathways maintain cellular homeostasis as the means by which biologic substrates are utilized and purposed to sustain "self," as a basic property of living systems. Therefore, physiology becomes more than a series of interrelated processes but represents a further enactment of "self" as a state function as it is achieved through multicellular networks through achievable thermodynamic states. Physiology can then be viewed as more than a means of maintaining protective cellular homeostasis, but further, as an active agency that both permits and sustains "self." This is accomplished through cell-cell communication as the exchange of information that underscores self-identity. This activity is an energy-intensive process. Multicellular organisms might then be assumed to represent that form that best utilizes energy transformation for information sharing, and thereby, self-organization can now be best understood in a biologic context as a series of linked processes by which thermodynamically advantaged solutions to environmental stresses are achieved through form.

Although not obvious, multicellularity need not have been a necessary evolutionary outcome. Intracellular engineering might have led to enormously large, efficient and capable single cells. There is an analogy in the viral realm to support this case. The giant mimiviruses are larger in size than some bacteria, and have larger genomes (Moreira and Brochier-Armanet, 2008; Raoult and Forterre, 2008). However, in the cellular realm, this has not been our known biologic outcome. A salient question within biology might be "why didn't evolution lead to single extremely large and efficient cells as the dominant biologic players?"

That answer lies within physiologic mechanisms that extend forward from unicellular roots. These are based upon stable principles of evolutionary development that can be traced from those unicellular origins as thermodynamically effective outcomes for the maintenance of cellular identities and cellular homeostasis directed by a sustaining "self."

So we know where consciousness/mind originated from, and how it appears in its fully formed state, but how did it transition from paramecia to Einstein? The key to such an analytic approach

to the evolution of mind through paths of complex physiology is the realization that any given physiologic trait is the permutations and combinations of pre-existing unicellular mechanisms and physiologic traits (Torday, 2015b). This does not proceed in a linear, arithmetic fashion, but as Boolean contingencies on previous events in the history of the organism. Knowing the molecular "parts list" and the contexts in which they have existed in previous iterations and persist as current physiology and metabolism provides important clues to how and why they are relevant to brain evolution.

One important contextual clue is the water-land transition that occurred some 300 million years ago due to the CO_2 "greenhouse effect" (Ward et al., 2006), drying up lakes, rivers, and ponds (Romer, 1949). That precipitated several specific gene duplications in vertebrates due to the existential stress of having to adapt to land (Torday, 2013), providing insights to pivotal physiologic changes in vertebrate physiology- the lung, kidney, skeleton, skin. This paper is predicated on the hypothesis that like those visceral organs, the brain also evolved under selection pressure, as first proposed by Hughlings-Jackson (Franz and Gillett, 2011), and reinforced by Gottlieb (2007), who pointed out the norms of reaction in brain structure/function. Thereby, it can be advanced that there is an underlying cellular apparatus that appraises self through boundaries and proscriptions, thereby purposing "self" toward common cellular solutions to maintain homeostasis in reaction to environmental stresses that affect all the cellular constituencies of macro-organisms. This is achieved through physiological/metabolic pathways.

GENETIC CHANGES ASSOCIATED WITH THE WATER-LAND TRANSITION FACILITATE VERTEBRATE EVOLUTION

The initiating factor for vertebrate evolution was the insertion of cholesterol into the phospholipid bilayer (Miao et al., 2002), rendering the membrane more compliant by thinning it out and making it more permeable for gas exchange (Torday and Rehan, 2012). This is the fundament of vertebrate evolution that enabled endo/exocytosis, increased metabolism, and facilitated locomotion (Perry and Carrier, 2006).

The beginnings of the evolution of the visceral organs in vertebrates is the adaptation of water-based life forms to a primary existence on land. In support of this, there were two gene duplications for the Parathyroid Hormone-related Protein Receptor (PTHrPR) (Pinheiro et al., 2012) and the Beta Adrenergic Receptor (βAR) (Aris-Brosou et al., 2009), accompanied by two gene mutations, for the Glucocorticoid Receptor (GR) (Bridgham et al., 2006) and type IV collagen (MacDonald et al., 2006). The first three of these genetic changes were critically important in the evolution of land adaptive visceral organ changes. The best known is the PTHrP signaling mechanism, which is necessary for the formation of lung alveoli (Rubin et al., 1994), the primary mechanism for lung evolution (Torday and Rehan, 2007)—deletion of PTHrP results in failed alveolar formation (Rubin et al., 1994). It also affects bone (Karaplis and Goltzman, 2000), skin (Wysolmerski et al., 1998), kidney (Hochane et al., 2013), and brain (Liu et al., 2013), though not as profoundly as the lung- mice lacking the PTHrP gene die at birth of pulmonary insufficiency (Karaplis et al., 1994). The βAR was necessary for the earlier stages of lung evolution, increased βAR density in the pulmonary microcirculation allowing for blood pressure regulation independent of the systemic circulation (West and Mathieu-Costello, 1999). This adaptation was critical for the increase in gas exchange surface area of the lung, without which, every time there was a physiologic reaction to stress the microcirculation of the nascent lung would have been damaged. The GR evolved from the Mineralocorticoid Receptor (MR) (Bridgham et al., 2006), likely due to the elevated blood pressure on land vs. water (Volkmann and Baluska, 2006). The addition of two amino acids to the MR resulted in the evolution of the GR (Bridgham et al., 2006); the other positive selection for the GR during the water-land transition was its molecular induction by activation of βARs (Maier et al., 1989), effectively reducing blood pressure under stress conditions. The type IV collagen mutation that causes Goodpasture's Syndrome was adaptive for both the alveolus and glomerulus because it is hydrophobic (MacDonald et al., 2006), forming a barrier against the loss of fluid and electrolytes from the lung and kidney on land. However, people expressing this isomer of type IV collagen can develop autoantibodies to it, inhibiting gas exchange in the alveolus and blood filtration in the glomerulus, eventually resulting in death (Greco et al., 2015).

There is an interesting fundamental mechanistic difference between the PTHrP and βAR receptor gene duplications and the GR mutation, and that of the type IV collagen matrix protein, all of which were responses to physiologic stress caused by increased blood pressure, generating oxygen radicals in the microcirculation (De Nigris et al., 2001). In the case of the receptors, their expression was constrained by their previously evolved down-stream signaling pathways (Jordan et al., 2000), referred to in conventional evolutionary biology as terminal addition (Jacobs et al., 2005), whereas the type IV collagen mutation had no such specific servo-regulatory constraints, explaining why the consequent disease.

ON THE EVOLUTION OF ENDOTHERMY

Given the described step-wise empiric adaptation to land, there were undoubtedly stages in this process at which the lung was inefficient for gas exchange, resulting in hypoxia, the most potent physiologic stressor known. Such stress conditions would have resulted in spates of catecholamine production by the adrenal gland, alleviating the constraint of an inefficient lung by stimulating surfactant production by the alveoli (Lawson et al., 1978), acutely increasing alveolar distensibility and thus oxygenation. Over time, such episodes of over-distension of the alveoli would have culminated in the formation of additional alveoli, PTHrP acting to form more alveolar units (Rubin et al., 1994), accommodating oxygen deficiency constitutively. In tandem, catecholamines would have stimulated fatty acid secretion by fat cells in the periphery (Lawson et al., 1978), increasing metabolism and body heat (Lee et al., 2015). Again,

over time this mechanism would have given rise to a constitutive increase in body temperature, or endothermy. As evidence for this mechanism of evolution, PTHrP signaling appears in the pituitaries of mammals (Mamillapalli and Wysolmerski, 2010) and birds (Nakayama et al., 2011), and stimulates corticosteroid production by the adrenal cortex (Mazzocchi et al., 2001) in association with increased microvascularization of the adrenal medulla (Wurtman, 2002). Consequently, in the pituitary PTHrP amplifies ACTH (Mamillapalli and Wysolmerski, 2010), and in the adrenal cortex it stimulates corticoid production (Kawashima et al., 2005). The associated increase in angiogenesis within the adrenal medulla amplifies corticosteroid stimulation of the rate-limiting step in catecholamine synthesis, Catecholamine-O-Methyltransferase (Nic a' Bháird et al., 1990). The enhanced microcirculation within the medulla was likely due to the increased production of PTHrP within the adrenal cortex passing through the adrenal medulla since PTHrP is angiogenic (Isowa et al., 2010). The other consequence of increased catecholamine production elevating body temperature was the evolution of lung surfactant in adaptation to endothermy (Torday, 2015a). The composition of the surfactant phospholipid changes to dipalmitoylphosphatidylcholine (Suri et al., 2012), which has a phase transition temperature of 41°C, rendering it 3-times more active than it is at 25°C (Lau and Keough, 1981). In support of this hypothetical interrelationship, catecholamines have an adaptive effect on peripheral cellular oxygenation, increasing the amount of unsaturated phospholipid in the cell membrane, making it more gas permeable (Ward et al., 2006). In contrast to this, under hibernation conditions when oxygen utilization and stress are at a minimum, there is decreased unsaturated phospholipid in the peripheral cell membranes (Lau and Keough, 1981), decreasing cellular oxygen uptake, and lung surfactant phospholipid composition reverts to its cold-blooded composition (Suri et al., 2013). Experimental support for this integrated mechanism comes from study of MAP turtles reared at different ambient temperatures, altering the composition of their lung surfactant consistent with the previously described evolutionary changes (Lau and Keough, 1981).

ON THE EVOLUTION OF THE BRAIN IN ENDOTHERMS

The evolution of endothermy/homeothermy is a milestone in vertebrate evolution shared by mammals and birds (Grigg et al., 2004). What else do these organisms share in common that might give insight to evolution? It may be of evolutionary relevance that mammals and birds are both bipedal. That trait may have been contingent on the evolution of endothermy since being warm blooded rendered metabolism much more efficient. Cold-blooded poikilotherms require multiple isoforms of the same metabolic enzyme to function efficiently at different ambient temperatures (Duarte et al., 2007). Bipedalism requires much more energy than being quadrupedal (Rodman and McHenry, 1980). Another trait held in common by mammals and birds that may be a consequence of endothermy and bipedalism is the freeing of the forelimbs for such adaptations as flight in birds, and manual manipulation in Man.

There are clues within this narrative toward the evolution of the central nervous system in the context of the convergence of thermodynamic, self-organizational and informational characteristics. Importantly, there are exaptive traits that fostered the higher consciousness of mammals and birds. The role of thermoregulation has been alluded to by Hobson et al. (2014), invoking the need to cool the brain during Rapid Eye Movement sleep. As for the self-organizational aspect, the evolution of lipid metabolism converged in endothermy and pulmonary alveolar respiration (Torday, 2015a), perhaps acting as positive selection for neuregulin, an intermediary in the Epidermal Growth Factor signaling pathway that mediates both alveolar and neuronal lipid utility. In the alveolus, neuregulin promotes surfactant phospholipid synthesis (Fiaturi et al., 2014), whereas in the brain it is the only known mechanism that mediates myelinization of axons by Schwann Cells (Li, 2015). As previously described, the co-evolution of the pulmonary and neuroendocrine systems via PTHrP signaling is more evident, and may have been the forerunner of the neuregulin exaptation. Neuregulin mediates Schwann cell myelinization of neurons in the skin (McKenzie et al., 2006), the latter bearing a strong homology with the lung and brain. Neuregulin' role in myelinization is consistent with the informational phenotype proffered by Shannon's Communication Theory (Shannon and Weaver, 1949). And the evolution of the forelimb reinforced such co-evolved mechanisms at multiple levels, referring all the way back to the advent of cholesterol in the cell membrane of unicellular eukaryotes promoting locomotion, respiration and metabolism (Perry and Carrier, 2006).

Another co-evolved molecular mechanism common to the lung, adipose tissue and brain is leptin, which is secreted by both fat cells (Adamczak and Wiecek, 2013) and the lipofibroblasts of the alveolar wall (Torday et al., 2002). The role of leptin in endothermy/homeothermy has been demonstrated by the experimental treatment of cold-blooded Fence Lizards with leptin, increasing their basal metabolic rate and body temperature (Niewiarowski et al., 2000). In the brain, leptin stimulates the arborization of the neurons (Moult and Harvey, 2008), increasing informational processing properties in the brain.

Thus, the arc of vertebrate physiologic evolution is better understood as a continuum emanating from the water-land transition (Torday, 2013, 2015a), through only a few crucial pleiotropic gene duplications and mutations that occurred in that era—the Parathyroid Hormone-related Protein (PTHrP) Receptor, the βAdrenergic Receptor (βAR), and the Glucocorticoid Receptor (GR). The PTHrPR in particular fostered the evolution of a number of key terrestrial adaptations, most importantly the lung alveolus. The βAR was similarly critical in adaptation to air breathing since its regulation of blood pressure in both the systemic and pulmonary circulations was a constraint on the expansion of the surface area of the gas exchanger (West and Mathieu-Costello, 1999).

The evolution of the vertebrate lung from the fish swim bladder (Zheng et al., 2011) was principally due to the progressive

reduction in the air space formed initially by the swim bladder of fish, resulting in the increase in the gas exchange surface area-to-vascular blood supply ratio (Torday and Rehan, 2012). In order for that to occur, the alveolar epithelial type II cells lining the air space/alveolus had to increase the efficiency of surfactant surface tension reducing capacity to counter the increasing surface tension due to the reduction in the alveolar diameter- the surface tension of a sphere being inversely proportional to its diameter (Law of Laplace). It is this dynamic interplay of epithelial-mesenchymal interactions mediated by soluble paracrine growth factors and their cognate receptors (Torday and Rehan, 2012) that orchestrated the evolution of air breathing, from the swim bladder of fish to the lungs of amphibians, reptiles, mammals and birds.

It is through understanding these sorts of biophysical mechanisms that the evolution of the brain, particularly as it relates to self-organizational processes, can be explored as a productive reconciliation between thermodynamic necessities and informational requirements. At each moment, organizational problems secondary to system wide epiphenomena are being solved. It has been hypothesized that the interaction between the evolving lung and neuroendocrine system gave rise to endothermy in a series of steps. Intermittent periods of hypoxia during the water-land transition would have caused physiologic stress, hypoxia being the most potent physiologic stressor known in vertebrates. The stress would have been alleviated by the production of catecholamines by the adrenal medulla, relieving the initial constraint on gas exchange by stimulating the alveolar secretion of surfactant, rendering the evolving alveoli more distensible, acutely increasing their surface area for gas exchange. Over the long haul, these bouts with hypoxia would have fostered more alveoli since the distension of the alveoli stimulates PTHrP signaling, fostering more alveoli. In tandem with the stimulation of lung evolution, catecholamines would have stimulated the secretion of fatty acids from fat cells, in turn leading to increased metabolism and body temperature. Consequently, endothermy evolves over time in adaptation to terrestrial life, as a thermodynamically efficient solution to the chain of informational transfer that supports life and upon which "self" is based. In support of this hypothesis, PTHrP signaling for ACTH appears in the mammalian and avian pituitaries and in the adrenal cortex in association with increased microvasculature in the adrenal medulla. These physiologic changes would have enhanced catecholamine production in response to stress, synergizing the evolution of both the alveoli and endothermy. Systematically, the physiologic effect of catecholamines on the gas permeability of the cell membrane promotes oxygenation- the catecholamines promote the population of the cell membrane by unsaturated phospholipids, rendering it more fluid and therefore more permeable to gas exchange. This property is likely causal since the opposite effect is seen during hibernation and torpor, regarding both the composition of the cell membrane phospholipids, and that of the lung surfactant.

Therefore, factors that seem separable on an evolutionary basis are in fact united. They extend through demonstrable evolutionary paths that must remain adherent to those basic principles that underscore all such connections. This then provides for neuregulin mediation of both lung development and axonal myelinization, or the skin-brain connection, with their joint embryological connections (Foster, 2012). The Friston-Hobson (Hobson et al., 2014) theory of brain cooling coalesces with endothermy as both become implicit with regard to physical requirements in support of information systems. Therefore, the suspension of thermoregulation during Rapid Eye Movement sleep infers a "reverse evolution" reflecting that endothermy evolved from poikilothermy, and "Brain Cooling" becomes functionally homologous with the implementation of catecholamines for thermoregulation within the brain.

Complex physiology emerged as exaptations from the unicellular realm (Torday, 2015a). Therefore, self-organization can be understood as a direct means of protecting cellular physiological homeostasis that in turn sustains self-identity within consonant thermodynamic roots. At every scope and scale, self-aware cells enact solutions to environmental stresses according to thermodynamically efficient paths. "Self" exists within those thermodynamic necessities at the cellular level and then proceeds through multicellular reiterative physiological mechanisms to become, in an eventual series, our human brain. This accounts for a mammalian brain that is, in thermodynamic terms, an open system on an entropic basis and is energetically dissipative (Freeman and Vitello, 2011; Varpula et al., 2013).

Therefore, the distributive nature of mammalian cognition across widely separated cellular networks can be concluded to be derivative of both cellular-based cognition and also physiological mechanisms permitting energy dissipation in a manner that yields neurohumoral coherence across space-time. Physiological mechanisms both permit and support cellular "self" as a state function that emanates from unicellular origins and perpetually exists within that frame. That resultant network of cellular self-identity is always adherent to thermodynamic limits and is further delimited by cellular homeostasis that is itself dependent upon physiological pathways extending forward from their unicellular origins. Multicellular networks reiterate toward mammalian cognition by purposing self-organization as responsiveness to cellular self-identification, ever dependent upon the maintenance of cellular homeostatic flux boundaries, and sustained and advanced by cellular physiologic mechanisms in continual adaptation to epiphenomena.

AUTHOR CONTRIBUTIONS

JT contributed 50%; WM contributed 50%.

FUNDING

JT has been supported by NIH Grant HL055268

REFERENCES

Adamczak, M., and Wiecek, A. (2013). The adipose tissue as an endocrine organ. *Semin. Nephrol.* 33, 2–13. doi: 10.1016/j.semnephrol.2012.12.008

Adolphs, J., and Renger, T. (2006). How proteins trigger excitation energy transfer in the FMO complex of green sulfur bacteria. *Biophys. J.* 91, 2778–2797. doi: 10.1529/biophysj.105.079483

Aerts, D., Broekaert, J., and Gabora, L. (2011). A case for applying an abstracted quantum formalism to cognition. *New Ideas Psychol.* 29, 136–146. doi: 10.1016/j.newideapsych.2010.06.002

Aledo, J. C., and del Valle, A. E. (2004). The ATP paradox is the expression of an economizing fuel mechanism. *J. Biol. Chem.* 279, 55372–55375. doi: 10.1074/jbc.M410479200

Al-Khalili, J., and McFadden, J. (2014). *Life on the Edge. The Coming of Age of Quantum Biology*. London: Bantam Press.

Aris-Brosou, S., Chen, X., Perry, S. F., and Moon, T. W. (2009). Timing of the functional diversification of alpha- and beta-adrenoceptors in fish and other vertebrates. *Ann. N.Y. Acad. Sci.* 1163, 343–347. doi: 10.1111/j.1749-6632.2009.04451.x

Baluška, F., and Mancuso, S. (2009). Deep evolutionary origins of neurobiology: turning the essence of 'neural' upside-down. *Commun. Integr. Biol.* 2, 60–65. doi: 10.4161/cib.2.1.7620

Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312, 97–101. doi: 10.1126/science.1123348

Conte, E., Khrennikov, A. Y., Todarello, O., Federici, A., Mendolicchio, L., and Zbilut, J. P. (2009). Mental states follow quantum mechanics during perception and cognition of ambiguous figures. *Open Syst. Inf. Dyn.* 16, 85–100. doi: 10.1142/S1230161209000074

Conant, R. C., and Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Craddock, T. J. A., Friesen, D., Mane, J., Hameroff, S., and Tuszynski, J. A. (2014). The feasibility of coherent energy transfer in microtubules. *J. R. Soc. Interface* 11:20140677. doi: 10.1098/rsif.2014.0677

Dahlberg, P. D., Norris, G. J., Wang, C., Viswanathan, S., Singh, V. P., and Engel, G. S. (2015). Communication: coherences observed *in vivo* in photosynthetic bacteria using two-dimensional electronic spectroscopy. *J. Chem. Phys.* 143, 101101. doi: 10.1063/1.4930539

De Loof, A. (2015). Organic and cultural evolution can be seamlessly integrated using the principles of communication and problem-solving: the foundations for an extended evolutionary synthesis (EES) as outlined in the Mega-Evolution concept. *Life Exc. Biol.* 2, 247–269. doi: 10.9784/leb2(4)deloof.01

De Nigris, F., Lerman, L. O., Condorelli, M., Lerman, A., and Napoli, C. (2001). Oxidation-sensitive transcription factors and molecular mechanisms in the arterial wall. *Antioxid. Redox Signal.* 3, 1119–1130. doi: 10.1089/152308601317203620

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1777–1782. doi: 10.1073/pnas.0610772104

Engel, G. S., Calhoun, T. R., Read, E. L., Ahn, T. K., Mančal, T., Cheng, Y. C., et al. (2007). Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature* 446, 782–786. doi: 10.1038/nature05678

Fiaturi, N., Castellot, J. J. Jr., and Nielsen, H. C. (2014). Neuregulin-ErbB4 signaling in the developing lung alveolus: a brief review. *J. Cell Commun. Signal.* 8, 105–111. doi: 10.1007/s12079-014-0233-y

Foster, P. P. (2012). The "Brain–Skin Connection" in protein misfolding and amyloid deposits: embryological, pathophysiological, and therapeutic common grounds? *Front. Neurol* 3:56. doi: 10.3389/fneur.2012.00056

Franz, E. A., and Gillett, G. (2011). John Hughlings Jackson's evolutionary neurology: a unifying framework for cognitive neuroscience. *Brain* 134, 3114–3120. doi: 10.1093/brain/awr218

Freeman, W. J., and Vitello, G. (2011). The dissipative brain and non-equilibrium thermodynamics. *J. Cosmol.* 14, 4461–4468.

Friesen, D. E., Craddock, T. J., Kalra, A. P., and Tuszynski, J. A. (2015). Biological wires, communication systems, and implications for disease. *Biosystems* 127, 14–27. doi: 10.1016/j.biosystems.2014.10.006

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cog. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K., SenGupta, B., and Auletta, G. (2014). Cognitive dynamics: from attractors to active inference. *Proc. IEEE* 102, 427–445. doi: 10.1109/JPROC.2014.2306251

Giuditta, A. (2010). The origin and philogenetic role of mind. *Hum. Evol.* 25, 221–227.

Gottlieb, G. (2007). Probabilistic epigenesis. *Dev. Sci.* 10, 1–11. doi: 10.1111/j.1467-7687.2007.00556.x

Greco, A., Rizzo, M. I., De Virgilio, A., Gallo, A., Fusconi, M., Pagliuca, G., et al. (2015). Goodpasture's syndrome: a clinical update. *Autoimmun. Rev.* 14, 246–253. doi: 10.1016/j.autrev.2014.11.006

Grigg, G. C., Beard, L. A., and Augee, M. L. (2004). The evolution of endothermy and its diversity in mammals and birds. *Physiol. Biochem. Zool.* 77, 982–997. doi: 10.1086/425188

Hameroff, S., and Penrose, R. (2014). Consciousness in the universe: a review of the 'Orch OR' theory. *Phys. Life Rev.* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002

Hobson, J. A., Hong, C. C., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Hochane, M., Raison, D., Coquard, C., Imhoff, O., Massfelder, T., Moulin, B., et al. (2013). Parathyroid hormone-related protein is a mitogenic and a survival factor of mesangial cells from male mice: role of intracrine and paracrine pathways. *Endocrinology* 154, 853–8564. doi: 10.1210/en.2012-1802

Ikeda, S., Namekawa, T., Sugimoto, M., and Kume, S. I. (2010). Expression of methylation pathway enzymes in bovine oocytes and preimplantation embryos. *J. Exp. Zool. A Ecol. Genet. Physiol.* 313, 129–136. doi: 10.1002/jez.581

Isowa, S., Shimo, T., Ibaragi, S., Kurio, N., Okui, T., Matsubara, K., et al. (2010). PTHrP regulates angiogenesis and bone resorption via VEGF expression. *Anticancer Res.* 30, 2755–2767.

Jacobs, D. K., Hughes, N. C., Fitz-Gibbon, S. T., and Winchell, C. J. (2005). Terminal addition, the Cambrian radiation and the Phanerozoic evolution of bilaterian form. *Evol. Dev.* 7, 498–514. doi: 10.1111/j.1525-142X.2005.05055.x

Jordan, J. D., Landau, E. M., and Iyengar, R. (2000). Signaling networks: the origins of cellular multitasking. *Cell* 103, 193–200. doi: 10.1016/S0092-8674(00)00112-4

Karaplis, A. C., and Goltzman, D. (2000). PTH and PTHrP effects on the skeleton. *Rev. Endocr. Metab. Disord.* 1, 331–341. doi: 10.1023/A:1026526703898

Karaplis, A. C., Luz, A., Glowacki, J., Bronson, R. T., Tybulewicz, V. L., Kronenberg, H. M., et al. (1994). Lethal skeletal dysplasia from targeted disruption of the parathyroid hormone-related peptide gene. *Genes Dev.* 8, 277–289. doi: 10.1101/gad.8.3.277

Kawashima, M., Takahashi, T., Yanai, H., Ogawa, H., and Yasuoka, T. (2005). Direct action of parathyroid hormone-related peptide to enhance corticosterone production stimulated by adrenocorticotropic hormone in adrenocortical cells of hens. *Poult. Sci.* 84, 1463–1469. doi: 10.1093/ps/84.9.1463

Larson, C. S. (2014). Evidence of macroscopic quantum phenomena and conscious reality selection. *Cosmos Hist.* 10, 34–47.

Lau, M. J., and Keough, K. M. (1981). Lipid composition of lung and lung lavage fluid from map turtles (Malaclemys geographica) maintained at different

environmental temperatures. *Can. J. Biochem.* 59, 208–219. doi: 10.1139/o81-029

Lawson, E. E., Brown, E. R., Torday, J. S., Madansky, D. L., and Taeusch, H. W. Jr. (1978). The effect of epinephrine on tracheal fluid flow and surfactant efflux in fetal sheep. *Am. Rev. Respir. Dis.* 118, 1023–1026.

Lee, J., Ellis, J. M., and Wolfgang, M. J. (2015). Adipose fatty acid oxidation is required for thermogenesis and potentiates oxidative stress-induced inflammation. *Cell Rep.* 10, 266–279. doi: 10.1016/j.celrep.2014.12.023

Li, J. (2015). Molecular regulators of nerve conduction - lessons from inherited neuropathies and rodent genetic models. *Exp. Neurol.* 267, 209–218. doi: 10.1016/j.expneurol.2015.03.009

Liu, X. L., Lu, Y. S., Gao, J. Y., Marshall, C., Xiao, M., Miao, D. S., et al. (2013). Calcium sensing receptor absence delays postnatal brain development via direct and indirect mechanisms. *Mol. Neurobiol.* 48, 590–600. doi: 10.1007/s12035-013-8448-0

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Front. Microbiol.* 6:264. doi: 10.3389/fmicb.2015.00264

MacDonald, B. A., Sund, M., Grant, M. A., Pfaff, K. L., Holthaus, K., Zon, L. I., et al. (2006). Zebrafish to humans: evolution of the alpha3-chain of type IV collagen and emergence of the autoimmune epitopes associated with Goodpasture syndrome. *Blood* 107, 1908–1915. doi: 10.1182/blood-2005-05-1814

Maier, J. A., Roberts, J. M., and Jacobs, M. M. (1989). Ontogeny of fetal adenylate cymechanisms for regulation of beta-adrenergic receptors. *J. Dev. Physiol.* 12, 249–261.

Mamillapalli, R., and Wysolmerski, J. (2010). The calcium-sensing receptor couples to Galpha(s) and regulates PTHrP and ACTH secretion in pituitary cells. *J. Endocrinol.* 204, 287–297. doi: 10.1677/JOE-09-0183

Mazzocchi, G., Aragona, F., Malendowicz, L. K., and Nussdorfer, G. G. (2001). PTH and PTH-related peptide enhance steroid secretion from human adrenocortical cells. *Am. J. Physiol. Endocrinol. Metab.* 280, E209–E213.

McKenzie, I. A., Biernaskie, J., Toma, J. G., Midha, R., and Miller, F. D. (2006). Skin-derived precursors generate myelinating Schwann cells for the injured and dysmyelinated nervous system. *J. Neurosci.* 26, 6651–6660. doi: 10.1523/JNEUROSCI.1007-06.2006

Miao, L., Nielsen, M., Thewalt, J., Ipsen, J. H., Bloom, M., Zuckermann, M. J., et al. (2002). From lanosterol to cholesterol: structural evolution and differential effects on lipid bilayers. *Biophys. J.* 82, 1429–1444. doi: 10.1016/S0006-3495(02)75497-0

Miller, W. B. Jr. (2013). *The Microcosm within: Evolution and Extinction in the Hologenome*. Florida, FL: Universal-Publishers.

Minami, N., Suzuki, T., and Tsukamoto, S. (2007). Zygotic gene activation and maternal factors in mammals. *J. Reprod. Dev.* 53, 707–715. doi: 10.1262/jrd.19029

Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* 8:12. doi: 10.1186/1471-2148-8-12

Moult, P. R., and Harvey, J. (2008). Hormonal regulation of hippocampal dendritic morphology and synaptic plasticity. *Cell Adh. Migr.* 2, 269–275. doi: 10.4161/cam.2.4.6354

Nakayama, H., Takahashi, T., Oomatsu, Y., Nakagawa-Mizuyachi, K., and Kawashima, M. (2011). Parathyroid hormone-related peptide directly increases adrenocorticotropic hormone secretion from the anterior pituitary in hens. *Poult. Sci.* 90, 175–180. doi: 10.3382/ps.2010-00860

Nic a' Bháird, N., Goldberg, R., and Tipton, K. F. (1990). Catechol-O-methyltransferase and its role in catecholamine metabolism. *Adv. Neurol.* 53, 489–495.

Niewiarowski, P. H., Balk, M. L., and Londraville, R. L. (2000). Phenotypic effects of leptin in an ectotherm: a new tool to study the evolution of life histories and endothermy? *J. Exp. Biol.* 203, 295–300.

Nurse, P. (2008). Life, logic and information. *Nature* 454, 424–426. doi: 10.1038/454424a

Perry, S. F., and Carrier, D. R. (2006). The coupled evolution of breathing and locomotion as a game of leapfrog. *Physiol. Biochem. Zool.* 79, 997–999. doi: 10.1086/507657

Pinheiro, P. L., Cardoso, J. C., Power, D. M., and Canário, A. V. (2012). Functional characterization and evolution of PTH/PTHrP receptors: insights from the chicken. *BMC Evol. Biol.* 6:110. doi: 10.1186/1471-2148-12-110

Raoult, D., and Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nat. Rev. Microbiol.* 6, 315–319. doi: 10.1038/nrmicro1858

Rodman, P. S., and McHenry, H. M. (1980). Bioenergetics and the origin of hominid bipedalism. *Am. J. Phys. Anthropol.* 52, 103–106. doi: 10.1002/ajpa.1330520113

Romer, A. S. (1949). *The Vertebrate Story*. Chicago: University of Chicago Press.

Rubin, L. P., Kifor, O., Hua, J., Brown, E. M., and Torday, J. S. (1994). Parathyroid hormone (PTH) and PTH-related protein stimulate surfactant phospholipid synthesis in rat fetal lung, apparently by a mesenchymal-epithelial mechanism. *Biochim. Biophys. Acta* 1223, 91–100. doi: 10.1016/0167-4889(94)90077-9

Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication*. Chicago: University of Illinois Press.

Shapiro, J. A. (2011). *Evolution: A View from the 21st Century*. New Jersey, NJ: FT Press Science.

Suri, L. N., Cruz, A., Veldhuizen, R. A., Staples, J. F., Possmayer, F., Orgeig, S., et al. (2013). Adaptations to hibernation in lung surfactant composition of 13-lined ground squirrels influence surfactant lipid phase segregation properties. *Biochim. Biophys. Acta* 1828, 1707–1714. doi: 10.1016/j.bbamem.2013.03.005

Suri, L. N., McCaig, L., Picardi, M. V., Ospina, O. L., Veldhuizen, R. A., Staples, J. F., et al. (2012). Adaptation to low body temperature influences pulmonary surfactant composition thereby increasing fluidity while maintaining appropriately ordered membrane structure and surface activity. *Biochim. Biophys. Acta* 1818, 1581–1589. doi: 10.1016/j.bbamem.2012.02.021

Takada, K., and Jameson, S. C. (2009). Naive T cell homeostasis: from awareness of space to a sense of place. *Nat. Rev. Immunol.* 9, 823–832. doi: 10.1038/nri2657

Tonello, L., Cocchi, M., Gabrielli, F., and Tuszynski, J. A. (2015). On the possible quantum role of serotonin in consciousness. *J. Integr. Neurosci.* 14, 295–308. doi: 10.1142/S021963521550017X

Torday, J. S. (2013). Evolutionary biology redux. *Perspect. Biol. Med. Autumn.* 56, 455–484. doi: 10.1353/pbm.2013.0038

Torday, J. S. (2015a). A central theory of biology. *Med. Hypotheses.* 85, 49–57. doi: 10.1016/j.mehy.2015.03.019

Torday, J. S. (2015b). The cell as the mechanistic basis for evolution. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 7, 275–284. doi: 10.1002/wsbm.1305

Torday, J. S., and Rehan, V. K. (2007). The evolutionary continuum from lung development to homeostasis and repair. *Am. J. Physiol. Lung Cell Mol. Physiol.* 292, L608–L611. doi: 10.1152/ajplung.00379.2006

Torday, J. S., and Rehan, V. K. (2012). *Evolutionary Biology, Cell-Cell Communication and Complex Disease*. New Jersey, NJ: Wiley.

Torday, J. S., Sun, H., Wang, L., Torres, E., Sunday, M. E., and Rubin, L. P. (2002). Leptin mediates the parathyroid hormone-related protein paracrine stimulation of fetal lung maturation. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 282, L405–L410.

Trewavas, A. J., and Baluška, F. (2011). The ubiquity of consciousness. *EMBO Rep.* 12, 1221–1225. doi: 10.1038/embor.2011.218

Varpula, S., Annila, A., and Beck, C. (2013). Thoughts about thinking: cognition according to the second law of thermodynamics. *Adv. Stud. Biol.* 5, 135–149.

Volkmann, D., and Baluska, F. (2006). Gravity: one of the driving forces for evolution. *Protoplasma* 229, 143–148. doi: 10.1007/s00709-006-0200-4

Walker, S. I., and Davies, P. C. (2013). The algorithmic origins of life. *J. R. Soc. Interface* 10:20120869. doi: 10.1098/rsif.2012.0869

Wang, Z., Busemeyer, J. R., Atmanspacher, H., and Pothos, E. M. (2013). The potential of using quantum theory to build models of cognition. *Top. Cogn. Sci.* 5, 672–688. doi: 10.1111/tops.12043

Ward, P., Labandeira, C., Laurin, M., and Berner, R. A. (2006). Confirmation of Romer's Gap as a low oxygen interval constraining the timing of initial arthropod and vertebrate terrestrialization. *Proc. Natl. Acad. Sci. U.S.A.* 103, 16818–16822. doi: 10.1073/pnas.0607824103

West, J. B., and Mathieu-Costello, O. (1999). Structure, strength, failure, and remodeling of the pulmonary blood-gas barrier. *Annu. Rev. Physiol.* 61, 543–572. doi: 10.1146/annurev.physiol.61.1.543

Wurtman, R. J. (2002). Stress and the adrenocortical control of epinephrine synthesis. *Metab. Clin. Exp.* 51, 11–14. doi: 10.1053/meta.2002.33185

Wysolmerski, J. J., Philbrick, W. M., Dunbar, M. E., Lanske, B., Kronenberg, H., and Broadus, A. E. (1998). Rescue of the parathyroid hormone-related protein knockout mouse demonstrates that parathyroid hormone-related protein is essential for mammary gland development. *Development* 125, 1285–1294.

Zheng, W., Wang, Z., Collins, J. E., Andrews, R. M., Stemple, D., and Gong, Z. (2011). Comparative transcriptome analyses indicate molecular homology of zebrafish swimbladder and mammalian lung. *PLoS ONE* 6:e24019. doi: 10.1371/journal.pone.0024019

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer KF and handling Editor declared a current collaboration and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 07 June 2016
doi: 10.3389/fnsys.2016.00049

Universal Darwinism As a Process of Bayesian Inference

John O. Campbell *

Independent Researcher, Victoria, BC, Canada

Many of the mathematical frameworks describing natural selection are equivalent to Bayes' Theorem, also known as Bayesian updating. By definition, a process of Bayesian Inference is one which involves a Bayesian update, so we may conclude that these frameworks describe natural selection as a process of Bayesian inference. Thus, natural selection serves as a counter example to a widely-held interpretation that restricts Bayesian Inference to human mental processes (including the endeavors of statisticians). As Bayesian inference can always be cast in terms of (variational) free energy minimization, natural selection can be viewed as comprising two components: a generative model of an "experiment" in the external world environment, and the results of that "experiment" or the "surprise" entailed by predicted and actual outcomes of the "experiment." Minimization of free energy implies that the implicit measure of "surprise" experienced serves to update the generative model in a Bayesian manner. This description closely accords with the mechanisms of generalized Darwinian process proposed both by Dawkins, in terms of replicators and vehicles, and Campbell, in terms of inferential systems. Bayesian inference is an algorithm for the accumulation of evidence-based knowledge. This algorithm is now seen to operate over a wide range of evolutionary processes, including natural selection, the evolution of mental models and cultural evolutionary processes, notably including science itself. The variational principle of free energy minimization may thus serve as a unifying mathematical framework for universal Darwinism, the study of evolutionary processes operating throughout nature.

Keywords: free energy, natural selection, information, Bayesian inference, Universal Darwinism

OPEN ACCESS

Edited by:
Biswa Sengupta,
University of Cambridge, UK

Reviewed by:
Nelson Jesús Trujillo-Barreto,
The University of Manchester, UK
Pedro Larrañaga,
Technical University of Madrid, Spain
Adeel Razi,
University College London, UK

***Correspondence:**
John O. Campbell
jockocampbell@gmail.com

Received: *11 December 2015*
Accepted: *25 May 2016*
Published: *07 June 2016*

Citation:
Campbell JO (2016) Universal Darwinism As a Process of Bayesian Inference.
Front. Syst. Neurosci. 10:49.
doi: 10.3389/fnsys.2016.00049

INTRODUCTION

Although Darwin must be counted amongst history's greatest scientific geniuses, he had very little talent for mathematics. His theory of natural selection was presented in remarkable detail, with many compelling examples but without a formal or mathematical framework (Darwin, 1872). Darwin did not think in mathematical terms; he found mathematics repugnant and it comprised only a small part of his Cambridge education (Darwin, 1958).

Generally, mathematics is an aid to scientific theories because a theory whose basics are described through mathematical relationships can be expanded into a larger network of predictive implications and the entirety of the expanded theory subjected to the test of evidence. As a bonus, any interpretation of the theory must also conform to this larger network of implications to ensure some consistency.

Natural selection describes the change in frequency or probability of biological traits over succeeding generations. One might suppose that a mathematical description—complete with

an insightful interpretation—would be straightforward, but even today this remains elusive. The current impasse involves conceptual difficulties arising from one of mathematics' bitterest interpretational controversies.

That controversy is between the Bayesian and Frequentist interpretations of probability theory. Frequentists assume probability or frequency to be a natural propensity of nature. For instance, the fact that each face of a dice will land with 1/6 probability is understood by frequentists to be a physical property of the dice. On the other hand, Bayesians understand that humans assign probabilities to hypotheses on the basis of the knowledge they have (and the hypotheses they can entertain); thus the probability of each side of a dice is 1/6 because the observer has no knowledge that would favor one face over the other; the only way that no face is favored is for each hypothesis to be assigned the same probability. Furthermore, the value 1/6 is conditioned upon the assumption that there are only six possible outcomes. This means that probabilities are an attribute of a hypothesis or model space—not of the world that is modeled.

The Bayesian framework is arguably more comprehensive and has been developed into the mathematics of Bayesian inference, at the heart of which is Bayes' theorem, which describes how probabilistic models gain knowledge and learn from evidence. In my opinion, the major drawback of the Bayesian approach is an anthropomorphic reliance on human agency, the assumption that inference is an algorithm performed only by humans that possess (probabilistic) beliefs. Despite this interpretational dispute there has been some progress in uniting Bayesian and frequentist mathematics (Bayarri and Berger, 2004).

Despite the lack of mathematics in Darwin's initial formulation it was not long before researchers began developing a mathematical framework describing natural selection. It is an historical curiosity that most of these frameworks involved Bayesian mathematics, yet no interpretations were offered, proposing natural selection as a process of Bayesian inference.

The first step in developing this mathematics was taken during Darwin's lifetime by his cousin, Francis Galton. Galton developed numerous probabilistic techniques for describing the variance in natural traits—as well as for natural selection in general. His conception of natural selection was intriguingly Bayesian; although he may never have heard of Bayes' theorem. Evidence of his Bayesian bent is provided by a visual aid that he built for a lecture on heredity and natural selection given to the Royal Society (Galton, 1877).

He used this device (see **Figure 1** below) to explain natural selection in probabilistic terms. It contains three compartments: a top compartment representing the frequency of traits in the parent population, a middle one representing the application of "relative fitness" to the child generation and a third representing the normalization of the resulting distribution in the child generation. Beads are loaded in the top compartment to represent the distribution in the parent generation and then are allowed to fall into the second compartment. The trick is in the second compartment, which contains a vertical division, in the shape of the relative fitness distribution. Some of the beads fall behind this division and are "wasted"; they do not survive and are removed from sight. The remaining beads represent the distribution of the "survivors" in the child generation.

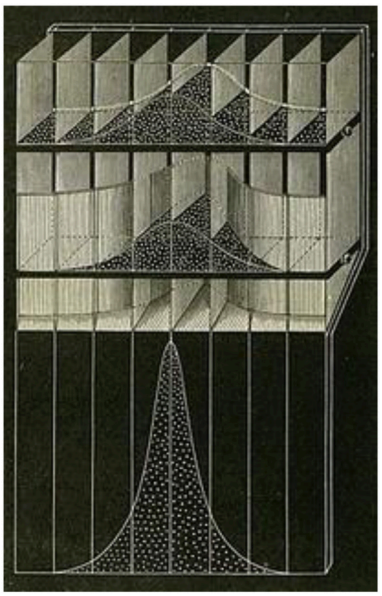


FIGURE 1 | A device constructed by Francis Galton as an aide in an 1877 talk he gave to the RoyalSociety. It is meant to illustrate generational change in the distribution of a population's characteristics due to natural selection.

Galton's device has recently been rediscovered and employed by Stephan Stigler and others in the statistics community as a visual aid, not for natural selection, but for Bayes' theorem. The top compartment represents the prior distribution, the middle one represents the application of the likelihood to the prior, and the third represents the normalization of the resulting distribution. The change between the initial distribution and the final one is the Bayesian update.

Fisher further developed the mathematics describing natural selection during the 1920s and 1930s. He applied statistical methods to the analysis of natural selection via Mendelian genetics and arrived at the fundamental theorem of natural selection which states (Fisher, 1930):

> The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.

Although Fisher was a fierce critic of the Bayesian interpretation (which he considered subjective) he pioneered—and made many advances with—the frequentist interpretation.

The next major development in the mathematics of natural selection came in 1970 with the publication of the Price equation, which built on the fundamental theorem of natural selection (Harper, 2010; Frank, 2012a). Although the Price equation fully describes evolutionary change, its meaning has only recently begun to be unraveled, notably by Steven A. Frank in a series of papers spanning the last couple of decades. Frank's insights into the meaning of the Price equation culminated in a 2012 paper

(Frank, 2012b) which derives a description of natural selection using the mathematics of information theory.

In my opinion, this paper represents a significant advance in the understanding of evolutionary change as it shifts the interpretation from the objective statistical description of frequentist probability to an interpretation in terms of Bayesian inference. Unfortunately, Frank does not share my appreciation of his accomplishment. While he understands that his mathematics are very close to those of Bayesian inference he does not endorse a Bayesian interpretation but prefers an interpretation in terms of information theory.

INFORMATION AND BAYESIAN INFERENCE

However, the mathematics of information theory and Bayesian probability are joined at the hip, as their basic definitions are in terms of one another. Information theory begins with a definition of information in terms of probability:

$$I\left(h_i|\mathrm{m}\right) = -log\left(P\left(h_i|\mathrm{m}\right)\right)$$

Here, we may view h_i as the i^{th} hypothesis or event in a mutually exclusive and exhaustive family of n competing hypotheses comprising a model m. $I\left(h_i|\mathrm{m}\right)$ is the information gained, under the model, on learning that hypothesis h_i is true. $P\left(h_i|\mathrm{m}\right)$ is the probability that had previously been assigned by the model that the hypothesis h_i is true. Thus, information is "surprise"; the less likely a model initially considers a hypothesis that turns out to be the case, the more surprise it experiences, and thus the more information it receives.

Information theory, starting with the very definition of information, is aligned with the Bayesian interpretation of probability; information is "surprise" or the gap between an existing state of knowledge and a new state of knowledge gained through receiving new information or evidence.

The model itself, composed of the distribution of the $p(h_i)$, may also be said to have an expectation. The information which the model "expects" is the weighted average of the information expected by the n individual $p(h_i)$, which is called the model's entropy.

$$S\left(H|m\right) = \sum_{1}^{n} p\left(h_i|\mathrm{m}\right) \ \left(-\log\left(p\left(h_i|\mathrm{m}\right)\right)\right)$$

Entropy is the amount of information that separates a model's current state of knowledge from certainty.

Bayes' theorem follows directly from the axioms of probability theory and may be understood as the implication that new evidence or information holds for the model described by the distribution of the $\mathrm{p(h_i)}$. This theorem states that on the reception of new information (I) by the model (m) the probability of each component hypothesis (h_i) making up the model updates according to:

$$\mathrm{P}\left(\mathrm{h_i} \mid \mathrm{I, m}\right) = \mathrm{P}\left(\mathrm{h_i} \mid \mathrm{m}\right) \frac{\mathrm{P}\left(\mathrm{I} \mid \mathrm{h_i m}\right)}{\mathrm{P}\left(\mathrm{I} \mid \mathrm{m}\right)}$$

Bayesian inference is commonly understood as any process which employs Bayes' theorem to accumulate evidence based knowledge (Wikipedia[1]): the quantity $\mathrm{P}\left(\mathrm{I} \mid \mathrm{m}\right)$ is called (Bayesian) model evidence and corresponds to the probability of observing some new information, under a particular model, averaged over all hypotheses. This is a crucial quantity in practice and can be used to adjudicate between good and bad models in statistical analysis. It is also the quantity approximated by (variational) free energy—as we will see below. Effectively, this equation provides the formal basis for Bayesian belief updating: in which prior beliefs about the hypotheses $\mathrm{P}\left(\mathrm{h_i} \mid \mathrm{m}\right)$ are transformed into posterior beliefs$\mathrm{P}\left(\mathrm{h_i} \mid \mathrm{I, m}\right)$, which are informed by new information. This updating rests upon the likelihood model; namely the likelihood of observing new information given the *i*-th hypothesis $\mathrm{P}\left(\mathrm{I} \mid \mathrm{h_i m}\right)$. This formalism highlights the information theoretic nature of Bayesian updating—and the key role of a (likelihood) model in accumulating evidence.

We may conclude from this short overview of the relationship between information and Bayesian inference that information has little meaning outside a Bayesian context. Information depends upon a model that assigns probabilities to outcomes and which is updated on the reception of new information. In short, there is no information unless there is something that can be informed. This something is a model.

Thus, we see that, contrary to Frank's view, Bayesian inference and information theory have the same logical structure. However, it is instructive to follow Frank's development of the mathematics of evolutionary change in terms of information theory, while keeping in mind his denial of its relationship to Bayesian inference. Frank begins his unpacking of the Price equation by describing the "simple model" he will develop:

> A simple model starts with n different types of individuals. The frequency of each type is q_i. Each type has w_i offspring, where w expresses fitness. In the simplest case, each type is a clone producing w_i copies of itself in each round of reproduction. The frequency of each type after selection is
>
> $$q_i^{'} = q_i \frac{w_i}{\mathrm{w}} \quad (1)$$
>
> Where $w = \sum_1^n q_i w_i$ is the average fitness of the trait in the population. The summation is over all of the n different types indexed by the i subscripts.

Equation (1) is clearly an instance of Bayes' theorem, where the new evidence or information is given in terms of relative fitness and thus Frank's development of this simple model is in terms of Bayesian inference.

While Frank acknowledges an isomorphism between Bayes' theorem and his simple model, he does not find this useful and prefers to describe the relationship as an analogy. He makes the somewhat dismissive remark:

[1]Wikipedia. *Bayesian Inference*. Available online at: https://en.wikipedia.org/wiki/Bayesian_inference (Accessed 26 September, 2015).

> I am sure this Bayesian analogy has been noted many times. But it has never developed into a coherent framework that has contributed significantly to understanding selection.

On the contrary, I would suggest that Frank's paper itself develops a coherent framework for natural selection in terms of Bayesian inference. In particular, he highlights the formal relationships between the Price equation (or replicator equation) and Bayesian belief updating (e.g., Kalman Filtering). This is potentially interesting because many results in evolutionary theory can now be mapped to standard results in statistics, machine learning and control theory. Although we will not go into technical details, a nice example here is that Fisher's fundamental theorem corresponds to the increase in Kalman gain induced by random fluctuations (this variational principle is well-known in control theory and volatility theory in economics). Despite this, Frank dismisses Bayesian formulations because they do not appear to bring much to the table. This is understandable in the sense that the mathematics traditionally used to describe natural selection already has a Bayesian form and merely acknowledging this fact does not lead to a new formalism. However, this conclusion might change dramatically if biological evolution was itself a special case of a Universal Darwinism that was inherently Bayesian in its nature. In what follows, we pursue this line of argument by appealing to the variational principle of least free energy.

FREE ENERGY MINIMIZATION PRINCIPLE

Baez and Pollard have recently demonstrated the similarities of a number of information-theoretic formulations, including the Bayesian replicator equation, evolutionary game theory, Markov processes and chemical reaction networks, that are applicable to biological systems as they approach equilibrium (Baez and Pollard, 2016). In general, any process of Bayesian inference may be cast in terms of (variational) free energy minimization (Roweis and Ghahramani, 1999; Friston, 2010) and—in this form—some important interpretative issues gain clarity. This approach has been used by Hinton, Friston, and others to describe the evolution of mental states as well as to describe pattern formation and general evolutionary processes. In its most general form, the free energy principle suggests that any weakly-mixing ergodic random dynamical system must be describable in terms of Bayesian inference. This means that the equivalence between classical formulations of evolution and Bayesian updating are both emergent properties of any random dynamical system that sustains measurable characteristics over time (i.e., is ergodic; Friston, 2013). This is quite important because it means that evolution is itself an emergent property of any such systems. Although conceptually intriguing, there may be other advantages to treating evolution in terms of minimizing variational free energy. In what follows, I will try to demonstrate this may be true.

In 1970 Ashby and Conant (Conant and Ashby, 1970) proved a theorem that any regulating mechanism for a complex system that is both successful and simple must be isomorphic with the system being regulated. In other words, it must contain a model of the system being regulated. As no model can be exactly isomorphic to its subject without being a clone and therefore exactly as complex as its subject, this theorem suggests a variational approach may be useful, one which optimizes the difference between the accuracy and the complexity of the model.

This is exactly a form in which the free energy minimization principle may be cast (Moran et al., 2014):

$$F(s,u) = D_{KL}\left[q(\psi|\mu)\,||\,p(\psi|m)\right] - E_q\left[\log p(s|\psi,m)\right]$$

Free Energy = Complexity-Accuracy

Where ψ are hidden states of the world or environment, s are their sensory consequences or samples (that can depend upon action), μ are internal states and m is the generative model. The distribution q is the current predictions of the states of the environment, the distribution p is the true states of the environment and the KL divergence is a measure of the distance between them. Crucially, free energy can also be expressed in terms of the surprise of sampled consequences:

$$F(s,u) = D_{KL}\left[q(\psi|\mu)\,||p(\psi|s,m)\right] - logp(s|m)$$

Free Energy = relative entropy+surprise

This formulation of evolutionary change may appear quite different from that of Bayesian inference as it has a focus on model quality rather than fitness. However, a sustained decrease in free energy (or increase in log model evidence) is equivalent to a decrease in model entropy and therefore contravenes the spirit, if not the letter, of the second law. The letter of that law allows a decrease in entropy for dynamic systems only if an environmental swap is conducted where low entropy inputs are exchanged for high entropy outputs. In short, the second law forbids the existence or survival of low entropy dynamic systems lacking such an ability—an ability that mandates a model of the environment and Bayesian inference under that model. This provides a focus for the model's knowledge accumulation; it must entail knowledge of its environment as well as a strategy to perform the required entropy swaps within that environment. Thus, the drive to fitness, which is explicit in the Bayesian formulation, is also implicit in the free energy formulation.

As descriptions of evolutionary processes in terms of free energy minimization have great general applicability it may be useful to consider some specific examples. In biological evolution we can associate the model (m) with a genotype. This means the genotype corresponds to the sufficient statistics of the prior beliefs a phenotype is equipped with on entering the world. Keeping in mind that organisms may sense their environments through both chemical and neural means, we may associate sensory exchanges with the environment (s) with adaptive states. Finally, the sufficient statistics of the posterior ($m\mu$) can be associated with a phenotype. In other words, the phenotype embodies probabilistic beliefs about states of its external milieu. This formulation tells us several fundamental things:

(i) everything that can change will change to minimize free energy. Here, the only things that can change are the sufficient statistics; namely, the genotype and phenotype. This means there are two optimizations in play: adaptive changes in the phenotype over somatic time (i.e., changes in $m\mu$) and adaptive changes in the genotype over evolutionary time (i.e., changes in m).

(ii) somatic changes will be subject to two forces: first, a maximization of accuracy that simply maximizes the probability of occupying adaptive states, and second, a minimization of complexity. This minimization corresponds to reducing the divergence between the beliefs about, or model of (hidden) environmental states (ψ) implicit in the phenotype and the prior beliefs implicit in the genotype. In other words, a good genotype will enable the minimization of free energy by equipping the phenotype with prior beliefs that are sufficient to maintain accuracy or a higher probability of adaptive states. Thus, the phenotype may be thought of as a type of experiment, which gathers evidence to test prior beliefs; i.e., gathers evidence for its own existence.

(iii) changes in the genotype correspond to Bayesian model selection (c.f., natural selection). This simply means selecting models or genotypes that have a low free energy or high Bayesian model evidence. Because the Bayesian model evidence is the probability of an adaptive state given a model or genotype ($p(s|m)$), natural selection's negative variational free energy becomes (free) fitness. At this level of free energy minimization, evolution is in the game of orchestrating multiple (phenotypic) experiments to optimize models of the (local) environment.

Another specific example of the general ability of the free energy minimization principle to describe evolutionary change is in neuroscience where it is fairly easy to demonstrate the centrality of this principle in explaining evolutionary, developmental and perceptual processes in a wide range of mental functions (Friston, 2010). The brain produces mental models which combine sensory information concerning the state of the environment, with possible actions with which the organism may intervene. The initiation of an action is a kind of experiment in the outside world testing the current beliefs about its hidden states. The overall drive of the free energy principle is to reduce the model complexity, while maximizing its accuracy in achieving the predicted outcome. Crucially, the ensuing self-organization can be seen at multiple levels of organization; from dendritic processes that form part of the single neuron—to entire brains. The principles are exactly the same, the only thing that changes is the way that the model is encoded (e.g., with intracellular concentrations of various substrates—or neuronal activity and connectivity in distributed brain systems). This sort of formulation has also been applied to self-organization and pattern formation when multiple systems jointly minimize their free energy (for example, in multi-agent games and morphogenesis at the cellular level).

Clearly, the application of variational (Bayesian) principles to ecological and cellular systems means we have to abandon the notion that only humans can make inferences. We will take up this theme below and see how freeing oneself from the tyranny of anthropomorphism leads us back to a universal Darwinism.

The free energy minimization principle may also be applied to processes of cultural evolution. A compelling example here is the evolution of scientific understanding itself. Science develops hypotheses or theoretical models of natural phenomena. These models are used to design experiments in the real world and the results of the experiment are used to update the probability of each hypothesis composing the model according to Bayes' theorem. In the process free energy is minimized through a balance which reduces the model's complexity (Occam's razor) while increasing the model's predictive accuracy and explanatory scope.

The evolutionary interaction between models and the systems they model, as described by the free energy minimization principle, may be applicable to additional natural phenomena beyond the examples above. Several attempts have been made to describe universal Darwinism in such terms. We have previously noted the wide range of scientific subject matter that has been identified within the literature as Darwinian processes—and have offered an interpretation in terms of inferential systems (Campbell, 2014); an interpretation closely related to that of the free energy minimization principle. Richard Dawkins offered a description of biological evolution in terms of replicators and vehicles (Dawkins, 1982), a description which Blackmore and Dennett have generalized to interpret universal Darwinism (Dennett, 1996; Blackmore, 1999). That description may also be understood as an interplay between internal models (replicators) and the experience of the "experiments" (vehicles) which they model in the external world.

The Price equation describing evolutionary change may be cast in a form which distinguishes between change due to selection and transmission. Changes due to selection tend to decrease model variation whereas changes due to transmission or copying of the model serve to increase variation. The transmission changes of biological models are often in the form of genetic mutations (Frank, 2011). From the perspective of universal Darwinism, we might expect a mechanism capable of increasing model variation within non-biological evolutionary processes that is analogous to biological mutation. As an example we might consider the process of evolutionary change in scientific models during transmission. These may appear less clear; there is less consensus on how new and sometimes improved scientific models are generated. It may seem this process has little in common with the somewhat random and undirected process of biological mutation.

The mental process by which researchers arrives at innovative models is largely hidden and might be considered closer to an art form than algorithmic but the development of inferential/Darwinian evolutionary computational processes have demonstrated a strong ability to discover innovative models in agreement with the evidence(Holland, 1975; Ibáñez et al., 2015). In some instances, these evolutionary approaches have inferred successful models for systems which have long eluded researchers (Lobo and Levin, 2015).

THE ARENA OF BAYESIAN INFERENCE

The reluctance of many researchers to endorse a Bayesian interpretation of evolutionary change may be somewhat puzzling. One reason for this is a peculiarity, and I would suggest a flaw, in the usual Bayesian interpretation of inference

that renders it unfit as a description of generalized evolutionary change. The consensus Bayesian position is that probability theory only describes inferences made by humans. As Jaynes put it (Jaynes, 1988):

> it is...the job of probability theory to describe human inferences at the level of epistemology.

Epistemology is the branch of philosophy that studies the nature and scope of knowledge. Since Plato the accepted definition of knowledge within epistemology has been "justified true beliefs" held by humans. In the Bayesian interpretation "justified" means justified by the evidence. "True belief" is the degree of belief in a given hypothesis which is justified by the evidence; it is the probability that the hypothesis is true within the terms of the model. Thus, knowledge is the probability, based on the evidence, that a given belief or model is true. I have proposed a technical definition of knowledge as 2^{-S} where S is the entropy of the model (Campbell, 2014).

A perhaps interesting interpretation of this definition is that knowledge occurs within the confines of entropy or ignorance. For example, in a model composed of a family of 64 competing hypotheses, where no evidence is available to decide amongst them, we would assign a probability of 1/64 to each hypothesis. The model has an entropy of six bits and has knowledge of $2^{-6} = 1/64$. Let's say some evidence becomes available and the model's entropy or ignorance is reduced to three bits. Then the knowledge of the updated model is 1/8, equivalent to the entropy of a model composed of only eight competing hypotheses that is maximally ignorant, which has no available evidence. The effect which evidence has on the model is to increase its knowledge by reducing the scope of its ignorance.

It is unfortunate that both Bayesian and Frequentist interpretations deny the existence of knowledge outside of the human realm because it forbids the application of Bayesian inference to phenomena other than models conceived by humans, it denies that knowledge may be accumulated in natural processes unconnected to human agency and it acts as a barrier in realizing our close relationship to the rest of nature. Thus, even though natural selection is clearly described in terms of the mathematics of Bayesian inference, neither Bayesians such as Jaynes nor frequentists such as Frank can acknowledge this fact due to another hard fact: natural selection is not dependent upon human agency. In both their views this may rule out a Bayesian interpretation.

I believe that the correct way out of this conundrum is to simply acknowledge that in many cases inference is performed by non-human agents as in the case of natural selection and that inference is an algorithm which we share with much of nature. The genome may for instance be understood as an example of a non-human conceived model involving families of competing hypotheses in the form of competing alleles within the population. Such models are capable of accumulating evidence-based knowledge in a Bayesian manner. The evidence involved is simply the proportion of traits in ancestral generations which make it into succeeding generations. In other words, we just need to broaden Jaynes' definition of probability to include non-human agency in order to view natural selection in terms of Bayesian inference.

In this view the accumulation of knowledge is a preoccupation we share with the rest of nature. It allows us to view nature as possessing some characteristics, such as surprise and expectations, previously thought by many as unique to humans or at least to animals. For instance, all organisms "expect" to find themselves in the type of environment for which they have been adapted and are "surprised" if they don't.

UNIVERSAL DARWINISM

Bayesian probability, epistemology and science in general tend to draw a false distinction between the human and non-human realms of nature. In this view the human realm is replete with knowledge and thus, infused with meaning, purpose and goals, and Bayesian inference may be used to describe its knowledge-accumulating attributes. On the other hand, the non-human realm is viewed as devoid of these attributes and thus Bayesian inference is considered inapplicable.

However, if we recognize expanded instances, such as natural selection, in which nature accumulates knowledge then we may also recognize that Bayesian inference, as well as equivalent mathematical forms, provides a suitable mathematical description in both realms. Evolutionary processes, as described by the mathematics of Bayesian inference, are those which accumulate knowledge for a specific purpose, knowledge required for increased fitness, for increased chances of continued existence. Thus, the mathematics implies purpose, meaning and goals, and provides legitimacy for Daniel Dennett's interpretation of natural selection in those terms (Dennett, 1996). If we allow an expanded scope for Bayesian inference, we may view Dennett's poetic interpretation of Darwinian processes as having support from its most powerful mathematical formulations.

An important aspect of these mathematics is that they apply not only to natural selection but also to any generalized evolutionary processes where inherited traits change in frequencies between generations. As noted in a cosmological context by Gardner and Conlon (2013):

> Specifically, Price's equation of evolutionary genetics has generalized the concept of selection acting upon any substrate and, in principle, can be used to formalize the selection of universes as readily as the selection of biological organisms.

At the core of Bayesian inference, underlying both the Price equation and the principle of free energy minimization we find an extremely simple mathematical expression: Bayes' theorem:

$$q_i^{`} = q_i \frac{w_i}{\mathrm{w}}$$

Simply put this equality says that the probabilities assigned to the hypotheses of a probabilistic model are updated by new data or experience according to a ratio, that of the probability of having the experience given that the specific hypothesis is correct to the average probability assigned by the model to having that experience. Those hypotheses supported by the data,

those that assign greater than average probability to having the actual experience, will be updated to greater values and those hypotheses not supported by the data will be updated to lesser values. This simple equation describes the accumulation of evidence-based knowledge concerning fitness.

When Bayes' theorem is used to describe an evolutionary process the ratio involved is one of relative fitness, the ratio of the fitness of a specific form of a trait to the average fitness of all forms of that trait. It is thus extremely general in describing any entity able to increase its chances of survival or to increase its adaptiveness. When cast in terms of the principle of free energy minimization some further implications of this simple equation are revealed (see above).

In a biological evolutionary context, the Price equation is traditionally understood as the mathematics of evolutionary change. However, the Price equation may be derived from a form of Bayes' theorem (Gardner, 2008; Shalizi, 2009; Frank, 2012b) which means it describes a process of Bayesian inference, a very general form of Bayesian inference which according to Gardner (Gardner, 2008) applies to any group of entities that undergo transformations in terms of a change in probabilities between generations or iterations. Even with this great generality it provides a useful model as it partitions evolutionary change in terms of selection and transmission (Frank, 2012a).

There are numerous examples of these equivalent mathematical forms used in the literature to describe evolutionary change across a wide scope of scientific subject matter, specifically evolutionary change in biology (Gardner, 2008; Frank, 2012b), neuroscience (Friston, 2010; Fernando et al., 2012) and culture (Hull, 1988; Jaynes, 2003; Mesoudi et al., 2006; Gintis, 2007).

It is interesting to speculate on the similarity of these mathematical forms to those which may be used to describe quantum physics. Quantum physics is also based upon probabilistic models which are updated by information received through interactions with other entities in the world. Wojciech Zurek, the founder of the theory of quantum Darwinism (Zurek, 2009), notes that the update of quantum states may be understood in terms of ratios acting to update probabilistic models (Zurek, 2005).

> Using this connection, we then infer probabilities of possible outcomes of measurements on S from the analogue of the Laplacian 'ratio of favorable events to the total number of equiprobable events', which we shall see in Section V is a good definition of quantum probabilities for events associated with effectively classical records kept in pointer states.

Unfortunately, many who have attempted to interpret quantum theory in terms of Bayesian inference, such as Caves, Fuchs, and Schack (Fuchs, 2010), have endorsed a common anthropomorphic Bayesian flaw and conclude that the probabilities involved with quantum phenomena are a "personal judgment" (Fuchs et al., 2015), and thus that the inferences involved take place within a human brain. A conceptual shift acknowledging that inference is a natural algorithm which may be performed in processes outside of the human brain may go some way to allowing quantum Darwinism to be understood as a process of Bayesian inference conducted at the quantum level (Campbell, 2010).

A vast array of phenomena is subject to evolutionary change and describable by the equivalent mathematical forms discussed here. These forms interpret evolutionary change as based on the accumulation of evidence-based knowledge. Conversely, many instances of evidence-based knowledge found in nature are describable using this mathematics. We might speculate that all forms of knowledge accumulation found in nature may eventually find accommodation within this paradigm. Certainly, the theorem proved by Cox (1946) identifies Bayesian inference as the unique method by which models may be updated with evidence.

It is somewhat ironic that in 1935 Fisher wrote (Fisher, 1937):

> Inductive inference is the only process known to us by which essentially new knowledge comes into the world.

Of course he was referring to experimental design and considered it unnecessary to specify that he was referring only to human knowledge. Probably he assumed that no other repositories of knowledge exist. The stage may now be set for us to understand his assertion as literally true in its full generality.

Ultimately the scope and interpretation of universal Darwinism, the study of phenomena which undergoes evolutionary change, will depend on the mathematical model underlying it. Those phenomena which are accurately and economically described by the mathematics must be judged to be within the scope of universal Darwinism. Given the great generality and substrate independence of current mathematical models, a unification of a wide range of scientific subject matters within this single paradigm may be possible.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

I appreciate discussions with Karl Friston clarifying the role of the free energy minimization principle in evolutionary change.

REFERENCES

Baez, J. C., and Pollard, B. S. (2016). Relative entropy in biological systems. *Entropy* 18, 46. doi: 10.3390/e18020046

Bayarri, M. J., and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Stat. Sci.* 19, 58–80. doi: 10.1214/088342304000000116

Blackmore, S. (1999). *The Meme Machine*. Oxford, UK: Oxford University Press.

Campbell, J. O. (2010). Quantum Darwinism as a Darwinian process. arXiv:1001.0745.
Campbell, J. O. (2014). *Darwin Does Physics.* CreateSpace Independent Publishing Platform.
Conant, R., and Ashby, R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220
Cox, R. (1946). Probability, frequency, and reasonable expectation. *Am. J. Phys.* 14, 1–13.
Darwin, C. (1872). *The Origin of Species*, *6th Edn*. New York, NY: The New American Library.
Darwin, C. (1958). *The Autobiography of Charles Darwin 1809-1882.* Edited by N. Barlow. New York, NY: W. W. Norton.
Dawkins, R. (1982). "Replicators and vehicles", in *Current Problems in Sociobiology*, ed King's College Sociobiology Group (Cambridge: Cambridge University Press), 45–64.
Dennett, D. C. (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York, NY: Simon and Schuster.
Fernando, C., Szathmary, E., and Husbands, P. (2012). Selectionist and evolutionary approaches to brain function: a critical appraisal. *Front. Comput. Neurosci.* 6:24. doi: 10.3389/fncom.2012.00024
Fisher, R. (1930). *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.
Fisher, R. A. (1937). *The Design of Experiments*, *9th Edn*. New York, NY: Macmillian.
Frank, S. A. (2011). Natural selection. III. Selection versus transmission and the levels of selection. *J. Evol. Biol.* 25, 227–243. doi: 10.1111/j.1420-9101.2011.02431.x
Frank, S. A. (2012a). Natural selection. IV. The Price equation. *J. Evol. Biol.* 25, 1002–1019. doi: 10.1111/j.1420-9101.2012.02498.x
Frank, S. A. (2012b). Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J. Evol. Biol.* 25, 2377–2396. doi: 10.1111/jeb.12010
Friston, K. (2010). The free energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787
Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475
Fuchs, C. (2010). QBism: the perimeter of quantum Bayesianism. arXiv:1003.5209.
Fuchs, C. A., Mermin, D. N., and Schack, R. (2015). Reading QBism: a reply to Nauenberg. *Am. J. Phys.* 83:198.
Galton, F. (1877). Typical laws of heredity. *Proc. R. Inst.* 8, 282–301.
Gardner, A. (2008). The price equation. *Curr. Biol.* 18, 198–202. doi: 10.1016/j.cub.2008.01.005
Gardner, A., and Conlon, J. (2013). Cosmological natural selection and the purpose of the universe. *Complexity* 18, 48–56. doi: 10.1002/cplx.21446
Gintis, H. (2007). A framework for the unification of the behavioral sciences. *Behav. Brain Sci.* 30, 1–61. doi: 10.1017/S0140525X07000581
Harper, M. (2010). The Replicator Equation as an Inference Dynamic. arXiv:0911.1763.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Cambridge, MA: MIT Press.
Hull, D. L. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science.* Chicago, IL; London: The University of Chicago Press.
Ibáñez, A., Armañanzas, R., Bielza, C., and Larrañaga, P. (2015). Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indicies. *J. Am. Soc. Inf. Sci. Tech.* doi: 10.1002/asi.23467. [Epub ahead of print].
Jaynes, E. T. (1988). "Clearing up the mysteries - the original goal," in *Maximum Entropy and Bayesian Methods,* ed J. Skilling (Dordrecht: Kluwer Academic Publishers), 1–27.
Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
Lobo, D., and Levin, M. (2015). Inferring regulatory networks from experimental morphological phenotypes: a computational method reverse-engineers planarian regeneration. *PLoS Comput. Biol.* 11:e1004295. doi: 10.1371/journal.pcbi.1004295
Mesoudi, A., Whiten, A., and Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behav. Brian Sci.* 29, 329–383.
Moran, R. J., Symmonds, M., Dolan, R. J., and Friston, K. J. (2014). The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. *PLoS Comput. Biol.* 10:e1003422. doi: 10.1371/journal.pcbi.1003422
Roweis, S., and Ghahramani, Z. (1999). A unifying view of linear Gaussian models. *Neural Comput.* 11, 305–345.
Shalizi, C. R. (2009). Dynamics of bayesian updating with dependent data and misspecified models. *Electronic J. Stat.* 3, 1039–1074. doi: 10.1214/09-EJS485
Zurek, W. H. (2005). Probabilities from entanglement, Born's rule from envariance. *Phys. Rev. A* 71:052105. doi: 10.1103/PhysRevA.71.052105
Zurek, W. H. (2009). Quantum Darwinism. *Nat. Phys.* 5, 181–188. doi: 10.1038/nphys1202

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 14 March 2017
doi: 10.3389/fnsys.2017.00010

Cinematic Operation of the Cerebral Cortex Interpreted via Critical Transitions in Self-Organized Dynamic Systems

Robert Kozma[1,2]* *and Walter J. Freeman*[3]

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA, [2]Department of Mathematical Sciences, University of Memphis, Memphis, TN, USA, [3]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA, USA

Measurements of local field potentials over the cortical surface and the scalp of animals and human subjects reveal intermittent bursts of beta and gamma oscillations. During the bursts, narrow-band metastable amplitude modulation (AM) patters emerge for a fraction of a second and ultimately dissolve to the broad-band random background activity. The burst process depends on previously learnt conditioned stimuli (CS), thus different AM patterns may emerge in response to different CS. This observation leads to our cinematic theory of cognition when perception happens in discrete steps manifested in the sequence of AM patterns. Our article summarizes findings in the past decades on experimental evidence of cinematic theory of cognition and relevant mathematical models. We treat cortices as dissipative systems that self-organize themselves near a critical level of activity that is a non-equilibrium metastable state. Criticality is arguably a key aspect of brains in their rapid adaptation, reconfiguration, high storage capacity, and sensitive response to external stimuli. Self-organized criticality (SOC) became an important concept to describe neural systems. We argue that transitions from one AM pattern to the other require the concept of phase transitions, extending beyond the dynamics described by SOC. We employ random graph theory (RGT) and percolation dynamics as fundamental mathematical approaches to model fluctuations in the cortical tissue. Our results indicate that perceptions are formed through a phase transition from a disorganized (high entropy) to a well-organized (low entropy) state, which explains the swiftness of the emergence of the perceptual experience in response to learned stimuli.

Keywords: cinematic theory of cognition, AM pattern, criticality, phase transition, Freeman K set, Hebbian assembly, graph theory, neuropercolation

OPEN ACCESS

Edited by:
Yan Mark Yufik, Virtual Structures Research, Inc., USA

Reviewed by:
Alianna JeanAnn Maren, Northwestern University, USA
Paul John Werbos, Retired, USA

***Correspondence:**
Robert Kozma
rkozma@memphis.edu

Received: *07 October 2016*
Accepted: *16 February 2017*
Published: *14 March 2017*

Citation:
Kozma R and Freeman WJ (2017) Cinematic Operation of the Cerebral Cortex Interpreted via Critical Transitions in Self-Organized Dynamic Systems. Front. Syst. Neurosci. 11:10. doi: 10.3389/fnsys.2017.00010

INTRODUCTION

It is now commonplace to regard cerebral cortex as an organ maintaining itself in a dynamic state at the edge of criticality (de Arcangelis et al., 2014; Plenz and Niebur, 2014). Criticality in mathematics and physics relates to a point of sudden transition from one state to another. In thermodynamics, the term denotes a point on the phase boundary between solid, liquid and gas phases. Near the critical point, the state of the system changes drastically with the variation of some control parameter, which behavior has been observed in the operation of the cortex

(Freeman, 2008; Fraiman and Chialvo, 2012; Freeman et al., 2012). Metastability is a related fundamental behavior employed in characterizing brain dynamics and cognition (Bressler and Kelso, 2001; Freeman and Holmes, 2005; Tognoli and Kelso, 2014). Metastability indicates a continuous interplay between phase synchrony and phase scattering in a system with many interacting components (van Straaten and Stam, 2013; Zalesky et al., 2014; Freeman, 2015).

How does the cortex maintain a critical state? Nuclear physicists use the concept of criticality to denote the threshold, at which nuclear fission reaction is maintained. The critical state of the fission chain reaction is achieved by a delicate balance between the material composition of the reactor and its geometrical properties. The criticality condition is expressed as the identity of geometrical curvature (buckling) and material curvature. Critical processes in nuclear reactors are designed in a way to satisfy strictly linear operational regimes, in order to guarantee stability of the underlying coupled reactor dynamic process (Upadhyaya et al., 1980; Kozma, 1985; March-Leuba and Rey, 1993). In brains, however, nonlinear feedback effects are of primary importance in sustaining complex cortical dynamics (Kozma and Freeman, 2001; Tagliazucchi and Chialvo, 2012). Our answer to the question on the origin of sustained critical state in brains is that mutual excitation between populations of cortical neurons maintains criticality, in combination with the refractory period that prevents exponential grow, thus stabilizes the dynamics (Freeman, 1975, 2004a).

In the past decade, neuroscientists successfully employed the concept of self-organized criticality (SOC) to neural processes (Beggs, 2008; Friston et al., 2012; Fingelkurts et al., 2013; Palva et al., 2013; Plenz and Niebur, 2014). These and many other studies point to scale-free dynamics in the cortex resembling cascades of sand piles during metastable states (Bak, 1996; Jensen, 1998; Petermann et al., 2009). SOC, however, cannot describe the existence of robust critical regions with sustained metastable dynamics, neither the rapid transitions from one metastable state to the other (Tognoli and Kelso, 2014). Bonachela et al. (2010) describe brains as "pseudo-critical" and suggest that we should "... look for more elaborate (adaptive/evolutionary) explanations, beyond simple self-organization." Reinforcement learning (RL) is crucial in producing rapid transitions from one metastable state to the other (Freeman, 1979). RL sensitizes the cortex selectively and creates spatially extended Hebbian cell assemblies (HCAs). Once HCAs are formed, they respond collectively to conditional stimuli. Stimulating any part of the assembly triggers a rapid increase in synaptic gain, leading to the explosive increase in the activity, until the activation density reaches saturation (Freeman, 2015). HCAs manifest emergent neural packets facilitating the understanding of perceptual experiences (Yufik and Friston, 2016).

Synchronized bursts of neural activity have been observed and analyzed extensively in the literature. This includes the description of spike bursts in interacting excitatory-inhibitory neural populations (see, e.g., Hindmarsh and Rose, 1984; Izhikevich, 2000; Coombes and Bressloff, 2005; Srinivasan et al., 2013). Mathematical models based on chaos theory have been proved to be useful to describe these bursts patterns (Hansel and Sompolinsky, 1992; Tsuda, 2001; Kozma, 2003). Recent breakthroughs include the comprehensive description of sharp wave ripples representing episodic memory effects (Buzsáki, 2015) and systematic analysis of spike bursts (Werbos and Davis, 2016). Our work addresses experimental and theoretical findings of transient synchronization in mesoscopic neural populations and their interpretation based on the concept of phase transitions in random graph theory (RGT) and statistical physics.

Since the early 2000s, phase transition in RGT has been employed as a useful mathematical concept to model the dynamics of the cortical tissue (Kozma et al., 2001). The random graph description of the cortex, called "*neuropercolation*," implements a hierarchy of cortical models (Kozma et al., 2005). Non-local interactions between neural populations via long axonal projections are crucial in describing cortical dynamics. There are extensive studies to model small-world effects (Watts and Strogatz, 1998) in structural and functional brain networks tuned to criticality (Bullmore and Sporns, 2009, 2012; Turova, 2012; Haimovici et al., 2013; Sporns, 2013; Alagapan et al., 2016). The level of system noise, the ratio of non-local connections corresponding to long axons, and the strength of inhibitory effects are key variables that allow controlling the transitions between opposite phases (Kozma and Puljic, 2015). In the absence of non-local connections, diffusion-like effects dominate the spatio-temporal dynamics, which fall short of producing the required rapid cortical transitions. With the help of non-local connections, we were able to generate and maintain phase transitions exhibiting rapid transitions between synchronized and desynchronized phases (Puljic and Kozma, 2008, 2010; Kozma and Puljic, 2015).

Phase transitions between disordered and ordered neural states provide key insights to understand and interpret the observed cortical space-time neurodynamics. Disordered states are characterized by random dispersion of active and inactive sites, while the emergence of metastable amplitude modulation (AM) patterns signify more ordered states. In the disorganized phase, the individual microscopic neurons are loosely coupled, which facilitates them processing sensory information individually. In the organized phase, the neurons are strongly coupled into populations producing metastable macroscopic AM patterns (Freeman, 2014). Transitions from one AM pattern to the other produce a sequence of metastable cortical states, which can be viewed as neural correlates of cognitive activity in the framework of the cinematic theory of cognition (Freeman, 2006, 2007; Kozma and Freeman, 2016). The cinematic theory of cognition is related to the concept of perception occurring in discrete epochs (Crick and Koch, 2003), and to the model of pulsating consciousness manifested via neuronal activity packages (Yufik, 2013).

This essay summarizes our decades-long experimental and theoretical studies supporting the concept of the cinematic theory of cognition. We review the theory of criticality in the cerebral cortex based on self-organized dynamics of neural populations, manifested in the form of sequential phase transitions between metastable AM patterns. In our interpretation, phase transitions are responsible for the rapid responses to sensory stimuli

observed in cognitive processing and for the emergence of our perceptual experiences according to the cinematic theory of cognition.

CONSTRUCTING THE SELF-ORGANIZED PERCEPTION CYCLE

Metastable AM Patterns Manifest the Organized Phase of Cortical Dynamics

From the variety of the available brain monitoring techniques, here we focus on recordings EEG and ECoG potentials. Intracranial experiments with electrode arrays over the cortex have been conducted in various laboratories, providing a window on the electrophysiological processes underlying brain functions (Freeman, 1975; Skarda and Freeman, 1987; Canolty et al., 2010; Panagiotides et al., 2011; Buzsáki et al., 2012). A state-of-art overview of brain imaging using EEG and ECoG monitoring techniques is given by Freeman and Quian-Quiroga (2013), including single trial experiments, high-density arrays, and spatio-temporal spectral analysis. More traditional Fourier analysis is often supplemented by Hilbert transform, which is especially beneficial in the characterization of rapidly changing, metastable activity patterns.

We illustrate the experimental results concerning the presence of highly organized metastable AM patterns and their intermittent collapse to a disorganized state using the example of rabbits, conducted in the Freeman neurophysiology laboratory at UC Berkeley (Freeman and Barrie, 2000). Rabbits were implanted with intracranial electrode arrays over their sensory cortices and trained using the well established, RL paradigm. In the experiment displayed in **Figure 1**, an ECoG array of 8 × 8 electrodes is fixed over the visual cortex of the rabbit. The measurement is 6 s long with a visual stimulus presented to the animal at time instant t = 3 s; thus there is a 3 s pre-stimulus and a 3 s post-stimulus period. **Figure 1**, upper plot, shows the 64 ECoG traces filtered in the gamma band 30–36 Hz (Davis et al., 2013). There is a base level of background activity during the 3 s expectancy state without stimulus. During the ∼1 s interval following the stimulus several gamma bursts appear. Finally, after about 1 s following the stimulus (at time instants >4 s), the activity returns to the background state. The novelty of the results lies in the development of quantitative measures to characterize the sequence of metastable states, using various pragmatic information indices (Freeman, 2004a; Davis et al., 2013).

Using Hilbert transform for each of the 64 ECoG signals, complex valued analytic signals are obtained with amplitude and phase components. The analytic amplitude represents the power of the ECoG signal, while the phase can be used to monitor synchronization effects. In **Figure 1**, lower plot, the amplitudes of the 64 analytic signals are shown. In the pre-stimulus period, the amplitudes fluctuate at a low level, indicating a sustained, disorganized background activity. There are several beats during the ∼1 s period following the stimulus, which demonstrate intermittent bursts of power in the gamma band. These bursts signify the emergence of metastable AM patterns (for details, see Freeman, 1975, 2004a, 2014).

The existence of an AM pattern indicates that the cortical dynamics is constrained to a narrow attractor basin in response to a given stimulus. This is a highly structured (organized) state with significant coordination between the 64 ECoG channels. In spite of the individual differences between the ECoG channels, they have significant commonality in their behaviors; namely, they rise, reach a maximum, and decrease in synchrony. This means that the AM pattern is largely time-invariant during the 100–200 ms of its existence, although its overall intensity varies in time. The relevance of AM patterns in defining the cognitive state of the animal has been demonstrated by using AM patterns as classification tools to discriminate between stimuli (Freeman, 1979; Kozma and Freeman, 2001). The AM patterns provide us with an observation window to monitor the cognitive process using ECoG/EEG techniques. When the input is removed, the cortical dynamics is released from its constrained state, the AM pattern disappears, and the cortex returns to the disorganized, background state.

The AM patterns do not represent the input stimuli in any practical sense; rather they correspond to the meaning of the input. They continuously change during the life of the animal through a learning process, as a result of past experiences, present state and future goals of the subject. If a new stimulus does not match a previously learnt experience, the response of the cortex

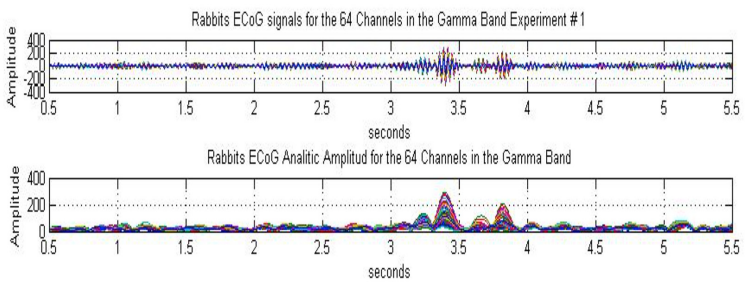


FIGURE 1 | Rabbit ECoG data measured over the visual cortex using an 8 × 8 array of electrodes. The duration of the experiment is 6 s, with a visual stimulus (light flash) presented to the animal at t = 3 s; the signals were filtered over the gamma band (30–36 Hz). The subplots show 64 curves corresponding to the ECoG signals (top) and the analytic signals (bottom), respectively. The analytic signals have been calculated using Hilbert transform, from Davis et al. (2013).

is a rapidly decaying oscillation. If the stimulus is presented again and again to the animal, the connections between excitatory neurons are strengthened in a process called Hebbian learning. As the result, the response decays less and less, which ultimately leads to sustained narrow-band oscillations due to the formation of a HCAs. The emergence of narrow-band oscillations is crucial for the efficient memory readout based on metastable AM patterns. The role of Hebbian reinforcement of connections between co-activated neurons has been demonstrated in large neuron populations, including the hippocampus, sensorimotor and speech areas (Buzsáki, 2005; Pulvermüller and Fadiga, 2010; Lopes-dos-Santos et al., 2013). In the computational domain, Hebbian RL has been implemented in various neural network models (see, e.g., Amit, 1995; Wennekers and Palm, 2009).

The example of the olfactory system with convergent-divergent connections is illustrated in **Figure 2** (Freeman, 1979). Input is transmitted via the primary olfactory nerve (PON) to the olfactory bulb, where the HCA is shown by black dots. By stimulating any subset of the HCA, the whole HCA is activated and produces narrow-band oscillations, thus exhibits the key property of generalization over the category of the sensory stimulus. Activations from the bulb are projected to the olfactory cortex through the lateral olfactory tract (LOT). The increased strengths of mutual excitatory connections (Kee) in the Hebbian assembly strongly enhance gamma oscillations in response to learned stimuli (Baird et al., 1991; Kozma and Freeman, 2001). In the context of the present work it is to be emphasized that he formation of HCAs and their rapid activation in response to learned stimuli are important conditions of cortical phase transitions (Freeman, 2015).

Background Activity and "Null Spikes"

The low overall magnitudes of the ECoG and analytic signals in **Figure 1** before the stimulus onset ($t < 3$ s) indicate that the background activity is a state of relatively low energy as compared to the high-energy burst of the AM patterns. Moreover, the energy of the background oscillations is distributed over a wide range of frequencies as opposed to the narrow-band (gamma) oscillations contributing the formation of AM patterns. In fact, the background conforms to power-law dynamics with a power exponent ranging between −2 and −4 (Freeman and Zhai, 2009). It is generated by mutual excitation among populations of cortical excitatory neurons, which activity places great demand on bodily metabolism even in brains at rest, sometimes referred to as "dark energy" (Raichle, 2006).

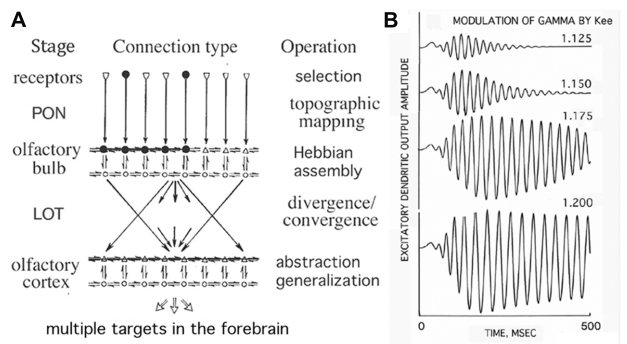


FIGURE 2 | Illustration of the topographic mapping that characterizes the olfactory input pathway. (A) Input is transmitted via the primary olfactory nerve (PON) to the olfactory bulb having excitatory (upper) and inhibitory (lower) layers, where the Hebbian assembly is shown (black dots). The Hebbian assembly is ignited with the stimulation of any of its subsets and leads to a phase transition from broad-band background activity to narrow-band oscillations, which are projected to the olfactory cortex through the lateral olfactory tract (LOT). **(B)** The increased strengths of mutual excitatory connections (Kee) in the Hebbian assembly strongly enhance gamma oscillations in response to learnt stimuli (Freeman, 1979).

The background activity is characterized by weak correlation and strong desynchronization between individual channels. The overall low background activity level may briefly drop to near zero for some channels, which phenomenon is called *"null spike"* (Freeman, 2008; Kozma and Freeman, 2008). During null spikes, the analytic phase of the background exhibits sudden changes, jumps, discontinuities; the channels have significant dispersion in their analytic phases. If the background is described as a disordered phase compared to the ordered phase with metastable AM patterns, then the null spikes clearly represent extreme disorder, which we characterize as singularity. The singularity is embedded in the background activity. At the singularity, we observe that the analytic amplitude diminishes and the analytic phase dispersion increases explosively. The very low power of the null spike means that the interactions between neural populations are suppressed. This provides favorable conditions for inputs to have a significant impact on the behavior of neural populations, especially through igniting relevant Hebbian assemblies, which facilitate a consequent rapid propagation of activities.

Null spikes are interpreted as the sites of nucleation initiating a phase transition, following the analogy of crystallization or condensation. For example, when a liquid is converted to a solid phase, the solidification starts as a specific point on the surface, and expands from that point rapidly as the liquid to solid phase transition progresses. Similarly, condensation of steam into the liquid phase starts at a point on the surface; the incipient drop grows from that location by expanding the boundary between the liquid and vapor phases. Following these examples, the initiation of null spike on the cortex may signify the start of the phase transition in the brain dynamics from disorganized phase to organized phase. In brains, the organized phase appears in the form of an emergent AM pattern with increasing power at the frequency of the carrier wave (gamma power).

The synchronized pattern emerges at the wake of a phase gradient rapidly propagating over the surface of the cortex. This phase gradient has the form of a cone and it is called *"phase cone"* (Freeman, 2004b). Note that there are many phase cones that appear and disappear all the time, however, those phase cones are mostly small (microscopic), and do not grow to the macroscopic size characteristic of a phase transition. Only when the drop of the analytic power coincides with the presence of a suitable stimulus, can we observe the rapid growth of a phase cone to sizes covering large cortical areas. The location of the apex of the cone varies randomly from each burst to the next and has no relation to the stimulus. The conic apex is in itself

a singularity, and there is some preliminary evidence that its location may correspond to the location of the preceding null spike (Freeman, 2015).

The Collapse of AM Patterns

AM patterns represent highly organized states of the cortex, which ultimately dissolve through gradual erosion under continual bombardment by sensory stimuli. The collapse of AM patterns can be viewed as a phase transitions from a synchronized to a disorganized state. In physics, such a conversion is described as evaporation of a liquid, or melting of a solid substance. This phase transition requires energy transferred to the system.

AM patterns are synchronized bursts of the activities of large masses of neurons, which emerge through phase transitions initiated by null spikes and exists for a fraction of a second (theta rates). There is a characteristic frequency of the burst in the gamma band due to the interaction of excitatory and inhibitory populations, but there is a marked distribution of frequencies of the myriads of individual feedback loops that contribute to the formation of the AM pattern (Kozma and Freeman, 2008). It is inevitable that variations in these frequencies produce oscillations that become less and less synchronized, thus the collective order of the neural populations decreases. As a result, the overall power of the oscillations diminishes and the AM patterns collapse (Freeman, 2014).

The elimination of the AM pattern drives the dynamics back to the background level, which will produce another AM pattern and the whole cycle starts again. The presence of the continual cycle of the emergence and destruction of metastable AM patterns is an important property of cortical dynamics, which is a lifelong process. In the next section, this cycle is discussed in the context of the cinematic theory of cognition, while energy considerations are described afterwards.

CINEMATIC MODEL OF PERCEPTION AS A SEQUENCE OF PHASE TRANSITIONS

ECoG measurements with intermittent transitions between synchronized and desynchronized brain states are interpreted in the framework of the cinematic theory of cognition (Freeman, 2007; Kozma and Freeman, 2016). Accordingly, neocortex processes information in frames like a cinema. Metastable AM patterns manifest the "*frames*," and the phase transitions provide the *"shutter"* from one frame to the next. Moving from one metastable pattern to the other corresponds to successive images in a movie, which we interpret using the synergetic approach to information processing (Haken, 1983). Haken proposed that state transitions are essential for information transfer between hierarchical levels, by which a collection of particles create an order parameter and in circular causality enslaves the activity of the particles. Cortical AM patterns are the manifestations of the enslavement of individual neural oscillations by collective EEG dynamics (Freeman, 2007).

Figure 3 illustrates the sequential processing in the cinematic model of cognition; the top two diagrams show the superimposed 64 ECoG signals (pass band: 20–28 Hz) and the corresponding curves of the analytic power, respectively. The time evolution displays a sequence of beats having relatively high power, separated by periods with diminishing analytic power (marked by blue vertical bars). The duration of a beat is about 100–200 ms, and a metastable AM pattern is sustained during this period. The blue bars correspond to brief time periods of transition from one beat to the other. During the transition, the AM patterns collapse to a singularity (null spike), when the synchrony disappears and the phase relationships exhibit high dispersion.

The cinematic theory employs two main components of cortical dynamics that occur sequentially, namely, the movie frame and the shutter.

- The frames are defined by the metastable AM patterns, which describe a phase with synchronous activity and macroscopic order. The metastable AM patterns represent a transmission mode of operation, i.e., they convey the knowledge contained in the meaning of the stimulus that gave rise to AM patterns. At the ordered phase, the cortex ignores the impact of the irrelevant input stimuli, until the AM pattern finally erodes and leads to the disorganized phase (shutter).
- The shutter is brief (∼20 ms) and it corresponds to the collapse of order due to the desynchronization of the neural activity. This is the receiving phase of the perception cycle, when the analytic power drops near zero and the dynamics becomes susceptible to input stimuli. Once a relevant stimulus is selected, it activates a HCA and induces rapid growth of a large phase cone, which extends over distant cortical areas.

The cinematic theory describes two types of phase transitions, one with the emergence of order from disorder in the form of AM patterns, and the other is the collapse of order manifested in the dissolution of the AM patterns.

- Transition from disorder to order: AM patterns emerge rapidly following the initiation by a null spike under the influence of a relevant stimulus. The large cones are initiated and maintained by corresponding HCAs. These large-scale phase cones enslave the cortical dynamics and lead to the emergence of order in the form of AM patterns. Without activating a HCA, the incipient phase cones cannot grow to macroscopic level, rather they remain localized, and the impact of the input stimuli rapidly fades away.
- Transitions from order to disorder: the degradation of the AM patterns is gradual, under the constant impact of input stimuli. At first, AM patterns are highly synchronized and resist to perturbations in the form of the emergence and collapse of small phase cones during the metastable state. Ultimately, however, the synchrony erodes, the power of the population activity decreases, and the dynamics returns to the disorganized background phase.

The existence of metastable AM patterns and their ultimate collapse can be interpreted in the context of SOC. There are incipient, smaller phase cones during the metastable AM patterns (Freeman, 2004b), which resemble avalanches of various sizes

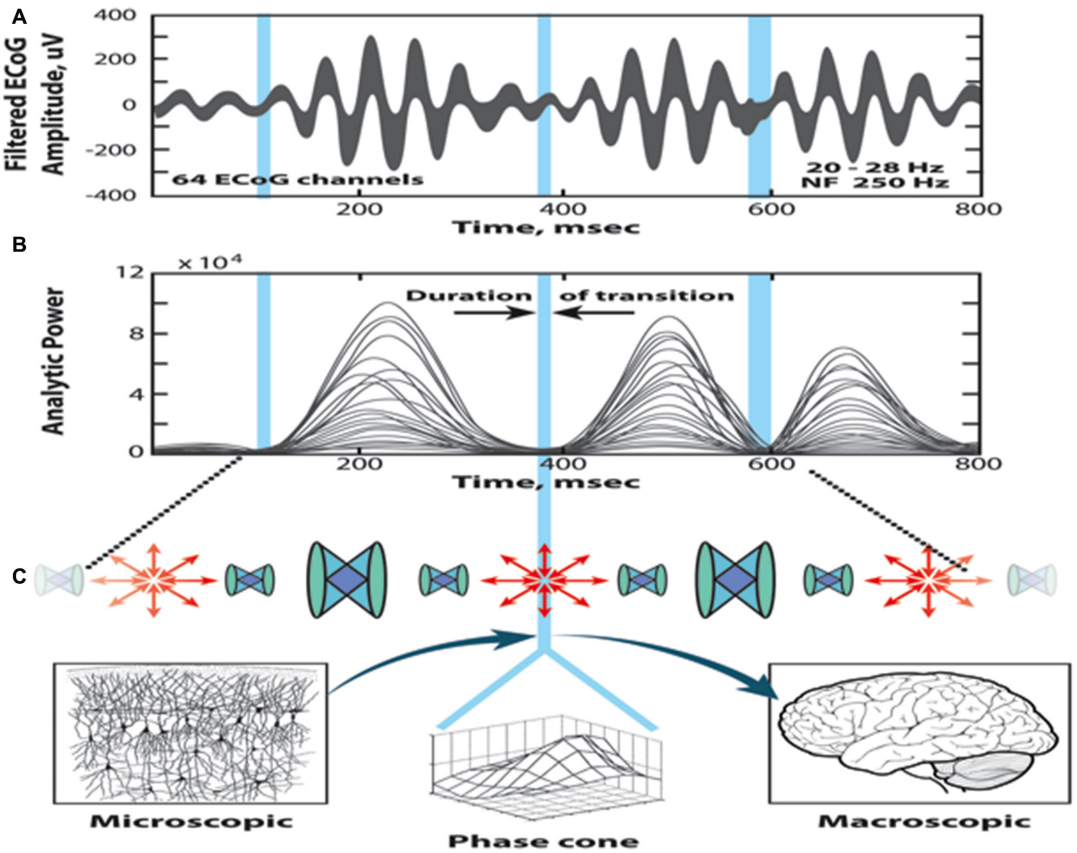


FIGURE 3 | Illustration of the self-organized perception cycle based on the cinematic theory of cognition. (A) Superimposed band-pass filtered ECoG signals. **(B)** The 64 analytic amplitudes show beats with high amplitudes, interrupted with periods of reduced power, marked by blue bars (null spikes). The high amplitudes between the blue bars correspond to metastable amplitude modulation (AM) patterns carrying the cognitive content (frames). The null spikes are singularities localized in space and time, with high dispersion of the phases (shutter). **(C)** Following the singularity, large phase cones emerge, which manifest transition from microscopic disorder to macroscopic order (illustration by Chris Gralapp), from Kozma and Freeman (2016).

that maintain the state of SOC (Bak, 1996; Jensen, 1998; Beggs and Plenz, 2003). The power law distribution of avalanche sizes suggests that the neural tissue is in the dynamic state of criticality. These incipient phase cones manifest the dissipation of energy in weak bursts. Such incipient cones may manifest the SOC metastable state, however, they are different from the large-scale phase cones emerging during the phase transitions. SOC cannot describe the sequence of transient patterns observed in the perception cycle and described here in the context of the cinematic theory of cognition. Neuropercolation is a suitable mathematical tool to describe cortical phase transitions, as summarized next (Kozma and Puljic, 2015).

DISCUSSION ON GRAPH THEORY INTERPRETATION OF CORTICAL PHASE TRANSITIONS

The perception cycle is a sequence of transitions between synchronized and desynchronized states. EEG and ECoG measurements provide a window of observation into this cycle by monitoring synchronization properties of the AM patterns. A prominent example of synchronization-desynchronization transitions in the cortex is depicted in **Figure 4**, where the analytical phase difference is shown in the vertical axis, against time and space (x and y axes). Uniformly distributed phase differences indicate synchrony across the array, while highly variable phase differences mark the presence of desynchronization. The upper segment of **Figure 4** is based on the 8×8 array of electrodes with rabbits, while the lower segment is based on intracranial measurements of the EEG of human volunteers using a linear array of 64 electrodes (Freeman, 2004b). One can see extended periods of global synchrony indicated by dominant blue colors, i.e., uniformly low values of phase differences. The periods of synchrony are interrupted by brief desynchronization events shown by a range of colors due to the large spread of the phase differences.

A family of hierarchical models of cortical dynamics has been developed originally for the olfactory system (Freeman, 1979), which is called now Freeman K (Katchalsky) sets. Freeman K sets have been applied as a general neural network model to describe chaotic dynamic memories using encoding of external data in a sequence of spatial oscillatory patterns, mimicking cortical

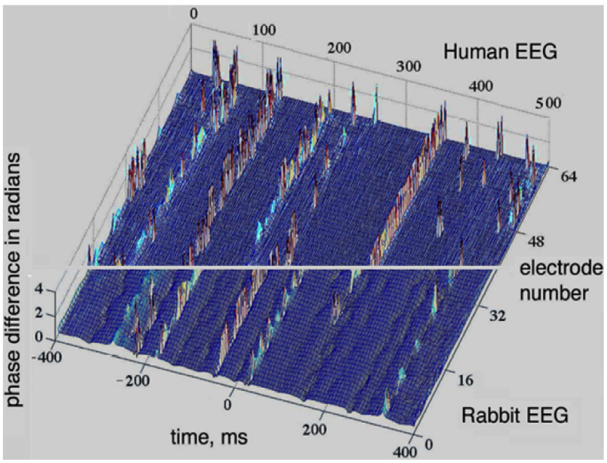


FIGURE 4 | Synchronization-desynchronization effects seen in EEG measurements with humans (lower part) and ECoG with rabbits (upper part); there are extended periods with low phase differences across space (blue color), interrupted by short periods with large phase differences (variable colors). The window of the 8 × 8 array was 5.6 × 5.6 mm for the rabbit data (upper half), while a 1 × 64 curvilinear array (189 mm long) was used over the scalp of normal human volunteers (lower half; Freeman, 2004b).

AM patterns. The original mathematical formulation of the model was based on a set of second-order ordinary differential equations (ODEs) with distributed parameters (for an overview, see Kozma and Freeman, 2001). Freeman K sets have been used in the past decades for pattern recognition, time series prediction, autonomous navigation and control, and clustering in cyber-security domains (Harter and Kozma, 2005; Kozma et al., 2007; Freeman and Kozma, 2010; Rosa and Piazentin, 2016).

An alternative implementation of Freeman K sets uses RGT instead of ODEs and it is called "neuropercolation" (Kozma et al., 2001, 2005; Kozma, 2007). Neuropercolation is based on a mathematical approach combining cellular automata on lattices and random graphs. Neuropercolation considers the interconnected network of neural populations as large-scale random graphs, which exhibit phase transitions near some well-defined critical states. Neuropercolation includes sparse rewiring of connections creating small-world effects (Watts and Strogatz, 1998), as well as the interaction of excitatory and inhibitory populations (Puljic and Kozma, 2008). It has clear advantages as compared to ODEs in characterizing rapid transients and phase transitions, due to the inherent flexibility of the graph theory framework (Kozma and Puljic, 2013, 2015; Janson et al., 2016).

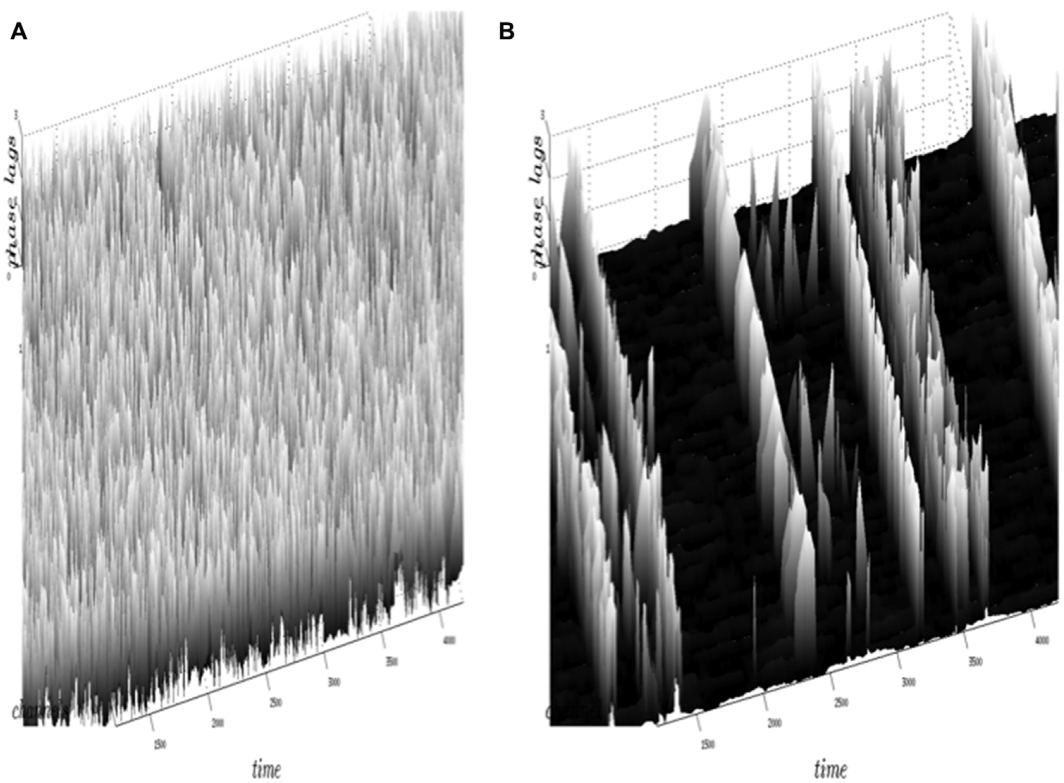


FIGURE 5 | Illustration of the results obtained by the neuropercolation model of Freeman K sets with excitatory and inhibitory neural populations. Phase lags (vertical axis) are depicted for individual channels across time (*x* axis) and space (*y* axis). In the model we use the noise level (p) as a control parameter, which allows tuning the system to criticality. The supercritical state **(A)** has highly variable phase differences without synchrony, corresponding to $p = 0.13$. Criticality is obtained in plot **(B)** with probability value $p = 0.15$, which drives the system to spontaneous, intermittent synchronization across the array (Kozma and Puljic, 2013).

Figure 5 illustrates the results obtained by the neuropercolation model using Freeman K sets with excitatory and inhibitory populations. **Figure 5A** shows the supercritical state with highly variable phase differences (no synchrony), while **Figure 5B** is an example of the near critical state with intermittent synchronization-desynchronization transitions. The criticality of the system is controlled by the overall noise level (p); $p = 0.13$ belongs to a supercritical state (no synchrony), while $p = 0.15$ results in critical state with synchronization-desynchronization transitions; from Kozma and Puljic (2013). Note that the calculated synchronization-desynchronization transitions across space and time resemble the dynamics observed in measurements with ECoG/EEG arrays. This result supports the hypothesis that the transitions between organized and disorganized phases in the cortex may be the consequence of the cortex residing in a metastable state near criticality.

CONCLUSIONS

Brains constitute only 2% of the human body but they use disproportionately high amount of energy (over 20%), which shows that creating intelligence requires a large amount of metabolic energy. Therefore, energy considerations are very important to understand the nature of biological intelligence in our brains, as well as in attempting to create artificial intelligence in machines.

The cortical energy cycle is summarized as follows, starting from a disordered background state of high entropy and low analytic amplitude. Upon the activation of a HCA by a meaningful stimulus, the synchronized activity of neural populations rapidly propagates across the cortex and creates highly structured AM patterns with low entropy states oscillating in a narrow frequency band (gamma). The formation of AM patterns can be viewed as a condensation process that leads to the dissipation of excess energy in the form of heat that is carried away in the blood stream.

The AM pattern is maintained for some time in a metastable dynamic state that seems to conform to SOC. Synchronized activity of extended neural populations is clearly documented through low phase dispersion between ECoG/EEG channels. Some disturbances in the analytic phase of the cortical tissue appear in the form of small-scale phase cones, which disappear soon after they are formed, obeying the rules of self-similar dynamics of sand piles. The energy released during the formation of the AM pattern is replenished through the metabolism, thus the oxygen debt is repaid (Freeman et al., 2012; Freeman, 2014).

The synchrony represented in the AM pattern is under constant threat by the bombardment of input stimuli and it leads to a degradation of the structure, which can be viewed as an evaporation process. Consequently, the neurons uncouple their dynamics as they are released from the binding represented by the structure. Ultimately, the AM pattern disintegrates, the overall level of firing activity decreases, and the analytic amplitude diminishes. The system returns to a chaotic background state and the cycle is completed (for a detailed description of the cycle, see Kozma and Freeman, 2016).

EEG/ECoG techniques provide insight on the perception cycle in the cortex. Synchronization-desynchronization transitions can be measured by noninvasive scalp EEG (Ruiz et al., 2010; Panagiotides et al., 2011), which allows monitoring the cognitive activity of normal subjects during routine daily activities (Freeman and Quian-Quiroga, 2013). This creates the opportunity to develop various brain-computer interfaces to improve the quality of life of the healthy human population and people with disabilities.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, as displayed in the publication.

ACKNOWLEDGMENTS

This work has been funded in part by National Science Foundation (NSF) CRCNS Program DMS-13-11165, and by Defense Sciences Office, DARPA HR0011-16-l-0006 Superior Artificial Intelligence Project. Fruitful discussions with Joshua Davis, Ray Noack, and Paul Werbos are greatly appreciated.

REFERENCES

Alagapan, S., Franca, E., Pan, L., Leondopulos, S., Wheeler, B. C., and DeMarse, T. B. (2016). Structure, function, and propagation of information across living two, four, and eight node degree topologies. *Front. Bioeng. Biotechnol.* 4:15. doi: 10.3389/fbioe.2016.00015

Amit, D. J. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behav. Brain Sci.* 18, 617–626. doi: 10.1017/s0140525x00040164

Baird, B., Freeman, W. J., Eeckman, F. H., and Yao, Y. (1991). "Applications of chaotic neurodynamics in pattern recognition," in *Applications of Artificial Neural Networks II* (Int. Society for Optics and Photonics), 12–23.

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*. New York, NY: Copernicus Press.

Beggs, J. M. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. *Philos. Trans. A Math. Phys. Eng. Sci.* 366, 329–343. doi: 10.1098/rsta.2007.2092

Beggs, J. M., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *J. Neurosci.* 23, 11167–11177.

Bonachela, J. A., de Franciscis, S., Torres, J. J., and Muñoz, M. A. (2010). Self-organization without conservation: are neuronal avalanches generically critical? *J. Stat. Mech. Theory Exp.* 2010:P02015. doi: 10.1088/1742-5468/2010/02/p02015

Bressler, S. L., and Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition. *Trends Cogn. Sci.* 5, 26–36. doi: 10.1016/s1364-6613(00)01564-3

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2618

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336–349. doi: 10.1038/nrn3214

Buzsáki, G. (2005). Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus* 15, 827–840. doi: 10.1002/hipo.20113

Buzsáki, G. (2015). Hippocampal sharp wave-ripple: a cognitive biomarker for episodic memory and planning. *Hippocampus* 25, 1073–1188. doi: 10.1002/hipo.22488

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. doi: 10.1038/nrn3241

Canolty, R. T., Ganguly, K., Kennerley, S. W., Cadieu, C. F., Koepsell, K., Wallis, J. D., et al. (2010). Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *Proc. Natl. Acad. Sci. U S A* 107, 17356–17361. doi: 10.1073/pnas.1008306107

Coombes, S., and Bressloff, P. C. (2005). *Bursting: The Genesis of Rhythm in the Nervous System.* Singapore: World Scientific.

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

Davis, J. J. J., Kozma, R., Freeman, W. J. (2013). "Neurophysiological evidence of the cognitive cycle and the emergence of awareness," in *International Joint Conference on Awareness Science and Technology & Ubi-Media Computing (iCAST-UMEDIA)* (Aizu-Wakamatsu, Japan: IEEE Press), 149–157.

de Arcangelis, L., Lombardi, F., and Herrmann, H. J. (2014). Criticality in the brain. *J. Stat. Mech. Theory Exp.* 2014:P03026. doi: 10.1088/1742-5468/2014/03/P03026

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. (2013). Consciousness as a phenomenon in the operational architectonics of brain organization: criticality and self-organization considerations. *Chaos Solitons Fractals* 55, 13–31. doi: 10.1016/j.chaos.2013.02.007

Fraiman, D., and Chialvo, D. R. (2012). What kind of noise is brain noise: anomalous scaling behavior of the resting brain activity fluctuations. *Front. Physiol.* 3:307. doi: 10.3389/fphys.2012.00307

Freeman, W. J. (1975). *Mass Action in the Nervous System.* New York, NY: Academic Press.

Freeman, W. J. (1979). Nonlinear dynamics of paleocortex manifested in the olfactory EEG. *Biol. Cybern.* 35, 21–37. doi: 10.1007/bf01845841

Freeman, W. J. (2004a). Origin, structure, and role of background EEG activity. Part 1. Analytic amplitude. *Clin. Neurophysiol.* 115, 2077–2088. doi: 10.1016/j.clinph.2004.02.029

Freeman, W. J. (2004b). Origin, structure, and role of background EEG activity. Part 2. Analytic phase. *Clin. Neurophysiol.* 115, 2089–2107. doi: 10.1016/j.clinph.2004.02.028

Freeman, W. J. (2006). A cinematographic hypothesis of cortical dynamics in perception. *Int. J. Psychophysiol.* 60, 149–161. doi: 10.1016/j.ijpsycho.2005.12.009

Freeman, W. J. (2007). "Proposed cortical 'shutter' mechanism in cinematographic perception," in *Neurodynamics of Cognition and Consciousness*, eds L. Perlovsky and R. Kozma (Heidelberg: Springer), 11–38.

Freeman, W. J. (2008). A pseudo-equilibrium thermodynamic model of information processing in nonlinear brain dynamics. *Neural Netw.* 21, 257–265. doi: 10.1016/j.neunet.2007.12.011

Freeman, W. J. (2014). "Thermodynamics of cerebral cortex assayed by measures of mass action," in *Chaos Information Processing and Paradoxical Game—The Legacy of John S. Nicolis*, eds G. Nicolis and V. Basios, (Singapore: World Scientific Publishing Co.), 275–298.

Freeman, W. J. (2015). Mechanism and significance of global coherence in scalp EEG. *Curr. Opin. Neurobiol.* 31, 199–205. doi: 10.1016/j.conb.2014.11.008

Freeman, W. J., and Barrie, J. M. (2000). Analysis of spatial patterns of phase in neocortical γ EEGs in rabbit. *J. Neurophysiol.* 84, 1266–1278.

Freeman, W. J., and Holmes, M. D. (2005). Metastability, instability, and state transition in neocortex. *Neural Netw.* 18, 497–504. doi: 10.1016/j.neunet.2005.06.014

Freeman, W. J., and Kozma, R. (2010). Freeman's mass action. *Scholarpedia* 5:8040. doi: 10.4249/scholarpedia.8040

Freeman, W. J., and Kozma, R., and Vitiello, G. (2012). "Adaptation of the generalized Carnot cycle to describe thermodynamics of cerebral cortex," in *The 2012 International Joint Conference Neural Networks (IJCNN)* (Brisbane, Australia: IEEE Press), 1–8.

Freeman, W. J., and Quian-Quiroga, R. (2013). *Imaging Brain Function with EEG: Advanced Temporal and Spatial Analysis of Electroencephalographic Signals.* New York , NY: Springer Verlag.

Freeman, W. J., and Zhai, J. (2009). Simulated power spectral density (PSD) of background electrocorticogram (ECoG). *Cogn. Neurodyn.* 3, 97–103. doi: 10.1007/s11571-008-9064-y

Friston, K., Breakspear, M., and Deco, D. (2012). Perception and self-organized instability. *Front. Comput. Neurosci.* 6:44. doi: 10.3389/fncom.2012.00044

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110:178101. doi: 10.1103/physrevlett.110.178101

Haken, H. (1983). *Synergetics: An Introduction.* Berlin: Springer-Verlag.

Hansel, D., and Sompolinsky, H. (1992). Synchronization and computation in a chaotic neural network. *Phys. Rev. Lett.* 68, 718–721. doi: 10.1103/physrevlett.68.718

Harter, D., and Kozma, R. (2005). Chaotic neurodynamics for autonomous agents. *IEEE Trans. Neural Netw.* 16, 565–579. doi: 10.1109/tnn.2005.845086

Hindmarsh, J. L., and Rose, R. M. (1984). A model of neuronal bursting using three coupled first order differential equations. *Proc. R. Soc. Lond. B Biol. Sci.* 221, 87–102. doi: 10.1098/rspb.1984.0024

Izhikevich, E. M. (2000). Neural excitability, spiking and bursting. *Int. J. Bifurcat. Chaos* 10, 1171–1266. doi: 10.1142/s0218127400000840

Janson, S., Kozma, R., Ruszinko, M., and Sokolov, Y. (2016). Bootstrap percolation on a random graph coupled with a lattice. *arXiv* 1507.07997v2.

Jensen, H. J. (1998). *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*. New York, NY: Cambridge University Press.

Kozma, R. (1985). Effect of temperature feedback on the neutron-noise field in PWRs. *Ann. Nucl. Energy* 12, 247–258. doi: 10.1016/0306-4549(85)90107-0

Kozma, R. (2003). On the constructive role of noise in stabilizing itinerant trajectories on chaotic dynamical systems. *Chaos* 11, 1078–1090. doi: 10.1063/1.1599991

Kozma, R. (2007). Neuropercolation. *Scholarpedia* 2:1360. doi: 10.4249/scholarpedia.1360

Kozma, R., Aghazarian, H., Huntsberger, T., Tunstel, E., and Freeman, W. J. (2007). Computational aspects of cognition and consciousness in intelligent devices. *IEEE Comput. Intell. Mag.* 2, 53–64. doi: 10.1109/MCI.2007.385369

Kozma, R., Balister, P., Bollobas, B., and Freeman, W. J. (2001). "Dynamical percolation models of phase transitions in the cortex," in *Proc. NOLTA 01 Nonlinear Theory and Applications Symposium* (Miyagi, Japan), (Vol. 1) 55–59.

Kozma, R., and Freeman, J. W. (2001). Chaotic resonance: methods and applications for robust classification of noisy and variable patterns. *Int. J. Bifurcat. Chaos* 10, 2307–2322. doi: 10.1142/s0218127401002870

Kozma, R., and Freeman, W. J. (2008). Intermittent spatio-temporal desynchronization and sequenced synchrony in ECoG signals. *Chaos* 18:037131. doi: 10.1063/1.2979694

Kozma, R., and Freeman, W. J. (2016). *Cognitive Phase Transitions in the Cerebral Cortex–Enhancing the Neuron Doctrine by Modeling Neural Fields.* Berlin: Springer Verlag.

Kozma, R., and Puljic, M. (2013). Hierarchical random cellular neural networks for system-level brain-like signal processing. *Neural Netw.* 45, 101–110. doi: 10.1016/j.neunet.2013.02.010

Kozma, R., and Puljic, M. (2015). Random graph theory and neuropercolation for modeling brain oscillations at criticality. *Curr. Opin. Neurobiol.* 31, 181–188. doi: 10.1016/j.conb.2014.11.005

Kozma, R., Puljic, M., Balister, P., Bollobas, B., and Freeman, W. J. (2005). Phase transitions in the neuropercolation model of neural populations with mixed local and non-local interactions. *Biol. Cybern.* 92, 367–379. doi: 10.1007/s00422-005-0565-z

Lopes-dos-Santos, V., Ribeiro, S., and Tort, A. B. L. (2013). Detecting cell assemblies in large neuronal populations. *J. Neurosci. Methods* 220, 149–166. doi: 10.1016/j.jneumeth.2013.04.010

March-Leuba, J., and Rey, J. M. (1993). Coupled thermohydraulic neutronic instabilities in boiling water nuclear reactors: a review of the state of the art. *Nucl. Eng. Des.* 145, 97–111. doi: 10.1016/0029-5493(93)90061-d

Palva, J. M., Zhigalov, A., Hirvonen, J., Korhonen, O., Linkenkaer-Hansen, K., and Palva, S. (2013). Neuronal long-range temporal correlations and avalanche dynamics are correlated with behavioral scaling laws. *Proc. Natl. Acad. Sci. U S A* 110, 3585–3590. doi: 10.1073/pnas.1216855110

Panagiotides, H., Freeman, W. J., Holmes, M. D., and Pantazis, D. (2011). Behavioral states may be associated with distinct spatial patterns in electrocorticogram (ECoG). *Cogn. Neurodyn.* 5, 55–66. doi: 10.1007/s11571-010-9139-4

Petermann, T., Thiagarajan, T. A., Lebedev, M., Nicoleli, M., Chialvo, D. R., and Plenz, D. (2009). Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. *Proc. Natl. Acad. Sci. U S A* 106, 15921–15926. doi: 10.1073/pnas.0904089106

Plenz, D., and Niebur, E. (Eds). (2014). *Criticality in Neural Systems*. Hoboken, NJ: John Wiley and Sons.

Puljic, M., and Kozma, R. (2008). Narrow-band oscillations in probabilistic cellular automata. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 78:026214. doi: 10.1103/physreve.78.026214

Puljic, M., and Kozma, R. (2010). Broad-band oscillations by probabilistic cellular automata. *J. Cell. Autom.* 5, 491–507.

Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360. doi: 10.1038/nrn2811

Raichle, M. E. (2006). The brain's dark energy. *Science* 314, 1249–1250. doi: 10.1126/science.%201134405

Rosa, J. L. G., and Piazentin, D. R. M. (2016). A new cognitive filtering approach based on freeman K3 neural networks. *Appl. Intell.* 45, 363–382. doi: 10.1007/s10489-016-0772-4

Ruiz, Y., Pockett, S., Freeman, W. J., Gonzales, E., and Li, G. (2010). A method to study global spatial patterns related to sensory perception in scalp EEG. *J. Neurosci. Methods* 191, 110–118. doi: 10.1016/j.jneumeth.2010.05.021

Skarda, C. A., and Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behav. Brain Sci.* 10, 161–195. doi: 10.1017/s0140525x00047336

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues Clin. Neurosci.* 15, 247–262.

Srinivasan, R., Thorpe, S., and Nunez, P. (2013). Top-down influences on local networks: basic theory with experimental implications. *Front. Comput. Neurosci.* 7:29. doi: 10.3389/fncom.2013.00029

Tagliazucchi, E., and Chialvo, D. R. (2012). Brain complexity born out of criticality. *arXiv Preprint* arXiv:1211.0309.

Tognoli, E., and Kelso, J. A. S. (2014). The metastable brain. *Neuron* 81, 35–48. doi: 10.1016/j.neuron.2013.12.022

Tsuda, I. (2001). Towards an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behav. Brain Sci.* 24, 793–810; discussion 810–848. doi: 10.1017/s0140525x01000097

Turova, T. S. (2012). The emergence of connectivity in neuronal networks: from bootstrap percolation to auto-associative memory. *Brain Res.* 1434, 277–284. doi: 10.1016/j.brainres.2011.07.050

Upadhyaya, B. R., Kitamura, M., and Kerlin, T. W. (1980). Multivariate signal analysis algorithms for process monitoring and parameter estimation in nuclear reactors. *Ann. Nucl. Energy* 7, 1–11. doi: 10.1016/0306-4549(80)90002-x

van Straaten, E. C. W., and Stam, C. J. (2013). Structure out of chaos: functional brain network analysis with EEG, MEG, and functional MRI. *Eur. Neuropharmacology* 23, 7–18. doi: 10.1016/j.euroneuro.2012.10.010

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Wennekers, T., and Palm, G. (2009). Syntactic sequencing in Hebbian cell assemblies. *Cogn. Neurodyn.* 3, 429–441. doi: 10.1007/s11571-009-9095-z

Werbos, P. J., and Davis, J. J. (2016). Regular cycles of forward and backward signal propagation in prefrontal cortex and in consciousness. *Front. Syst. Neurosci.* 10:97. doi: 10.3389/fnsys.2016.00097

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M., and Friston, K. (2016). Life and understanding: the origins of "understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., and Breakspear, M. (2014). Time-resolved resting-state brain networks. *Proc. Natl. Acad. Sci. U S A* 111, 10341–10346. doi: 10.1073/pnas.1400181111

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer PJW declared a past co-authorship with one of the authors RK to the handling Editor, who ensured that the process met the standards of a fair and objective review.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 15 June 2017
doi: 10.3389/fnsys.2017.00033

Neural Cross-Frequency Coupling Functions

Tomislav Stankovski[1,2†], Valentina Ticcinelli[1†], Peter V. E. McClintock[1] and Aneta Stefanovska[1]*

[1] Nonlinear and Biomedical Physics Group, Department of Physics, Lancaster University, Lancaster, United Kingdom,
[2] Faculty of Medicine, Ss Cyril and Methodius University, Skopje, Macedonia

Although neural interactions are usually characterized only by their coupling strength and directionality, there is often a need to go beyond this by establishing the functional mechanisms of the interaction. We introduce the use of dynamical Bayesian inference for estimation of the coupling functions of neural oscillations in the presence of noise. By grouping the partial functional contributions, the coupling is decomposed into its functional components and its most important characteristics—strength and form—are quantified. The method is applied to characterize the δ-to-α phase-to-phase neural coupling functions from electroencephalographic (EEG) data of the human resting state, and the differences that arise when the eyes are either open (EO) or closed (EC) are evaluated. The δ-to-α phase-to-phase coupling functions were reconstructed, quantified, compared, and followed as they evolved in time. Using phase-shuffled surrogates to test for significance, we show how the strength of the direct coupling, and the similarity and variability of the coupling functions, characterize the EO and EC states for different regions of the brain. We confirm an earlier observation that the direct coupling is stronger during EC, and we show for the first time that the coupling function is significantly less variable. Given the current understanding of the effects of e.g., aging and dementia on δ-waves, as well as the effect of cognitive and emotional tasks on α-waves, one may expect that new insights into the neural mechanisms underlying certain diseases will be obtained from studies of coupling functions. In principle, any pair of coupled oscillations could be studied in the same way as those shown here.

Keywords: coupling function, cross-frequency coupling, dynamical Bayesian inference, effective connectivity, EEG, neural oscillations, resting brain, eyes-open

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research, Inc.,
United States

Reviewed by:
Adenauer Girardi Casali,
Federal University of São Paulo, Brazil
Mehdi Adibi,
University of New South Wales,
Australia

***Correspondence:**
Aneta Stefanovska
aneta@lancaster.ac.uk

† These authors have contributed equally to this work.

Received: *02 October 2016*
Accepted: *04 May 2017*
Published: *15 June 2017*

Citation:
Stankovski T, Ticcinelli V, McClintock PVE and Stefanovska A (2017) Neural Cross-Frequency Coupling Functions. Front. Syst. Neurosci. 11:33. doi: 10.3389/fnsys.2017.00033

1. INTRODUCTION

The complexity of the human brain makes its function exceptionally challenging to analyse and understand. Its electrophysiological activity emanates from the dynamics of large-scale cell ensembles (Traub et al., 1996; Klausberger et al., 2003; Breakspear et al., 2010) which oscillate synchronously within characteristic frequency intervals. The ensembles communicate with each other to integrate their local information flows into a common brain network. Arguably, one of the most promising ways of describing communication of that kind is through cross-frequency coupling, and there has been a large number of such studies in recent years to elucidate the functional activity of the brain underlying e.g., cognition, attention, learning and working memory (Jensen and Colgin, 2007; Musizza et al., 2007; Stam et al., 2009; Axmacher et al., 2010; Belluscio et al., 2012; Jirsa and Müller, 2013; Purdon et al., 2013; van Wijk et al., 2013; Blain-Moraes et al., 2015; Sotero, 2016).

The different types of cross-frequency coupling (Jensen and Colgin, 2007; Canolty and Knight, 2010; Voytek et al., 2010; Jirsa and Müller, 2013) depend on the dynamical properties of the oscillating systems that are coupled, e.g., phase, amplitude/power and frequency. The most studied to date in brain dynamics have been the phase-to-phase (Varela et al., 2001) and phase-to-power (Canolty et al., 2006) cross-frequency couplings. The θ-γ coupling has attracted considerable attention and its neurophysiological correlates, especially those related to the working memory (Axmacher et al., 2010; Belluscio et al., 2012), have been largely understood; there are relatively fewer studies of the coupling between δ and α waves (Jirsa and Müller, 2013). These types of investigation are usually based on the statistics of the cross-frequency relationship e.g., in terms of correlation or phase-locking, or on a quantification of the coupling amplitude. Not much has yet been done, however, to assess systematically, *in vivo*, the coupling functions that describe the *functional forms* of individual cross-frequency interactions between neural oscillations.

Coupling functions describe in great detail the physical rule specifying how the interactions occur and manifest themselves. The coupling function as a whole can be described in terms of its strength and form. It is the functional form that has provided the new dimension and perspective on which we focus below. It probes directly the functional *mechanisms* of the interactions. In this way the coupling function can determine the possibility of qualitative transitions between states of the composite system e.g., routes into and out of synchronization, thus playing an active role in the possible self-organization of the systems. Decomposition of a coupling function can also facilitate a description of the functional contributions from each separate subsystem within the coupling relationship.

Recent progress directed toward the extraction and reconstruction of the coupling functions between interacting oscillatory processes has led to a diversity of applications. These include chemical interactions (Kiss et al., 2005; Miyazaki and Kinoshita, 2006; Tokuda et al., 2007), cardiorespiratory interactions (Stankovski et al., 2012; Iatsenko et al., 2013; Kralemann et al., 2013), mechanical interactions (Kralemann et al., 2008), social sciences (Ranganathan et al., 2014) and secure communications (Stankovski et al., 2014b). The study of coupling functions is a very active and expanding field of research (Stankovski et al., 2017). In this paper we evaluate coupling functions between brain waves. We focus on δ-to-α phase-to-phase interactions during eyes opened and closed and illustrate the underlying methodology. Moreover, we clearly show the difference in form of the coupling function between these two states, thereby paving the way to further applications and advancing the understanding of brain function.

2. MATERIALS AND METHODS

2.1. Wavelet Spectral Analysis

We computed the wavelet transform (WT) (Kaiser, 1994; Bračič and Stefanovska, 1998; Stefanovska et al., 1999) in order to evaluate the power content within the 0.8–40 Hz range, converting the signals $s(t)$ to the time-frequency domain:

$$WT(\omega, t) = \int_0^\infty \psi(\omega(u-t))s(u)\omega du, \tag{1}$$

where ω denotes angular frequency, t is time, and $\psi(u) = 1/(2\pi)$ $(e^{(i2\pi f_0 u)} - e^{(2\pi f_0)^2/2})e^{-u^2/2}$ (with $\int \psi(t)dt = 0$) with central frequency $f_0 = 1$. The power within each frequency interval was assessed by averaging the spectra over the corresponding frequency ranges.

2.2. Model of Phase Dynamics

Amplitude dynamics in living systems is often multidimensional, which can create complications in analysis. In contrast, the phase dynamics of a periodic process in such systems is describable in terms of a single-dimensional observable, which is usually much easier to detect and extract from data. It is well known that brain activity carries the signatures of several distinct neural oscillations that manifest themselves within characteristic frequency intervals (Buzsáki and Draguhn, 2004). The signals extracted from these intervals are periodic, enabling the underlying oscillatory processes and their interactions to be studied effectively through phase dynamics (Kuramoto, 1984), and leading to extraction of phase-to-phase cross-frequency couplings (Jensen and Colgin, 2007; Jirsa and Müller, 2013). The cross-frequency phase couplings coexist in a multivariate and multidimensional space, so we will consider a network model of N coupled phase oscillators, each described by

$$\begin{aligned}\dot{\phi}_i(t) &= \omega_i(t) + q_i(\phi_i, \phi_j, \phi_k, \ldots, \phi_N, t) + \xi_i(t) \\ &= \omega_i(t) + \sum_n q_i^{(1)}(\phi_n, t) + \sum_{nm} q_i^{(2)}(\phi_n, \phi_m, t) \\ &+ \sum_{nml} q_i^{(3)}(\phi_n, \phi_m, \phi_l, t) + \ldots + \xi_i(t),\end{aligned} \tag{2}$$

for all $l, m, n, \ldots$, where $\dot{\phi}_i(t)$ is the time derivative of the phase (i.e., the instantaneous frequency), $\omega_i(t)$ is the natural frequency and the external stochastic dynamics $\xi_i(t)$ is treated as Gaussian white noise $\langle \xi_i(t)\xi_j(\tau)\xi_k(\tau)...\rangle = \delta(t-\tau)D_{ijk...}$, where D is the matrix of noise diffusion and $D_{i,j,k..}$ gives the noise strength for the particular $i, j, k...$ element. Although we will discuss the inference of neural coupling functions from phase dynamics, the method that we will describe is in principle also applicable to their inference from amplitude dynamics (Stankovski et al., 2014b).

The coupling functions $q_i^{(\kappa)}$ describe the dynamics in terms of the phases of κ interacting oscillators. As can be seen from Equation (2), the coupling functions q_i act in such a way as to modify the natural frequency $\omega_i(t)$: in physical terms, a positive coupling coefficient will accelerate the oscillation in question (by increasing its instantaneous frequency $\dot{\phi}_i(t)$), whilst a negative coupling coefficient will decelerate it (by decreasing $\dot{\phi}_i(t)$). Thus a coupling function is able to describe in detail, within a single cycle, *how one oscillator is accelerated or decelerated as a result of the influence from the other oscillators*. This carries important implications for the interpretation of the mechanisms underlying the coupling functions, as will be discussed below. Each function

$q_i^{(\kappa)}$ is periodic, for $\kappa \geq 2$ on the κ-dimensional torus, and can be decomposed into a sum of κ-dimensional Fourier series of trigonometric functions. In practice it is assumed that the dynamics can be well-described by a finite number K of Fourier terms (Kralemann et al., 2011; Duggento et al., 2012): $\dot{\phi}_i = \sum_{k=-K}^{K} c_k^{(i)} \Phi_{i,k}(\phi_1, \phi_2, \ldots, \phi_N) + \xi_i = \sum_{k=-K}^{K} c_k^{(i)} \exp[i(k_1\phi_1 + k_2\phi_2 + \ldots + k_N\phi_N)] + \xi_i$, where $i = 1, \ldots, N$, $\Phi_{i,0} = 1$ so that $c_0^{(l)} = \omega_l$, and the rest of $\Phi_{i,k}$, scaled by $c_k^{(i)}$, are the k most important Fourier components. Such Fourier series $\Phi_{i,k}(\phi_1, \phi_2, \ldots, \phi_N)$ act as base functions for the dynamical inference method.

2.3. Dynamical Inference

Our aim is to reconstruct a dynamical model describing the interactions through the analysis of data, so that the model can then be used for extraction of the coupling functions. Our approach is based on the method of *dynamical inference*, often referred to as *dynamical modeling* or *dynamical filtering* (Kalman, 1960; Sanjeev Arulampalam et al., 2002; Friston et al., 2003; Voss et al., 2004; von Toussaint, 2011).

Note that inference of cross-frequency couplings from the statistics of the coupled signals, e.g., through correlation, (bi-)coherence and Granger causality measures (Geweke, 1982; Baccala and Sameshima, 2001; Kamiński et al., 2001), yields the *functional connectivity* but it provides no information about the mechanisms of causality. These latter methods are designed to infer statistical effects rather than dynamical mechanisms (Barrett and Barnett, 2013). In what follows, however, with the aid of dynamical inference we discuss how the mechanisms of the associated causality can be inferred from data, thus yielding an *effective connectivity* (Friston, 2011).

In particular, coupling functions represent one type of dynamical mechanism and their inference yields the effective connectivity. More specifically, the form of the coupling function defines the functional law under which some input of the interactions (i.e., the mutual influence between the oscillations) is translated into an appropriate output. This is related, not only to the quantitative parameters of the net coupling strength i.e., net information flow, but also to how this information is functionally structured to give an effective mechanism. For example, as we will see below, the interactions can be such that the *form of the coupling function varies* in time (see e.g., Section 3.3.2 and Stankovski et al., 2012). This dynamical change can cause a qualitative transition (like synchronization), irrespectively of the value and the variations of the net coupling strength. This is an example of a case where functional connectivity methods (e.g., Granger causality) will detect only the net coupling strength and not the possible reason for a qualitative transition, unlike coupling functions analysis which can do so (see below).

A number of different techniques are available for estimating a model from data, based on different procedures and theories, and resulting in slightly different properties and characteristics. They include e.g., least-squares and kernel smoothing fits (Rosenblum and Pikovsky, 2001; Kralemann et al., 2013), dynamical Bayesian inference (Smelyanskiy et al., 2005; Stankovski et al., 2012), maximum likelihood (multiple-shooting) methods (Voss et al., 2004; Tokuda et al., 2007), and dynamic causal modeling (Friston et al., 2003).

In what follows we use the dynamical Bayesian inference technique (Smelyanskiy et al., 2005; Stankovski et al., 2012). Briefly, the method applies Bayesian probability theory to the multidimensional time-series to infer the dynamical model in terms of stochastic differential equations. Assuming a normal multivariate distribution for the prior of the scale parameters, by the use of the model base functions,the method constructs a log-likelihood function which also ensures that the posterior probability is normally distributed. Evaluation of the current distribution relies on the evaluation of the previous block of data in the sequence, i.e., informative priors are used and the current prior depends on the previous posterior. For the first time window, in the absence of an earlier block, we set the initial prior to a flat (zero) distribution; which might effect the precision with which the initial coupling function is inferred for that window. To account for the time-variability of the interacting dynamics, the covariance matrix of the next prior is the convolution of the current posterior with the current diffusion matrix which describes how much the parameters can change. Further details of the method can be found in the Supplementary Material, in Smelyanskiy et al. (2005), Stankovski et al. (2012), Duggento et al. (2012), and Stankovski et al. (2014a) and in the references therein.

2.4. Coupling Quantifications and Decomposition

Using the inferred parameters we can calculate the coupling quantities and characteristics. The coupling functions $q_i(\phi_i, \phi_j, \phi_k, \ldots, \phi_N)$ acting on the oscillator from each of the i phases are evaluated on a $2\pi \times 2\pi \times \ldots \times 2\pi$ grid by selecting the relevant base functions, i.e., Fourier components scaled by the corresponding inferred coupling parameters. The coupling strength is calculated as the Euclidean norm $\|q_i\| = \langle q_i q_i \rangle^{1/2}$ of the inferred parameters for a particular coupling, and therefore carries the same unit of measure as the natural frequency (Hz). The correlation $\rho_i(q_i, q_j) = \langle \tilde{q}_i \tilde{q}_j \rangle / (\|\tilde{q}_i\| \, \|\tilde{q}_j\|)$, of two coupling functions where $\tilde{q}_i$ are the deviations from the mean, $\tilde{q}_i = q_i - \langle q_i \rangle$, gives the similarity of their forms, irrespectively of their amplitudes (Kralemann et al., 2013). Here, we propose a further extension of this index. By calculating the correlation of a coupling function q with a sequence of numerically-generated forms Q having specific shape features, taken from a bank, one can determine which of those features is dominant in q. The numerical set simulates the shape of a direct coupling from the slower oscillation to the faster, phase-shifted by an angle ϑ along the 2π axes. Thus, the numerical form Q_ϑ generating the highest ρ carries dual information: the extent of the similarity (described by ρ itself) and the corresponding phase, given by ϑ. See **Figure 1** and the animation video 1 in the Supplementary Material. A natural way of presenting this information is by plotting it on the complex plane to provide a polar representation of the similarity index $P_q = \rho_q e^{i\vartheta}$.

In neuroscience, the cross-frequency analyses reported to date have mostly focused on the *net* coupling. In contrast, coupling functions enable one to study the functional dependences of

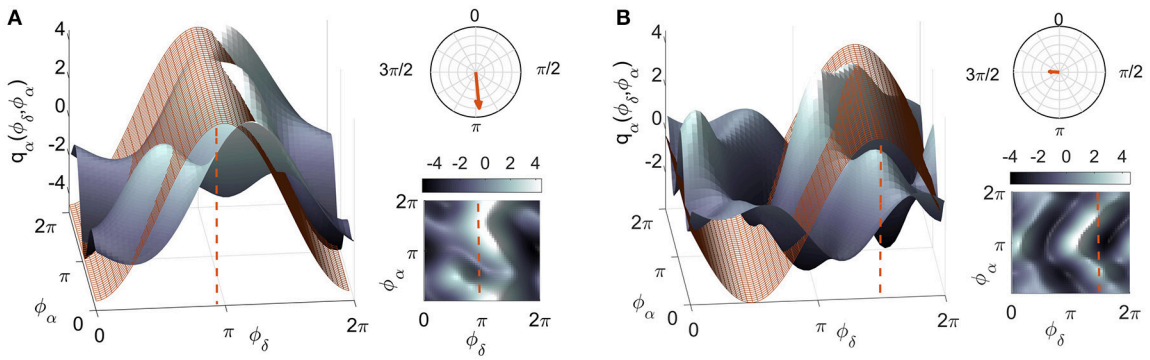


FIGURE 1 | The meaning of the polar similarity index. Two examples of coupling functions, plotted in blue, are compared with numerically-generated sinusoidal functions, plotted in red. The latter have been selected for being as similar as possible to the coupling functions: the only degree of freedom in the selection was the shift in phase (marked by the red dashed lines). The arrows in the polar planes in the top right corners have moduli equal to the similarity indices, and point to the corresponding phase values for: **(A)** a coupling function with high similarity ($\rho = 0.82$) and **(B)** one with a low value ($\rho = 0.23$). A complementary 2D color-contour plot of the coupling function is given in the bottom right-hand corner of each panel.

the distinct contributions from the individual oscillations. This procedure, referred to as *coupling decomposition*, separates a *net* pairwise coupling $q_i(\phi_i, \phi_j)$ into its *partial* self-coupling $\bar{q}_i(\phi_i)$, direct-coupling $\bar{q}_i(\phi_j)$, and common (or indirect) $\bar{q}_i(\phi_i, \phi_j)$ coupling components (Iatsenko et al., 2013; Stankovski et al., 2016). The inference of both net and partial coupling has been validated numerically (Stankovski et al., 2015). The direct-coupling $\bar{q}_i(\phi_j)$ describes the influence of the direct unidirectional driving exerted by one oscillator on the other. Arguably, this is the most observed interaction in physiology, often linked to modulation mechanisms; it dominates in a number of the coupling functions discussed below. Similarly, for a triplet coupling function $q_i(\phi_i, \phi_j, \phi_k)$ one can decompose the self, direct, and common components depending on either one or two phase variables. Additionally, one can have the direct component $\bar{q}_i(\phi_j, \phi_k)$ from two phase variables exerting a joint influence, and the common component between all three phases $\bar{q}_i(\phi_i, \phi_j, \phi_k)$. Generalization to higher κ-dimensional couplings is implicit. These couplings in a κ-dimensional network could reflect a joint functional influence from a cluster subnetwork.

2.5. EEG Recordings and Signal Processing

The multichannel EEG recordings analyzed in this work were downloaded from the Neurophysiological Biomarker Toolbox (NBT) dataset (O'Gorman et al., 2013; Poil et al., 2013). The signals were recorded for a group of 16 subjects (of which 10 were female, median age 27 years, range 21–48) in the resting state for 8 min, with a sampling frequency of 200 Hz. During the first 4 min, subjects were asked to keep their eyes open, and in the following 4 min to keep them closed. Signals from 19 EEG electrodes corresponding to the international 10–20 system were selected from the dataset for the analysis.

The cross-frequency intervals were extracted by a standard (FIR and no-phase-shift) filtering procedure. The boundaries for the conventional frequency intervals were: delta $\delta = 0.8$–4 Hz, theta $\theta = 4$–7.5 Hz, alpha $\alpha = 7.5$–14 Hz, beta $\beta = 14$–22 Hz, and gamma $\gamma = 22$–40 Hz. Special care was taken to minimize cardiac components and powerline interference (Lehnertz et al., 2014; Iatsenko et al., 2015). The phases of the filtered δ and α were estimated by use of the Hilbert transform, followed by the protophase-phase transformation (Kralemann et al., 2008).

2.6. Eyes-Open and Eyes-Closed States

The extensive changes that the simple closing of the eyes triggers in the brain caught the attention of the very first electroencephalographers (Berger, 1933). It is now known that exclusion of visual input from the central system causes the power of brain activity to increase instantaneously across all the conventional frequency ranges (Barry et al., 2007). The most striking change occurs within the α rhythm, and it has its strongest effect on the occipital part of the scalp, over the visual cortex area. It has been argued that, with eyes open, the desynchronization of α, resulting in a lower power, might occur in order to give way to a more sophisticated pattern of information processing (Klimesch, 1999).

2.7. The δ-to-α Coupling Functions

The δ-to-α interaction reflects how δ activity, associated with deep dreamless sleep (Feinberg et al., 1987), influences the α oscillations related to information processing (Pfurtscheller and Lopes da Silva, 1999). Other findings have also suggested that the δ-to-α coupling is mostly located within the frontal regions, that it is stronger during the eyes-closed resting state (Deco et al., 2010; Jirsa and Müller, 2013), and that a strong δ-to-α link exists during non-REM sleep (Bashan et al., 2012).

Cross-frequency interactions are usually mediated by the slower oscillations modulating the faster ones (Brunel and Wang, 2003; Lakatos et al., 2005; Händel and Haarmeier, 2009). In particular, task-based studies suggest that slow oscillations, which are extended across the scalp, modulate the spatial extent of the faster oscillations, which are more localized (Palva et al., 2005; Isler et al., 2008; Canolty and Knight, 2010).

In the light of this, and because of the crucial role that the α oscillation (Klimesch et al., 2007; Eidelman-Rothman et al.,

2016) plays in the eyes open (EO) and eyes closed (EC) states, we focused on the analysis of δ-to-α coupling functions. In doing so, we are able to assess, quantify, and describe in detail the functional mechanisms that define the interaction in question.

Moreover, the multichannel recordings allowed us to investigate couplings between δ and α oscillations extracted from different probes, and hence to create connectivity maps illustrating how the δ-to-α modulation differs in the EO and EC states. The coupling strength was first quantified. Note that, in earlier work (Musizza et al., 2007; Jirsa and Müller, 2013; Lehnertz et al., 2014) the use of the terms "coupling causality" and "directionality" refers to the *net* coupling strength.

2.8. Surrogate Testing

When applying non-linear analysis techniques, one should bear in mind that the linear properties of the signals, like autocorrelation or spectral features, are likely to affect the measure. To discriminate the genuine results from the ones happened by chance, one can apply surrogate testing (Theiler et al., 1992; Schreiber and Schmitz, 1996, 2000; Paluš and Hoyer, 1998; Kreuz et al., 2004). The idea behind this technique is to apply the non-linear method in question to independent time series that have the same, or as close as possible, statistical properties as the original time signals, while randomizing the expressions of the non-linear property being measured. This procedure allows one to define a threshold beneath which any result is considered spurious.

In practice, when inferring couplings even from very weakly-coupled (or completely uncoupled) systems, the methods always detect some non-zero values of apparent coupling strength. Surrogate testing can then be used to establish the "zero-level" of apparent coupling corresponding to uncoupled signals. In order not to bias the threshold with effects due to inter-subject or inter-probe variability, we applied the surrogate techniques to the same signals for which the coupling was to be measured, and we therefore define different thresholds for different subjects, pairs of probes and states.

We generated the necessary surrogates by use of the phase-shuffling (PS) method (Schreiber and Schmitz, 2000; Jirsa and Müller, 2013). This acts on the time evolution of the phase of an oscillation, wrapped between 0 and 2π, by randomizing the sequence of full phase-periods that it contains. With this technique, the linear structures of the signals are preserved but the nonlinear properties are changed. Non-stationarities appearing within each period of the oscillations are preserved. The method was applied for each subject, state, and pair of probes, thereby providing pairs of surrogate phases (δ and α). These pairs were used as input for the Bayesian inference to compute the surrogate coupling. The significance thresholds, calculated independently for each subject and combination of probes, were then set as the mean+2 standard deviations of the resultant distributions.

2.9. Statistical Analysis

The surrogate populations were tested for normality with the Shapiro-Wilk test, with the null hypothesis that the data come from a normal distribution of unknown mean and variance. The test rejected the null hypothesis at the 5% significance level in only 3% of the surrogates, and we therefore accepted the assumption of a normal distribution. Hence, we could test the coupling from the original signal by comparison with the significance threshold.

The non-parametric Wilcoxon paired test was used to determine the significance of differences between the EO and EC distributions for each frequency within the power spectra, for the averaged power within each frequency interval, for the coupling strength and for the similarity of coupling functions.

3. RESULTS

3.1. Spectral Analysis

Figure 2A shows the difference in spectral power between the EO (blue) and EC (red) states, for spectra averaged across all the probes. The shaded significance area coincides closely with the α band, indicating an increase of power in that interval for EC compared to EO. This increase was independent of scalp location. **Figure 2B** shows the statistical distributions of the averaged power within each frequency band, with a pairwise probe-by-probe statistical approach. The statistical analysis confirmed the increase of amplitude across all the frequency intervals when comparing EC with EO.

3.2. Coupling Analysis

3.2.1. Significance against Surrogate Data

Figure 3 shows the results of applying PS surrogate technique for the states of EO (in blue, **Figure 3A**) and EC (in red, **Figure 3B**). Only couplings whose δ-to-α direct-coupling strength was higher than the mean+2STD surrogate thresholds (gray shades) are indicated by dots. The average values of the surrogates and of the validated couplings (horizontal lines) are inversely proportional to the power trend, with both values for the EC being below the EO average surrogate level. For EO, however, a smaller number of probe pairs generated a coupling strength which was significant against surrogates (767 over 5,776 possible connections for EO against 1,323 over 5,776 for EC). The inter-subject variability is evident in **Figure 3**, where the different width of the x-axes portion for each subject corresponds to different number of significant connections detected.

3.2.2. Inter-Subject Variability

In order to evaluate the spatial patterns of significant coupling, the dots shown in **Figure 3** have been converted into the corresponding connections over a head-shaped map (**Figure 4**). The directionality of each connection is shown with an arrow starting from the probe where the δ oscillation was extracted, and ending on the corresponding location of the probe for the α oscillation.

The color-scale in **Figure 4** represents the number of recurrences of significant direct coupling strength among the subjects for EO (in blue, **Figure 4A**) and for EC (in red, **Figure 4B**). For clarity of visualization, arrows corresponding to less than 4 subjects for EO (for which 767 couplings were detected as non-surrogates) and 6 for EC (for which 1,323 couplings were detected as non-surrogates) are not shown. The

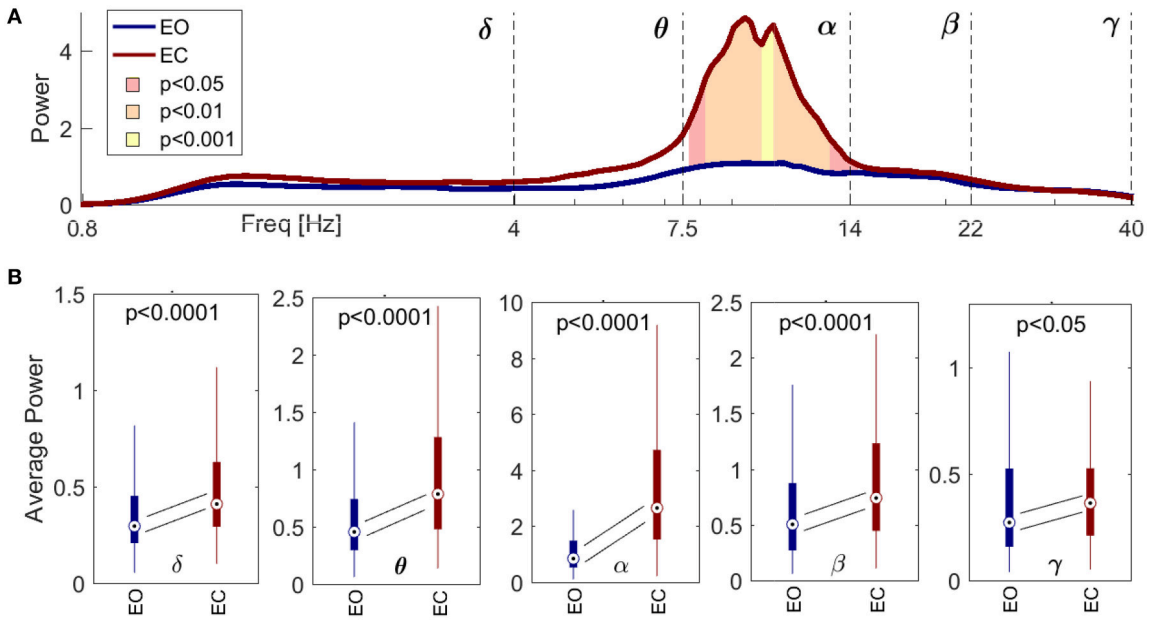


FIGURE 2 | Spectral comparison between signals recorded during the eyes open (EO, red) and eyes closed (EC, blue) conditions, for all the probes from all the subjects. **(A)** Paired statistical comparison between the inter-probe average power spectra from each subject in EO and EC, respectively. The lines show inter-subject medians, and the ranges of significance are shaded pink for $p < 0.05$, orange for $p < 0.01$ and yellow for $p < 0.001$. **(B)** Boxplots for the average power within the five frequency intervals. Diagonal lines symbolize statistical analyses pairing corresponding values for every probe and subject, and follow the changes in the medians. The p-value is indicated in each case. Note that the significance of the power in **(A)** corresponds closely to the boundaries of the α interval, and that the power in **(B)** increases significantly between EO and EC for every frequency band.

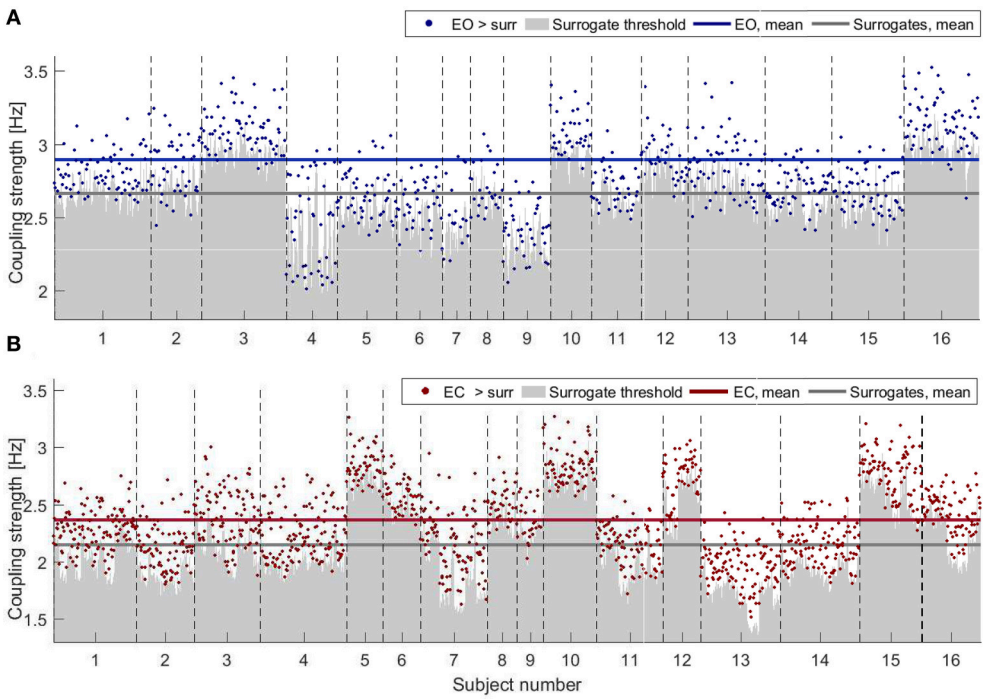


FIGURE 3 | Strengths of the couplings for **(A)** EO (blue) and **(B)** EC (red) for all the subjects, shown as consecutive intervals on the x-axes. Only values higher than the corresponding PS surrogate threshold are shown. Couplings are selected when their strengths are higher than the mean+2STD of the corresponding surrogate distribution (gray shading). Horizontal lines indicate the average values of the surrogates and of the validated couplings (color-scheme as explained above).

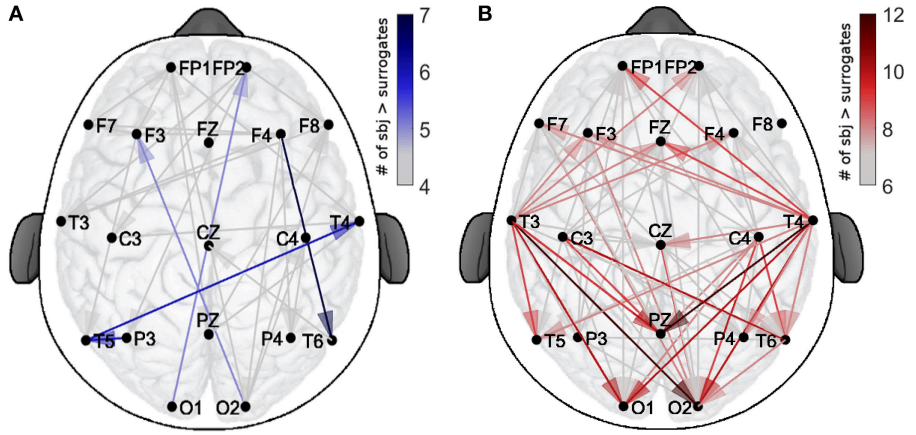


FIGURE 4 | Spatial distribution of the validated coupling strengths. The color codes indicates the number of subjects with a higher direct-coupling strength than the corresponding surrogate threshold for **(A)** EO and **(B)** EC. Note the different scalings of the two color-bars, used for clarity.

more intense colors correspond to larger numbers of subjects exhibiting significant coupling strength for a specific arrow for each state, e.g., 7 for EO and 12 for EC.

The figure shows how, for EO, two inter-hemispheric occipital-to-frontal δ-to-α couplings were exhibited by 5 subjects and one inter-hemispheric temporal long range connection, plus two intra-hemispheric, were detected in groups of 6 or 7 subjects. For EC, besides being in higher number, the significant connections were detected especially from temporal to occipital locations, and from temporal to the parietal Pz (for groups of 10–12 subjects). A clear pattern of temporal-to-frontal coupling was also detected, for smaller groups (8–9 subjects).

3.3. Coupling Functions Analysis

3.3.1. Form of the Coupling Function

To complement the coupling strength analysis, we now focus on the coupling functions themselves and discuss their unique properties. The results are summarized in **Figure 5**. The panels show the coupling functions corresponding to the links having the highest and lowest similarity indices for the intersubject average, for EO and EC. First, we describe in detail the δ-to-α coupling function as a 3D surface characterizing the EO state, as shown in **Figure 5A**. The form of this function indicates that much of the δ-to-α coupling is attributable to the direct contribution of the δ oscillation. It has a sine-like waveform along the ϕ_δ-axis, but is mostly constant along the ϕ_α-axis. This reveals the underlying functional mechanism i.e., shows that, when δ oscillations are between π and 2π, the sine-wave coupling function is higher and the δ activity accelerates the α oscillations; similarly, when the δ oscillations are between 0 and π, the coupling function is decreased and δ decelerates the α oscillations. The highest acceleration i.e., the ridge of the 3D function plot is around $3\pi/2$. The form of the coupling function of **Figure 5C** for the EC state is similar to the one for EO, but it is shifted with the highest acceleration being between 0 and π. In contrast to these two, the coupling functions shown in **Figure 5B** for EO and **Figure 5D** for EC, have uncharacteristic and undefined rippled form of lower amplitude.

These qualitative observations can be quantified and presented in terms of the polar similarity index. In **Figure 5** these are shown as a circle-map in the top-right corner of each plot. For the polar similarity index of EO (**Figure 5A**) one can note that the values for individual subjects (the dots in the circle-map), are distributed around a certain direction, and that the arrow for the average similarity index has the quite high value of 0.93. Also, the direction of the average arrow has an angle of about $3\pi/2$, which is the ridge of the average coupling function for the highest acceleration of α oscillations (compare the 3D plot in **Figure 5A**). The polar similarity index for the EC state (**Figure 5C**) shows a similar trend, with a high index of 0.91, but a different arrow direction. For the least-similar forms (**Figures 5B,D**) the similarity indices are very low with moduli close to zero (the dots are distributed sparsely), leading to almost unnoticeably small arrows at the center of the circle. Because these coupling functions come from inter-subject averages, it can be seen how the plot of polar similarity indices explains not only their morphology, but also their origin and the inter-subject variability.

3.3.2. Time-Variability of Neural Coupling Functions

Physiological systems and processes, including neural oscillations, do not exist in isolation. They can be affected by a variety of external influences making their dynamics, to a greater or lesser extent, time-varying. In such cases, one can use the dynamical Bayesian method to infer time-varying neural dynamics, as demonstrated in **Figure 6**. The coupling functions for the EC state in **Figure 6** (top), inferred at four different times, show that not only the strength but also the form of the coupling functions can vary in time. This time-variability is a representative example and it was not correlated with the coupling function time-variability of other subjects' EEG signals. It is more pronounced for the four EO coupling functions in **Figure 6** (bottom), which vary even

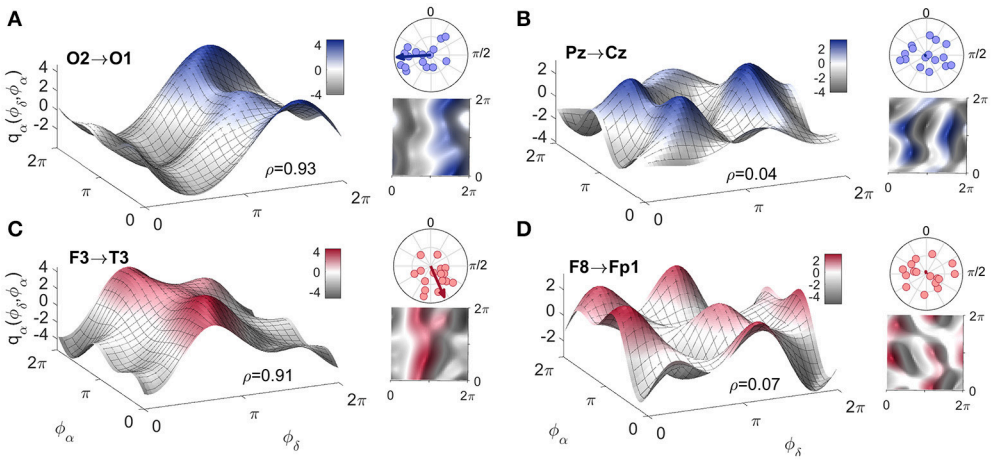


FIGURE 5 | Examples of inter-subject averages of coupling functions between particular pairs of probes. They have been selected for generating **(A,C)** the highest and **(B,D)** the lowest similarity indices, as shown. The arrows in the polar plots in the top right corners of each panel indicate the similarity indices for the averaged coupling functions, while the dots indicate the similarity indices for individuals. Note that in B and D the arrows are of negligible dimension. A complementary 2D color-contour plot of the coupling function is given in the bottom right-hand corner of each panel.

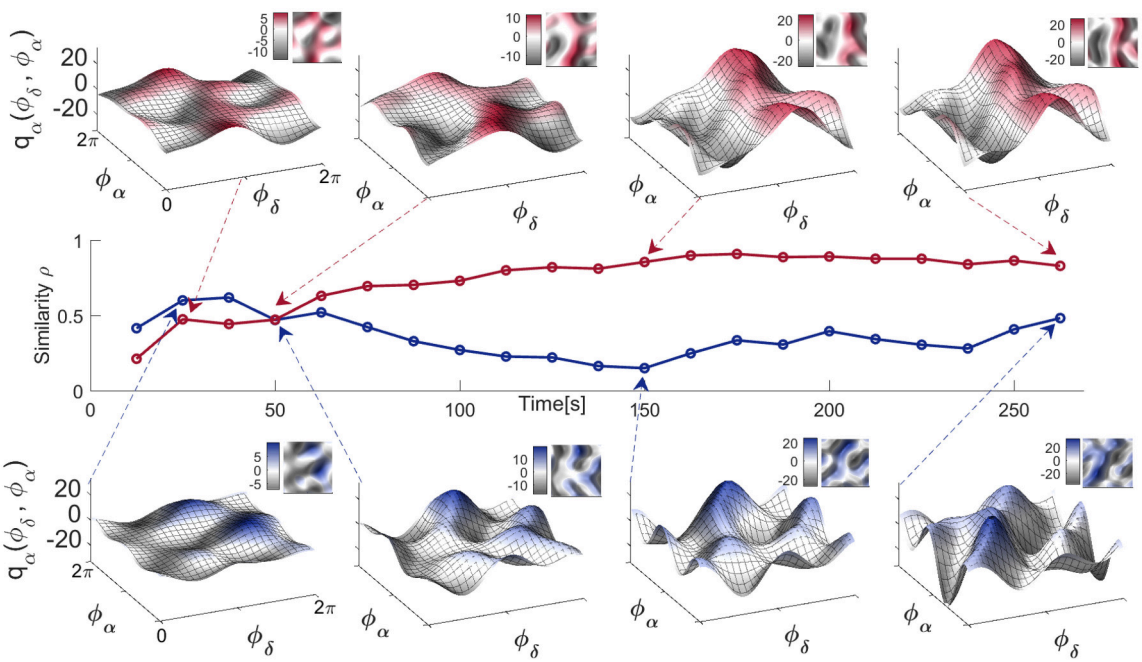


FIGURE 6 | Time-evolution of the δ-to-α coupling functions in the resting state. Middle panel: Time-evolution of the similarity index $\rho_\alpha(\delta,\alpha)$ for the EO and EC states of a single representative subject. Top panel: The δ-to-α coupling functions for EC inferred at four particular moments in time, as indicated by the arrows. Bottom panel: The δ-to-α coupling functions for EO inferred at four particular moments in time. Complementary 2D color-contour plots of the coupling functions are given in the top right-hand corner of their respective panels.

more. Consequently, the similarity index **Figure 6** (middle) which quantifies the effect is also time-varying, with higher values for the EC state resulting in more-similar forms of coupling function—compare for example the last two coupling functions in **Figure 6** (top). This time-variability and the evolution of the resting state δ-to-α coupling functions can be appreciated even better through the animation video 2 in the Supplementary Material, generated for each of the times in **Figure 6**.

3.4. Quantitative Group Analysis

To investigate the quantitative statistics of each group of subjects we calculated the average values of the significant coupling strengths, with the corresponding surrogates' value subtracted,

and the moduli of the polar similarity indices for the coupling functions of all the links for each subject. Then we compared statistically the distributions of these values for the two groups of subjects. To present the differences between the distributions visually, we use standard boxplots which refer to the descriptive statistics (median, quartiles, maximum and minimum).

The results in **Figure 7** show that there were statistically significant differences for both the coupling strengths and the similarity of coupling functions between the EO and EC states. **Figure 7A** shows that the coupling strengths detected for the EC is significantly higher than the EO. Similarly, **Figure 7B** shows that the similarity index for δ-to-α coupling functions for the EC were significantly higher than for the EO. The latter also means that there was larger variability of the coupling functions for the EO state, compared to the EC state. Overall, the similarity of coupling functions for the EO and EC states was not very high (in the interval of [0,1]), indicating that there is relatively high variability of coupling functions for both of the resting states.

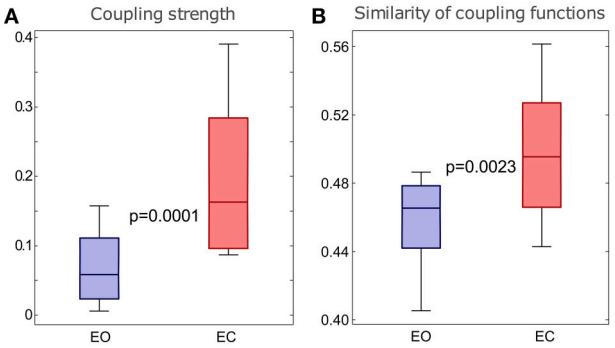


FIGURE 7 | Differences in the δ-to-α coupling strength above surrogates **(A)** and in the similarity of form of the δ-to-α coupling functions **(B)** for the two groups of subjects with EO and EC. The p-values indicated within each panel represent the statistical differences between the EO and EC states. Whiskers indicate ± 2.7 standard deviations of the distribution.

4. DISCUSSION

4.1. The EO and EC Resting States

Much has already been done, mostly through fMRI and EEG analysis, to demonstrate the existence of resting state interactions, including the formation and dissolution of resting state functional network configurations around a stable anatomical connectivity (Berger, 1933; Barry et al., 2007; Deco et al., 2010; Jirsa and Müller, 2013; O'Gorman et al., 2013). Our application of coupling functions to the resting state revealed the underlying mechanisms of interaction and has identified a number of differences between the EO and the EC states.

As there were more significant couplings in the EC than in the EO state (**Figure 3**), it is obvious that there will be more coupling links for the EC than for the EO state when presented spatially (**Figure 4**). What is interesting is that, for EC, different subjects seem to have a preferential pattern of directions, with the δ oscillation from the anterior temporal lobes (probes $T3$ and $T4$) acting as "hubs," influencing the phase of α in both the frontal and occipital directions. The occipital probes $O1$ and $O2$ are the most susceptible to the difference in α power (**Figure 2**) as they are placed over the visual area of the cortex. In EO, they act as a starting point for δ modulating long range connections toward the frontal cortex, which existed in five subjects, and then disappeared in EC. In contrast, for EC these probes receive the influence in their α rhythm from temporal and central probes.

The δ-to-α coupling functions had a specific shape, showing that the coupling is predominantly like a direct sine wave due to the δ influence, which accelerates and decelerates the α oscillations. Importantly, the form was similar for the EO and EC states (**Figure 5**), with distinctive variations and shifts along the δ oscillation. This similarity implies that the same underlying interaction mechanism exists in the EO and EC states, and that the difference between these two resting states corresponds to increasing and decreasing some of the connection strengths (or to switching them on-off).

Because we reconstructed the form of the coupling functions, we were able to observe what they look like for both individual and averaged connections and subjects. Even though we found relatively similar forms of function, we also observed a certain degree of variability, both inter-subject variability (**Figure 5**) and time variability (**Figure 6**) of the form. These should be taken into account when average values are used, for example in making multi-subject statistics.

Finally, for the comparison of the EO and EC states (Barry et al., 2007) our analysis confirmed that the spectral power of the α oscillations in EC is significantly larger than that of EO (Klimesch, 1999). It also showed that there are a larger number of real (i.e., validated by surrogate testing) δ-to-α couplings for the EC state (Jirsa and Müller, 2013), that the form of the coupling functions was similar for EO and EC, that the coupling functions were somewhat less variable for EC than for EO, and that this dominance of the EC state in the interactions was confirmed also by the quantitative boxplot statistics for the whole groups of subjects.

4.2. Methodological Aspects and Generalizations

The assessment of neural coupling functions through the phase dynamics of interacting neural oscillations enables us to study their acceleration/deceleration, i.e., timing and coordination. The generalization to amplitude coupling functions is implicit. In such cases, one should be able to determine a plausible state model in relation to the dimensionality of the signals. Amplitude neural coupling functions can reveal the mechanism through which the strength and power of one neural oscillation are affected by the influence of the other oscillations.

Earlier effective connectivity methods for the inference of neural interactions have in principle contained coupling functions within their models of the interacting dynamical systems. The question we address here, in addition to presenting an efficient Bayesian method for determination of coupling functions, is that of how to *assess* the neural coupling functions. We have shown how to unify a functional unit which can be quantified and compared with other such units, and whose

evolution can be followed in time. The key characteristic that distinguishes this assessment is the form of the neural coupling functions. A unified and effective coupling function analysis can provide insights that go far beyond just knowing that neural interactions exist.

The pairwise investigation can further be generalized to higher degrees of network complexity (Kralemann et al., 2014; Stankovski et al., 2015). One might, for example, study the coupling functions between the brain and other physiological oscillations, forming a physiological network (Musizza et al., 2007; Stefanovska, 2007; Bashan et al., 2012; Stankovski et al., 2016). The brain is a heavily connected network (Park and Friston, 2013) and coupling functions could be applied to reveal the functional mechanisms operative at different levels and sublevels of the interactions. In network topology with nodes and edges (Albert and Barabási, 2002) this would mean that, not only could the existence, strength and direction of the edge be studied, but also the underlying functional mechanism giving rise to the edge. The multivariate coupling function assessment can then be linked to hypergraphs (Karypis and Kumar, 2000; Zass and Shashua, 2008), though it was argued recently that, for larger networks ($N > 10$), there is no significant benefit from using multivariate inference of coupling (functions) and partialization (Rings and Lehnertz, 2016).

The time-varying form of the coupling functions (**Figure 6**) can be a cause of self-organization transitions, like the emergence of network clustering, or of the systems going into-and-out-of synchronization (Stefanovska et al., 2000; Varela et al., 2001), even for an invariant net coupling strength (Stankovski, 2017). More importantly, having detected and characterized a neural coupling function, one can then use this knowledge to detect, or even to predict, the onset of phase synchronization (Kiss et al., 2005). In such cases, the key feature is the known form of the coupling function which, depending on parameters like frequency, coupling strength, or polar similarity index, can predict the synchronization transition. This could have important implications for the prediction of epileptic seizures (Lehnertz and Elger, 1998; Fell et al., 2001) which occur or disappear as synchronous activity in the brain.

4.2.1. Limitations

The limitations of the method should also be borne in mind. First, the whole analysis starts with the extraction of one-dimensional vectors of phases from data which probably have a non-trivial distribution of spectral content. Especially when the coupling mode is extracted from a single signal, the filtering must be done with extreme care: spillage between different frequency intervals, as well as splitting of one mode into two intervals, will result in an artificial "common" coupling. Whenever bandpass-filtering is involved, one should exclude the possibility of investigating high-to-low frequency coupling, because any modulation of the lower frequency due to the phase of the higher one will probably be erased from the filtered mode. In any case, these couplings will usually turn out to be insignificant compared to surrogates later in the analysis.

The windowed nature of dynamical Bayesian inference carries its own limitations, too, as the length of the window is fixed for every computation. This parameter must be chosen with care, and should be adjusted so as to include a sufficient number of periods of the lower frequency involved. We found that 6–10 periods is a reasonable lower limit for this number. Due to the uninformative flat prior used for the initial window, the resultant inference of the first window should be interpreted with care. Moreover, the signals' own particular features must also be taken into account: a high degree of time-variability would need a correspondingly shorter window for the dynamical inference to follow the evolution correctly. If the method is to be generalized for use other than with a phase dynamics model, one should be careful not to infer dynamics due to non-specific, non-stationary, processes instead of genuine coupling.

4.3. Conclusion

In conclusion, coupling functions bring a novel perspective to neuroscience that is unique in that it provides access to the functional *form* of a coupling. The polar similarity index that we have introduced allows one to describe the form in quantitative detail. The comparisons of δ-to-α phase-to-phase coupling functions in the EO and EC resting states demonstrate how neural coupling functions can be reconstructed from spatially distributed sources, and what benefits and possibilities are introduced by their assessment. We have confirmed the previous result that the direct coupling is stronger during EC, and we have shown for the first time that the coupling function is significantly less variable in that state. The EO/EC states were taken as an example on which to base a discussion of methodological issues and, in so doing, to point to the wider implications and possibilities of the method itself. One may hope to gain new insights into the neuronal mechanisms underlying certain diseases from studies of coupling functions. In principle, the method can equally be applied to the time series created by *any* pair of coupled oscillatory processes.

AUTHOR CONTRIBUTIONS

TS and VT did the analysis. TS and VT wrote the draft of the paper assisted by PM. AS planned and oversaw the entire enterprize. All authors edited the text and contributed ideas and content.

FUNDING

This work was supported by the Engineering and Physical Sciences Research Council (UK) [Grant No. EP/100999X1], the EU projects BRACCIA [517133] and COSMOS [642563], and the Action Medical Research (UK) MASDA Project [GN1963]. VT is supported by a Ph.D. grant from the Department of Physics, Lancaster University.

ACKNOWLEDGMENTS

We are very grateful to Lars Michels and Simon-Shlomo Poil for their insightful comments on the results of our analyses of the eyes-open/eyes-closed data, and to them and the NBT research team for making the dataset available. Grateful thanks are also

due to Klaus Lehnertz and Andreas Daffertshofer for valuable discussions. We are also grateful to Lall Hussain for pointing out the dataset, to Christopher Orrell for performing the initial analysis, and to Bastian Pietras and Federico Devalle for their useful comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnsys.2017.00033/full#supplementary-material

REFERENCES

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74:47. doi: 10.1103/RevModPhys.74.47

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., and Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3228–3233. doi: 10.1073/pnas.0911531107

Baccala, L. A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.* 84, 463–474. doi: 10.1007/PL00007990

Barrett, A. B., and Barnett, L. (2013). Granger causality is designed to measure effect, not mechanism. *Front. Neuroinform.* 7:6. doi: 10.3389/fninf.2013.00006

Barry, R. J., Clarke, A. R., Johnstone, S. J., Magee, C. A., and Rushby, J. A. (2007). EEG differences between eyes-closed and eyes-open resting conditions. *Clin. Neurophysiol.* 118, 2765–2773. doi: 10.1016/j.clinph.2007.07.028

Bashan, A., Bartsch, R. P., Kantelhardt, J. W., Havlin, S., and Ivanov, P. C. (2012). Network physiology reveals relations between network topology and physiological function. *Nat. Commun.* 3:702. doi: 10.1038/ncomms1705

Belluscio, M. A., Mizuseki, K., Schmidt, R., Kempter, R., and Buzsáki, G. (2012). Cross-frequency phase–phase coupling between theta and gamma oscillations in the hippocampus. *J. Neurosci.* 32, 423–435. doi: 10.1523/JNEUROSCI.4122-11.2012

Berger, H. (1933). Über das elektrenkephalogramm des menschen. *Archiv Psychiat. Nerven.* 99, 555–574. doi: 10.1007/BF01814320

Blain-Moraes, S., Tarnal, V., Vanini, G., Alexander, A., Rosen, D., Shortal, B., et al. (2015). Neurophysiological correlates of sevoflurane-induced unconsciousness. *Anesthesiology* 122, 307–316. doi: 10.1097/ALN.0000000000000482

Bračič, M., and Stefanovska, A. (1998). Wavelet based analysis of human blood flow dynamics. *Bull. Math. Biol.* 60, 919–935. doi: 10.1006/bulm.1998.0047

Breakspear, M., Heitmann, S., and Daffertshofer, A. (2010). Generative models of cortical oscillations: neurobiological implications of the Kuramoto model. *Front. Hum. Neurosci.* 4:190. doi: 10.3389/fnhum.2010.00190

Brunel, N., and Wang, X.-J. (2003). What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitation-inhibition balance. *J. Neurophysiol.* 90, 415–430. doi: 10.1152/jn.01095.2002

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313, 1626–1628. doi: 10.1126/science.1128115

Canolty, R. T., and Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends Cogn. Sci.* 14, 506–515. doi: 10.1016/j.tics.2010.09.001

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2010). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961

Duggento, A., Stankovski, T., McClintock, P. V. E., and Stefanovska, A. (2012). Dynamical Bayesian inference of time-evolving interactions: from a pair of coupled oscillators to networks of oscillators. *Phys. Rev. E* 86:061126. doi: 10.1103/physreve.86.061126

Eidelman-Rothman, M., Levy, J., and Feldman, R. (2016). Alpha oscillations and their impairment in affective and post-traumatic stress disorders. *Neurosci. Biobehav. Rev.* 68, 794–815. doi: 10.1016/j.neubiorev.2016.07.005

Feinberg, I., Floyd, T. C., and March, J. D. (1987). Effects of sleep loss on delta (0.3–3 Hz) EEG and eye movement density: new observations and hypotheses. *Electroenceph. Clin. Neurophysiol.* 67, 217–221. doi: 10.1016/0013-4694(87)90019-8

Fell, J., Klaver, P., Lehnertz, K., Grunwald, T., Schaller, C., Elger, C. E., et al. (2001). Human memory formation is accompanied by rhinal–hippocampal coupling and decoupling. *Nat. Neurosci.* 4, 1259–1264. doi: 10.1038/nn759

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain. Connect.* 1, 13–36. doi: 10.1089/brain.2011.0008

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Amer. Statist. Assoc.* 77, 304–313. doi: 10.1080/01621459.1982.10477803

Händel, B., and Haarmeier, T. (2009). Cross-frequency coupling of brain oscillations indicates the success in visual motion discrimination. *Neuroimage* 45, 1040–1046. doi: 10.1016/j.neuroimage.2008.12.013

Iatsenko, D., Bernjak, A., Stankovski, T., Shiogai, Y., Owen-Lynch, P. J., Clarkson, P. B. M., et al. (2013). Evolution of cardio-respiratory interactions with age. *Philos. Trans. R. Soc. Lond. A* 371:20110622. doi: 10.1098/rsta.2011.0622

Iatsenko, D., Stefanovska, A., and McClintock, P. V. E. (2015). Nonlinear mode decomposition: a noise-robust, adaptive, decomposition method. *Phys. Rev. E.* 92:032916. doi: 10.1007/978-3-319-20016-3

Isler, J. R., Grieve, P. G., Czernochowski, D., Stark, R. I., and Friedman, D. (2008). Cross-frequency phase coupling of brain rhythms during the orienting response. *Brain Res.* 1232, 163–172. doi: 10.1016/j.brainres.2008.07.030

Jensen, O., and Colgin, L. L. (2007). Cross-frequency coupling between neuronal oscillations. *Trends Cogn. Sci.* 11, 267–269. doi: 10.1016/j.tics.2007.05.003

Jirsa, V., and Müller, V. (2013). Cross-frequency coupling in real and virtual brain networks. *Front. Comput. Neurosci.* 7:78. doi: 10.3389/fncom.2013.00078

Kaiser, G. (1994). *A Friendly Guide to Wavelets*. Boston, MA: Birkhäuser.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Fluid. Eng.* 82, 35–45. doi: 10.1115/1.3662552

Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biol. Cybern.* 85, 145–157. doi: 10.1007/s004220000235

Karypis, G., and Kumar, V. (2000). Multilevel k-way hypergraph partitioning. *VLSI Des.* 11, 285–300. doi: 10.1155/2000/19436

Kiss, I. Z., Zhai, Y., and Hudson, J. L. (2005). Predicting mutual entrainment of oscillators with experiment-based phase models. *Phys. Rev. Lett.* 94:248301. doi: 10.1103/PhysRevLett.94.248301

Klausberger, T., Magill, P. J., Márton, L. F., Roberts, J. D. B., Cobden, P. M., Buzsáki, G., et al. (2003). Brain-state-and cell-type-specific firing of hippocampal interneurons *in vivo*. *Nature* 421, 844–848. doi: 10.1038/nature01374

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain Res. Rev.* 53, 63–88. doi: 10.1016/j.brainresrev.2006.06.003

Kralemann, B., Cimponeriu, L., Rosenblum, M., Pikovsky, A., and Mrowka, R. (2008). Phase dynamics of coupled oscillators reconstructed from data. *Phys. Rev. E* 77:066205. doi: 10.1103/PhysRevE.77.066205

Kralemann, B., Frühwirth, M., Pikovsky, A., Rosenblum, M., Kenner, T., Schaefer, J., et al. (2013). *In vivo* cardiac phase response curve elucidates human respiratory heart rate variability. *Nat. Commun.* 4:2418. doi: 10.1038/ncomms3418

Kralemann, B., Pikovsky, A., and Rosenblum, M. (2011). Reconstructing phase dynamics of oscillator networks. *Chaos* 21:025104. doi: 10.1063/1.3597647

Kralemann, B., Pikovsky, A., and Rosenblum, M. (2014). Reconstructing effective phase connectivity of oscillator networks from observations. *New J. Phys.* 16:085013. doi: 10.1088/1367-2630/16/8/085013

Kreuz, T., Andrzejak, R. G., Mormann, F., Kraskov, A., Stögbauer, H., Elger, C. E., et al. (2004). Measure profile surrogates: a method to validate the performance of epileptic seizure prediction algorithms. *Phys. Rev. E* 69:061915. doi: 10.1103/PhysRevE.69.061915

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*. Berlin: Springer-Verlag.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., and Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* 94, 1904–1911. doi: 10.1152/jn.00263.2005

Lehnertz, K., Ansmann, G., Bialonski, S., Dickten, H., Geier, C., and Porz, S. (2014). Evolving networks in the human epileptic brain. *Physica D* 267, 7–15. doi: 10.1016/j.physd.2013.06.009

Lehnertz, K., and Elger, C. E. (1998). Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity. *Phys. Rev. Lett.* 80:5019. doi: 10.1103/PhysRevLett.80.5019

Miyazaki, J., and Kinoshita, S. (2006). Determination of a coupling function in multicoupled oscillators. *Phys. Rev. Lett.* 96:194101. doi: 10.1103/PhysRevLett.96.194101

Musizza, B., Stefanovska, A., McClintock, P. V. E., Paluš, M., Petrovčič, J., Ribarič, S., et al. (2007). Interactions between cardiac, respiratory, and EEG-δ oscillations in rats during anæsthesia. *J. Physiol.* (London) 580, 315–326. doi: 10.1113/jphysiol.2006.126748

O'Gorman, R. L., Poil, S.-S., Brandeis, D., Klaver, P., Bollmann, S., Ghisleni, C., et al. (2013). Coupling between resting cerebral perfusion and EEG. *Brain Topog.* 26, 442–457. doi: 10.1007/s10548-012-0265-7

Paluš, M., and Hoyer, D. (1998). Detecting nonlinearity and phase synchronization with surrogate data. *IEEE Eng. Med. Biol. Mag.* 17, 40–45. doi: 10.1109/51.731319

Palva, J. M., Palva, S., and Kaila, K. (2005). Phase synchrony among neuronal oscillations in the human cortex. *J. Neurosci.* 25, 3962–3972. doi: 10.1523/JNEUROSCI.4250-04.2005

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342, 1238411. doi: 10.1126/science.1238411

Pfurtscheller, G., and Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/S1388-2457(99)00141-8

Poil, S.-S., de Haan, W., van der Flier, W. M., Mansvelder, H. D., Scheltens, P., and Linkenkaer-Hansen, K. (2013). Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. *Front. Aging Neurosci.* 5:58. doi: 10.3389/fnagi.2013.00058

Purdon, P. L., Pierce, E. T., Mukamel, E. A., Prerau, M. J., Walsh, J. L., Wong, K. F. K., et al. (2013). Electroencephalogram signatures of loss and recovery of consciousness from propofol. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1142–E1151. doi: 10.1073/pnas.1221180110

Ranganathan, S., Spaiser, V., Mann, R. P., and Sumpter, D. J. T. (2014). Bayesian dynamical systems modelling in the social sciences. *PLoS ONE* 9:e86468. doi: 10.1371/journal.pone.0086468

Rings, T., and Lehnertz, K. (2016). Distinguishing between direct and indirect directional couplings in large oscillator networks: Partial or non-partial phase analyses? *Chaos* 26:093106. doi: 10.1063/1.4962295

Rosenblum, M. G., and Pikovsky, A. S. (2001). Detecting direction of coupling in interacting oscillators. *Phys. Rev. E.* 64:045202. doi: 10.1103/PhysRevE.64.045202

Sanjeev Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* 50, 174–188. doi: 10.1109/78.978374

Schreiber, T., and Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.* 77, 635–638. doi: 10.1103/PhysRevLett.77.635

Schreiber, T., and Schmitz, A. (2000). Surrogate time series. *Physica D* 142, 346–382. doi: 10.1016/S0167-2789(00)00043-9

Smelyanskiy, V. N., Luchinsky, D. G., Stefanovska, A., and McClintock, P. V. E. (2005). Inference of a nonlinear stochastic model of the cardiorespiratory interaction. *Phys. Rev. Lett.* 94:098101. doi: 10.1103/PhysRevLett.94.098101

Sotero, R. C. (2016). Topology, cross-frequency, and same-frequency band interactions shape the generation of phase-amplitude coupling in a neural mass model of a cortical column. *PLoS Comput. Biol.* 12:e1005180. doi: 10.1371/journal.pcbi.1005180

Stam, C. J., de Haan, W., Daffertshofer, A., Jones, B. F., Manshanden, I., van Walsum, A. M. V., et al. (2009). Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease. *Brain* 132, 213–224. doi: 10.1093/brain/awn262

Stankovski, T. (2017). Time-varying coupling functions: dynamical inference and cause of synchronization transitions. *Phys. Rev. E* 95:022206. doi: 10.1103/PhysRevE.95.022206

Stankovski, T., Duggento, A., McClintock, P. V. E., and Stefanovska, A. (2012). Inference of time-evolving coupled dynamical systems in the presence of noise. *Phys. Rev. Lett.* 109:024101. doi: 10.1103/PhysRevLett.109.024101

Stankovski, T., Duggento, A., McClintock, P. V. E., and Stefanovska, A. (2014a). A tutorial on time-evolving dynamical Bayesian inference. *Eur. Phys. J.* 223, 2685–2703. doi: 10.1140/epjst/e2014-02286-7

Stankovski, T., McClintock, P. V. E., and Stefanovska, A. (2014b). Coupling functions enable secure communications. *Phys. Rev. X* 4:011026. doi: 10.1103/PhysRevX.4.011026

Stankovski, T., Pereira, T., McClintock, P. V. E., and Stefanovska, A. (2017). Coupling functions: universal insights into dynamical interaction mechanisms. arXiv:1706.01810. Available online at: https://arxiv.org/abs/1706.01810

Stankovski, T., Petkoski, S., Raeder, J., Smith, A. F., McClintock, P. V. E., and Stefanovska, A. (2016). Alterations in the coupling functions between cortical and cardio-respiratory oscillations due to anaesthesia with propofol and sevoflurane. *Philos. Trans. R. Soc. A* 374:20150186. doi: 10.1098/rsta.2015.0186

Stankovski, T., Ticcinelli, V., McClintock, P. V. E., and Stefanovska, A. (2015). Coupling functions in networks of oscillators. *New J. Phys.* 17:035002. doi: 10.1088/1367-2630/17/3/035002

Stefanovska, A. (2007). Coupled oscillators: complex but not complicated cardiovascular and brain interactions. *IEEE Eng. Med. Bio. Magazine* 26, 25–29. doi: 10.1109/EMB.2007.907088

Stefanovska, A., Bračič, M., and Kvernmo, H. D. (1999). Wavelet analysis of oscillations in the peripheral blood circulation measured by laser Doppler technique. *IEEE Trans. Bio. Med. Eng.* 46, 1230–1239. doi: 10.1109/10.790500

Stefanovska, A., Haken, H., McClintock, P. V. E., Hožič, M., Bajrović, F., and Ribarič, S. (2000). Reversible transitions between synchronization states of the cardiorespiratory system. *Phys. Rev. Lett.* 85, 4831–4834. doi: 10.1103/PhysRevLett.85.4831

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D* 58, 77–94. doi: 10.1016/0167-2789(92)90102-S

Tokuda, I. T., Jain, S., Kiss, I. Z., and Hudson, J. L. (2007). Inferring phase equations from multivariate time series. *Phys. Rev. Lett.* 99:064101. doi: 10.1103/PhysRevLett.99.064101

Traub, R. D., Whittington, M. A., Colling, S. B., Buzsaki, G., and Jefferys, J. G. R. (1996). Analysis of gamma rhythms in the rat hippocampus *in vitro* and *in vivo*. *J. Physiol. (London)* 493, 471–484. doi: 10.1113/jphysiol.1996.sp021397

van Wijk, B. C., Litvak, V., Friston, K. J., and Daffertshofer, A. (2013). Nonlinear coupling between occipital and motor cortex during motor imagery: a dynamic causal modeling study. *NeuroImage* 71, 104–113. doi: 10.1016/j.neuroimage.2012.12.076

Varela, F., Lachaux, J.-P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

von Toussaint, U. (2011). Bayesian inference in physics. *Rev. Mod. Phys.* 83, 943–999. doi: 10.1103/RevModPhys.83.943

Voss, H. U., Timmer, J., and Kurths, J. (2004). Nonlinear dynamical system identification from uncertain and indirect measurements. *Int. J. Bifurcat. Chaos* 14, 1905–1933. doi: 10.1142/S02181274040 10345

Voytek, B., Canolty, R. T., Shestyuk, A., Crone, N. E., Parvizi, J., and Knight, R. T. (2010). Shifts in gamma phase–amplitude coupling frequency from theta to alpha over posterior cortex during visual tasks. *Front. Hum. Neurosci.* 4:191. doi: 10.3389/fnhum.2010.00191

Zass, R., and Shashua, A. (2008). "Probabilistic graph and hypergraph matching," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)* (Anchorage: IEEE), 1–8.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 28 February 2017
doi: 10.3389/fnsys.2017.00006

Brief Mental Training Reorganizes Large-Scale Brain Networks

Yi-Yuan Tang[1], Yan Tang[1], Rongxiang Tang[2] and Jarrod A. Lewis-Peacock[3]*

[1] Department of Psychological Sciences, Texas Tech University, Lubbock, TX, USA, [2] Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA, [3] Department of Psychology, University of Texas at Austin, Austin, TX, USA

Emerging evidences have shown that one form of mental training—mindfulness meditation, can improve attention, emotion regulation and cognitive performance through changing brain activity and structural connectivity. However, whether and how the short-term mindfulness meditation alters large-scale brain networks are not well understood. Here, we applied a novel data-driven technique, the multivariate pattern analysis (MVPA) to resting-state fMRI (rsfMRI) data to identify changes in brain activity patterns and assess the neural mechanisms induced by a brief mindfulness training—integrative body–mind training (IBMT), which was previously reported in our series of randomized studies. Whole brain rsfMRI was performed on an undergraduate group who received 2 weeks of IBMT with 30 min per session (5 h training in total). Classifiers were trained on measures of functional connectivity in this fMRI data, and they were able to reliably differentiate (with 72% accuracy) patterns of connectivity from before vs. after the IBMT training. After training, an increase in positive functional connections (60 connections) were detected, primarily involving bilateral superior/middle occipital gyrus, bilateral frontale operculum, bilateral superior temporal gyrus, right superior temporal pole, bilateral insula, caudate and cerebellum. These results suggest that brief mental training alters the functional connectivity of large-scale brain networks at rest that may involve a portion of the neural circuitry supporting attention, cognitive and affective processing, awareness and sensory integration and reward processing.

Keywords: integrative body–mind training (IBMT), multivariate pattern analysis (MVPA), resting-state fMRI, functional connectivity, large-scale brain networks

OPEN ACCESS

Edited by:
Yan Mark Yufik, Virtual Structures Research, Inc., USA

Reviewed by:
Yu Liu, University of Tennessee Health Science Center, USA
Zoran Josipovic, New York University, USA

****Correspondence:***
Yi-Yuan Tang
yiyuan.tang@ttu.edu

Received: *19 October 2016*
Accepted: *07 February 2017*
Published: *28 February 2017*

Citation:
Tang Y-Y, Tang Y, Tang R and Lewis-Peacock JA (2017) Brief Mental Training Reorganizes Large-Scale Brain Networks. Front. Syst. Neurosci. 11:6. doi: 10.3389/fnsys.2017.00006

INTRODUCTION

Mindfulness meditation is one form of mental training methods including several key components, such as body relaxation, breathing practice, mental imagery and mindfulness practice (Tang et al., 2015a; Acevedo et al., 2016), and has been reported to reduce stress, improve attention, emotion regulation and cognitive performance (Tang et al., 2007). The integrative body–mind training (IBMT; or simply integrative meditation) is one form of mindfulness meditation originated from ancient eastern contemplative traditions and includes techniques of body relaxation, mental imagery and mindfulness guided by an IBMT coach. Cooperation between the body and the mind is emphasized in facilitating and achieving a meditative state. The trainees concentrated on achieving a balanced state of body and mind. The method stresses no effort to control thoughts, but instead a state of restful alertness that allows a high degree of awareness of body, mind, and external instructions (Tang et al., 2007, 2010, 2012).

Our previous randomized studies have shown that short-term IBMT can improve attention, emotion regulation and cognitive performance through changing brain activity and white matter structural connectivity (Tang et al., 2007, 2009, 2010, 2012, 2013, 2015a,b). However, whether and how IBMT alters large-scale brain networks remains unknown.

The resting-state fMRI (rsfMRI) measures spontaneous neuronal activity of the brain and has been proven as an effective method for measuring large-scale functional networks in neuropsychology conditions. Therefore rsfMRI may be helpful for exploring the network alterations induced by short-term IBMT (Fox and Raichle, 2007; Tang et al., 2009, 2013).

Multivariate pattern analysis (MVPA) is a novel data-driven technique (Haynes and Rees, 2006; Norman et al., 2006; Pereira et al., 2009; Tong and Pratte, 2012; Lewis-Peacock and Norman, 2013, 2014) and has been paid increasing attention in rsfMRI analysis (De Martino et al., 2008; Haxby, 2012). MVPA has been applied in cognitive processing, brain aging, and mental disorders such as depression, antisocial personality disorder, attention-deficit disorder and schizophrenia (Dosenbach et al., 2010; Shen et al., 2010; Lewis-Peacock et al., 2012; Zeng et al., 2012). Studies suggested that MVPA could potentially detect spatially distributed information to further highlight the neural mechanisms underlying the behavioral symptoms (Zeng et al., 2012). Furthermore, MVPA based on whole-brain rsfMRI data can complement seed-based analyses. The whole-brain functional connectivity, unlike those analyzing several predefined regions or networks of interest, can ensure the optimal use of the wealth of information present in the brain imaging data (Zeng et al., 2012).

Hence, by using MVPA, our study employed whole-brain rsfMRI data to investigate the significant training-induced brain pattern changes in an undergraduate group who received 2 weeks of IBMT with 30 min per session for 10 sessions (5 h training in total). We hypothesize that the altered functional connections will be observed in the large-scale whole-brain resting-state networks including areas associated with attention, cognitive and emotional processing, awareness and sensory integration, and reward processing (Tang et al., 2007, 2009, 2010, 2012, 2013, 2015a,b; Acevedo et al., 2016). This exploration will be helpful in further discovering the neural mechanisms underlying the altered brain states, and may offer additional information for advancing our understanding of meditation training.

MATERIALS AND METHODS

Participants

Twenty-five (13 males, 21 ± 1.6 years old) healthy undergraduates at Dalian University of Technology (DUT) without any meditation experience were recruited and completed 2 weeks of IBMT training with 30 min per session for 10 sessions (5 h training in total). This study was carried out in accordance with the recommendations of DUT Institutional Review committee. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the DUT Institutional Review committee.

Data Acquisition

Imaging data collection was performed with a Philips-Achieva 3T scanner (Eindhoven, Netherlands) at Dalian Municipal Central Hospital. During the experiments, the subjects were instructed to relax, and lie still with eyes focused on a central white cross on a black screen during the resting scan. Foam pads with a standard birdcage head coil were used to fix the subject's head (Tang et al., 2013). Functional images were acquired using a gradient-echo EPI sequence (TR = 2000 ms, TE = 30 ms, flip angle = 80°). Whole-brain volumes were acquired with 36 contiguous 4-mm-thick transverse slices without gap. Functional resting-state session lasted 6 min and 10 s, and 180 volumes were obtained. For each subject, we collected the data before and after training.

Preprocessing

All resting-state images were pre-processed using the SPM8 package (Wellcome Trust Center for Neuroimaging, University College London, London, UK[1]) and Data Processing Assistant for Resting-State fMRI (DPARSF)[2]. For each subject, the first five volumes of the scanned data were discarded due to magnetic saturation. The remaining volumes were corrected for within-scan acquisition time differences between slices, and realigned to the first volume to correct for inter-scan head motions. All subjects in this study had less than 1.5 mm translation in the *x*, *y*, or *z*-axes and less than 1.5° of rotation in each axis. Next, the volumes were normalized to a standard echo planar imaging template in the Montreal Neurological Institute (MNI) space. Then, smoothing and filtering were performed using a Gaussian filter of 8 mm full-width half-maximum kernel and a Chebyshev band-pass filter (0.01–0.08 Hz) respectively. Considering several potential sources of physiological noise in the functional connectivity analysis, nuisance covariates including head motion parameters, global mean signals, white matter signals and cerebrospinal fluid signals were regressed out from the image (Dosenbach et al., 2010).

The processed images were divided into 116 regions according to the automated anatomical labeling (AAL) atlas (Schmahmann et al., 1999). Regional mean time series were obtained for each subject by averaging the fMRI time series over all the voxels in each of the 116 regions (Shen et al., 2010). Pearson's correlation coefficients were used to evaluate functional connectivity between each pair of regions and we obtained a resting-state functional network that was expressed as a 116 × 116 symmetrical matrix for each subject. By removing the 116 diagonal elements, the 6670 upper triangular elements of the functional connectivity matrix were normalized using Fisher's *z*-transform, and were then used as the features in the subsequent MVPA.

[1]www.fil.ion.ucl.ac.uk/spm
[2]http://www.restfmri.net

Features with High Discriminative Power

Reducing the number of features in a pattern classification problem can diminish noise, reduce overfitting and accelerate computation. In our analysis, feature selection reconstructs the feature space for classification by retaining the most discriminating functional connections and eliminating the rest. The discriminative power of a feature can be quantitatively measured by its relevance to classification (Guyon and Elisseeff, 2003). Therefore, the highly discriminating functional connections principally represented the alterative resting-state functional connectivity patterns. We can use these connections, rather than the full set of 6670 functional connections, to classify different brain states in the rsfMRI data before vs. after IBMT training.

In this study, we used the Kendall tau rank correlation coefficient (Kendall and Gibbons, 1990; Shen et al., 2010; Zeng et al., 2012), which provides a distribution-free test of independence between two variables to measure the relevance of each feature for classification. Suppose that there are n samples in the subjects after 2 weeks of IBMT. Let x_{ij} denotes the functional connectivity feature i of the jth sample and y_j denotes the class label of this sample (+1 for "post-training" and −1 for "pre-training"). The Kendall tau correlation coefficient of the functional connectivity feature i can be defined as:

$$\tau_i = \frac{n_c - n_d}{n^2} \tag{1}$$

Where n_c and n_d are the number of concordant and discordant pairs, respectively. Because we do not consider the relationship of two samples, the total number of sample pairs is n^2. For a pair of observation datasets $\{x_{ij}y_j\}$ and $\{x_{ik}y_k\}$, it is a concordant pair when

$$sgn(x_{ij} - x_{ik}) = sgn(y_j - y_k) \tag{2}$$

Correspondingly, it is a discordant pair when

$$sgn(x_{ij} - x_{ik}) = -sgn(y_j - y_k) \tag{3}$$

Thus, a positive correlation coefficient τ_i represents the ith functional connectivity feature that exhibits a significant increase after IBMT training, while a negative correlation coefficient τ_i represents the ith functional connectivity feature that exhibits a significant decrease after training. We defined the "discriminative power" of a given feature as the absolute value of its Kendall tau correlation coefficient. When the absolute value of τ_i was larger, the discriminative power was stronger. We ranked every τ_i according to its discriminative power and then selected those features with scores above a certain threshold as the final feature set for classification. Because a leave-one-out cross-validation strategy was used to test the generalizability of the classifier (**Figure 1**), the final feature sets differed slightly across iterations of the classification procedure. Cross-validation ensures that the classifier is trained on tested on independent data, thus avoiding concerns of double-dipping or circularity in the classification results (Kriegeskorte et al., 2009). Next, we defined the "consensus functional connectivity" as the functional connectivity features that appeared (i.e., showed sufficiently strong discriminative power) in every cross-validation iteration (Dosenbach et al., 2010; Zeng et al., 2012). Finally, we calculated the "region weight" of each feature by counting the number of times that feature appeared in the consensus functional connections in this study. Region weights represented the relative contribution of each feature to the classifier's discrimination of functional connectivity patterns in the rsfMRI data before vs. after IBMT training.

Support Vector Classification and Permutation Tests

After obtaining the data set of features with high discriminative power, we used support vector machines (SVM) with radial basis kernel function to perform the classification. The kernel function we used was:

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{4}$$

Here, sigma equaled 2. Due to our limited number of samples, we used a leave-one-out cross-validation strategy to estimate the performance of our classifier. Classification performance can be quantified using the generalization rate (GR), sensitivity and specificity based on the results of cross-validation. Note that the sensitivity represents the proportion of "post-training" samples correctly identified, while the specificity represents the proportion of "pre-training" samples correctly identified. The overall proportion of samples correctly predicted defines the GR.

Permutation tests were conducted to assess the performance of the classifier. In this study, the GR was chosen as the statistic to estimate the statistical significance of the classifier's performance. For each classification iteration, we randomly permuted 1000× the class labels ("pre-training" or "post-training") of the data being used to train the classifier. Importantly, the entire classification operation, including the feature selection and SVM, was carried out on every set of randomized class labels. We defined the GR as the performance of the classifier trained on permuted class labels, and we defined GR0 as the performance of the classifier trained on valid class labels. The p-values reported for classifier performance represent the probability of GR being no less than GR0. Therefore, when $p < 0.05$, this would indicate that the classifier could reliably decode whether the functional connectivity data was a pre-training or post-training sample.

Reliability of the Algorithm

Recent attention has focused on the possibility for systematic bias in fMRI scans resulting from in-scanner motion (Satterthwaite et al., 2013). As the optimal procedures for removing motion artifacts are still an ongoing area of research, and it is unclear exactly how different methods impact downstream analyses, we chose to test our main hypotheses on motion-corrected ("scrubbed") data. We implemented a scrubbing procedure as part of fMRI preprocessing. An estimate of motion at each time point was calculated as the frame-wise displacement (FD), using the three translational and three

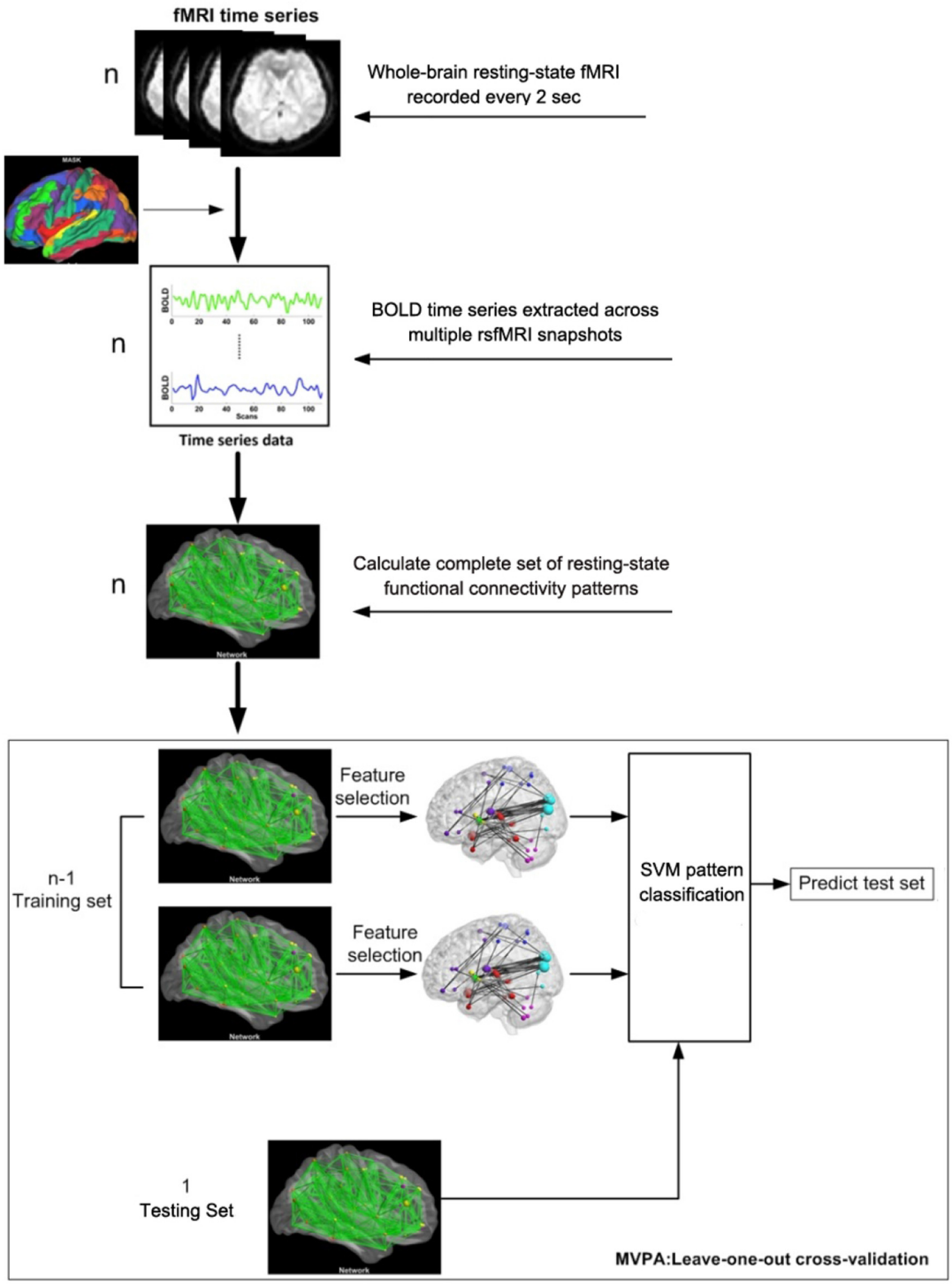


FIGURE 1 | Flow chart of the multivariate pattern analysis (MVPA) algorithm.

rotational displacements from rigid body motion correction procedure. Rotational displacements were converted from degrees to millimeters by calculating displacement on the surface of a sphere of radius 50 mm. Any frame i with $FD_i > 0.5$ mm was linearly interpolated. We found there was no material difference in the results obtained from scrubbed vs. unscrubbed data, confirming the reliability of our algorithm.

RESULTS

Classification Results

To estimate the effect of the selected parameters on the performance of the classifier, the cross-validation calculation was explored using different parameters. We repeated this calculation with a varying number of different features (from 40 to 300) in the feature selection and found that the classifier's best performance was achieved at 160 features (**Figure 2**) Therefore, we selected 160 as the optimal size of the final feature space for the classification analysis (i.e., the threshold was set at 160). We used this threshold value because many studies have used the same method for establishing the threshold (Corbetta and Shulman, 2002; Dosenbach et al., 2010; Shen et al., 2010; Zeng et al., 2012). In addition, this procedure was also used to choose the optimal value for the parameter C for the SVM algorithm. We repeated this calculation with a range of different values (dimension: 2–20 and C: 0.005:0.05:2). Then, we identified the values when the classifier achieved the maximum GR. We identified the optimal C as 0.01, which is consistent with previous studies (Besga et al., 2012; Zeng et al., 2012). When using 160 features in the feature selection (**Figure 2A**) and $C = 0.01$ for the SVM, the classifier achieved maximum performance (GR: 72%; sensitivity: 76%; specificity: 68%; **Figure 2B**). Permutation tests revealed that the classifier successfully learned the relationship between the resting-state functional connectivity data and the pre-training/post-training class labels ($p < 0.0001$).

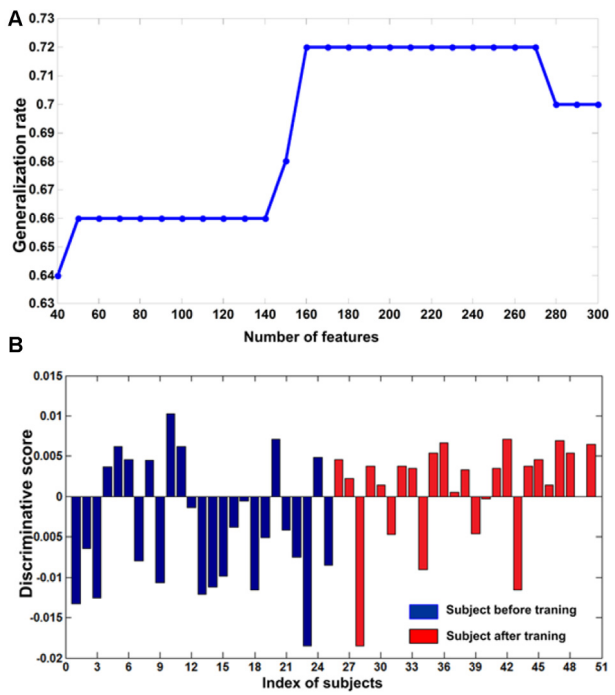


FIGURE 2 | (A) The curve of the generalization rate (GR) to the number of features. The horizontal axis represents the number of selected features and the vertical axis represents the GR. **(B)** The discriminative scores of all subjects. The first 25 samples represented subjects before training (blue bar). The remaining samples represented corresponding subjects after training (red bar).

Altered Resting-State Functional Connections after Training

Although 160 features were selected during a leave-one-out cross-validation iteration, the functional connectivity feature set selected in each iteration was slightly different (Dosenbach et al., 2010). In this investigation, 105 consensus functional connections were identified across the 50 (25 + 25 = 50) iterations of the cross-validation procedure (Dosenbach et al., 2010; Zeng et al., 2012). According to the Kendall tau rank correlation coefficient above, a positive correlation coefficient τ_i represents the ith functional connectivity feature that exhibits a significant increase after IBMT training, while a negative correlation coefficient τ_i represents the ith functional connectivity feature that exhibits a significant decrease after training. Comparing the consensus functional connectivity in subjects post-training vs. pre-training, we found more positive functional connections (60 connections) than negative connections (45 connections). This result indicates there are more increased functional connections after IBMT training. When analyzing the brain regions underlying this increase in functional connectivity, we found that occipital cortex (primarily including the superior and middle occipital gyrus) was functionally connected to many regions (**Figure 3**). Obviously, a large number of increased connections were encompassed between the occipital and temporal cortex (mainly comprising the superior temporal gyrus and its pole, and the insula), and between the occipital and the frontal cortex (mainly comprising frontal operculum). In addition, increased consensus functional connections between cerebellum and caudate were also detected (all $P < 0.05$). But we did not find significant lateralization differences among these bilateral areas.

DISCUSSIONS

Short-term mindfulness training induces a brain state that requires communication between multiple brain regions that collectively mediate the encoding and maintenance of sensory information (Tang et al., 2007, 2009, 2010, 2012, 2013, 2015a,b; Tang and Posner, 2014; Acevedo et al., 2016). Our results showed that 2 weeks of IBMT (5 h in total) reorganized the functional connectivity of large-scale brain networks involved in attention, cognitive and affective processing, awareness and sensory integration, and reward processing (e.g., the bilateral superior occipital/middle gyrus, bilateral frontal operculum, bilateral superior temporal gyrus, right superior temporal pole, bilateral insula, caudate and cerebellum.

Visual inputs contribute to over 90% of the total information (from all sensors) entering the brain. In literature, increased activity and connectivity in visual cortex are reported following

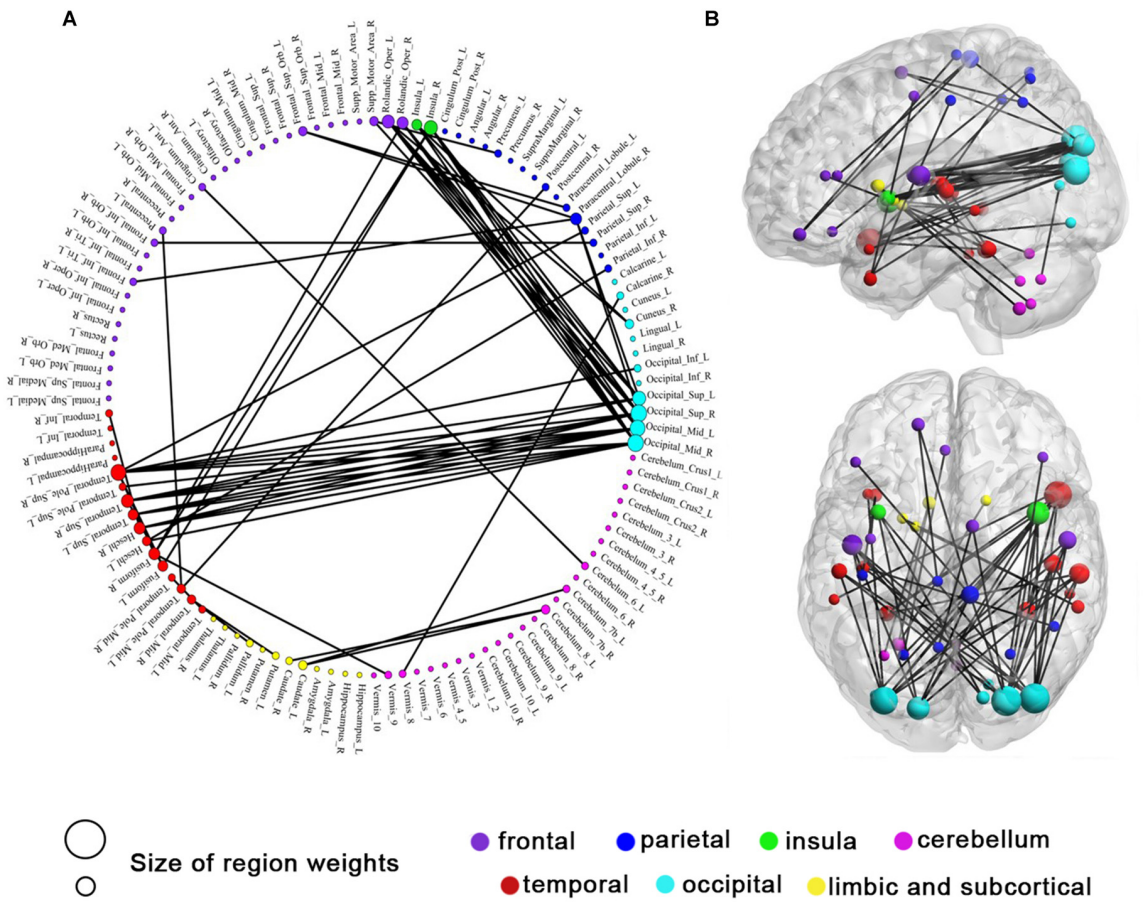


FIGURE 3 | Sixty consensus increased functional connections. Regions are color-coded by category. The line colors represent the relative consensus functional connections. **(A)** Region weights and the distribution of consensus increased functional connections in a circle graph. **(B)** Consensus increased functional connections demonstrated in left sagittal, and top axial view. The colors represent structural categories of brain regions and the size of circles represent region weights.

short- and long-term mindfulness meditation (Tang et al., 2009, 2015a; Kilpatrick et al., 2011; Xu et al., 2014; Berkovich-Ohana et al., 2016). However, the underlying mechanism of visual cortex involvement during mindfulness remains unclear. One possibility might be that when meditators close eyes and focus on the inside world, the sensory processes are amplified. When they continuously observe the inner thoughts entwined with mental images, the mental processes of visual areas are heavily involved in. Another possibility might be the relaxation effect following meditation because the activity and functional connectivity of the visual cortex is also increased during light sleep, sedation and alcohol consumption (Kiviniemi et al., 2005; Horovitz et al., 2008; Esposito et al., 2010).

IBMT includes components of body relaxation, mental imagery and mindfulness (maintaining a high degree of awareness of body, mind and external instructions guided by an IBMT coach). One of our studies also detected greater activity in visual cortex following only five sessions of IBMT (Tang et al., 2009). It makes sense that the component of body relaxation and mental imagery could induce greater activity in visual areas, consistent with previous reports (Tang et al., 2009, 2015a; Kilpatrick et al., 2011; Xu et al., 2014; Berkovich-Ohana et al., 2016). However, mindfulness is different from sleep or sedation state with low level of arousal, and it requires to maintain high level of vigilance state for meditators. This is in line with our results that a large number of increased connections were encompassed between the occipital and temporal cortex (mainly comprising the superior temporal gyrus and its pole), and between the occipital and the frontal cortex (mainly comprising frontal operculum and insula).

Recent studies indicated that meditation modified subsystems of attention (Jha et al., 2007; Tang et al., 2007). It is worth mentioning that the frontal cortex participates both dorsal and ventral attention network (Petersen and Posner, 2012; Schmidt et al., 2013; Tang et al., 2015a). This network is believed to modulate externally directed attention by amplifying or attenuating the saliency of relevant and irrelevant cues (Corbetta and Shulman, 2002). It has been shown in the

monkey that the combined actions of frontal eye fields and the occipital gyrus improved cross-area communication with attention (Gregoriou et al., 2009), and enhanced visual short-term memory performance (Liebe et al., 2012). In previous studies, we found that IBMT improves executive and altering attention networks compared to a well-controlled relaxation training (Tang et al., 2007, 2012, 2015a). Hence, we speculate that the long-range coupling between the occipital gyrus and frontal gyrus may improve and optimize global information processing helpful for the maintenance of a meditative state (Tang et al., 2007, 2015a; Tang and Posner, 2014).

Furthermore, we also found increased functional connectivity in adjacent occipital-temporal regions. These regions are often implicated in associative and item-recognition memory, a semantic network for both words and pictures, and self-cognition and awareness (Menon and Uddin, 2010). It might be possible that meditation training increases connections of temporal and occipital regions to allocate cognitive resources in order to improve performance. This idea is consistent with prior results showing that meditation improves attention and working memory performance (Tang et al., 2012, 2015a; Tang and Posner, 2014). The increased functional connectivity within temporal cortex is often associated with mood regulation and affective processing. Superior temporal sulcus was active in loving-kindness-compassion meditation (Lutz et al., 2008) and light modulation (Vandewalle et al., 2010). Insula was involved in interoceptive awareness, emotional responses and high-level attentional processes (Landtblom et al., 2011), consistent with our previous report that IBMT improves insula activity (Tang et al., 2009, 2015a). Importantly, using the Profile of Mood State and Attention Network Test, we found that IBMT improves attention and emotion regulation (Tang et al., 2007). The present results may indicate that IBMT improves emotion regulation through increased functional connectivity within temporal cortex.

In addition, increased consensus functional connections between cerebellum and caudate were also detected. Previous studies showed that the caudate nucleus plays a vital role in reward and learning, and the cerebellum may contribute to emotion and cognitive processing (Tang et al., 2009; Bostan et al., 2010; Ding et al., 2015, 2014). A recent study also showed that the basal ganglia and cerebellum may be linked together to form an integrated functional network that influences cognitive and affective processing (Bostan et al., 2010), and may support the brain state associated with meditation.

Taken together, our study indicates that MVPA of functional connectivity patterns in rsfMRI data effectively discriminates the different brain states in individuals before vs. after short-term meditation training. We found significantly increased functional connectivity between occipital, temporal and frontal regions, which may suggest that meditation training mainly improves attention, emotional, cognitive and reward processing. Our results provide new insights into the underlying neural mechanisms of mental training such as mindfulness, identifying complex changes in resting-state functional integration across the brain as a result of brief mindfulness training. It should be noted that we are aware of the potential issue of reverse inference when interpreting results (Poldrack, 2006). This is a legitimate first step in attempting to understand the significance of the observed changes in functional connectivity patterns following mindfulness training, and these results can be strengthened by future work focused on the functional selectivity and specificity of these changes in neural connectivity.

AUTHOR CONTRIBUTIONS

Y-YT designed and conducted research; YT, Y-YT, RT and JAL-P analyzed data and wrote the article.

ACKNOWLEDGMENTS

We thank lab members for assistance with data collection. This study was supported by the Office of Naval Research (award no. N000141310628).

REFERENCES

Acevedo, B. P., Pospos, S., and Lavretsky, H. (2016). The neural mechanisms of meditative practices: novel approaches for healthy aging. *Curr. Behav. Neurosci. Rep.* 3, 328–339. doi: 10.1007/s40473-016-0098-x

Berkovich-Ohana, A., Harel, M., Hahamy, A., Arieli, A., and Malach, R. (2016). Alterations in task-induced activity and resting-state fluctuations in visual and DMN areas revealed in long-term meditators. *Neuroimage* 135, 125–134. doi: 10.1016/j.neuroimage.2016.04.024

Besga, A., Termenon, M., Graña, M., Echeveste, J., Pérez, J. M., and Gonzalez-Pinto, A. (2012). Discovering Alzheimer's disease and bipolar disorder white matter effects building computer aided diagnostic systems on brain diffusion tensor imaging features. *Neurosci. Lett.* 520, 71–76. doi: 10.1016/j.neulet.2012.05.033

Bostan, A. C., Dum, R. P., and Strick, P. L. (2010). The basal ganglia communicate with the cerebellum. *Proc. Natl. Acad. Sci. U S A* 107, 8452–8456. doi: 10.1073/pnas.1000496107

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43, 44–58. doi: 10.1016/j.neuroimage.2008.06.037

Ding, X., Tang, Y. Y., Cao, C., Deng, Y., Wang, Y., Xin, X., et al. (2015). Short-term meditation modulates brain activity of insight evoked with solution cue. *Soc. Cogn. Affect. Neurosci.* 10, 43–49. doi: 10.1093/scan/nsu032

Ding, X., Tang, Y. Y., Tang, R., and Posner, M. I. (2014). Improving creativity performance by short-term meditation. *Behav. Brain Funct.* 10:9. doi: 10.1186/1744-9081-10-9

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., et al. (2010). Prediction of individual brain maturity using fMRI. *Science* 329, 1358–1361. doi: 10.1126/science.1194144

Esposito, F., Pignataro, G., Di Renzo, G., Spinali, A., Paccone, A., Tedeschi, G., et al. (2010). Alcohol increases spontaneous BOLD signal fluctuations in the visual network. *Neuroimage* 53, 534–543. doi: 10.1016/j.neuroimage.2010.06.061

Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700–711. doi: 10.1038/nrn2201

Gregoriou, G. G., Gotts, S. J., Zhou, H., and Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science* 324, 1207–1210. doi: 10.1126/science.1171402

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* 62, 852–855. doi: 10.1016/j.neuroimage.2012.03.016

Haynes, J. D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi: 10.1038/nrn1931

Horovitz, S. G., Fukunaga, M., de Zwart, J. A., van Gelderen, P., Fulton, S. C., Balkin, T. J., et al. (2008). Low frequency BOLD fluctuations during resting wakefulness and light sleep: a simultaneous EEG-fMRI study. *Hum. Brain Mapp.* 29, 671–682. doi: 10.1002/hbm.20428

Jha, A. P., Krompinger, J., and Baime, M. J. (2007). Mindfulness training modifies subsystems of attention. *Cogn. Affect. Behav. Neurosci.* 7, 109–119. doi: 10.3758/cabn.7.2.109

Kendall, M., and Gibbons, J. D. (1990). "Rank correlation methods edward arnold," in *A Division of Hodder and Stoughton, A Charles Griffin Title* (London), 29–50.

Kilpatrick, L. A., Suyenobu, B. Y., Smith, S. R., Bueller, J. A., Goodman, T., Creswell, J. D., et al. (2011). Impact of mindfulness-based stress reduction training on intrinsic brain connectivity. *Neuroimage* 56, 290–298. doi: 10.1016/j.neuroimage.2011.02.034

Kiviniemi, V. J., Haanpää, H., Kantola, J. H., Jauhiainen, J., Vainionpää, V., Alahuhta, S., et al. (2005). Midazolam sedation increases fluctuation and synchrony of the resting brain BOLD signal. *Magn. Reson. Imaging* 23, 531–537. doi: 10.1016/j.mri.2005.02.009

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Landtblom, A. M., Lindehammar, H., Karlsson, H., and Craig, A. (2011). Insular cortex activation in a patient with "sensed presence"/ecstatic seizures. *Epilepsy Behav.* 20, 714–718. doi: 10.1016/j.yebeh.2011.01.031

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., and Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *J. Cogn. Neurosci.* 24, 61–79. doi: 10.1162/jocn_a_00140

Lewis-Peacock, J. A., and Norman, K. A. (2013). "Multi-voxel pattern analysis of fMRI data," in *The Cognitive Neurosciences*, 4th Edn. eds M. S. Gazzaniga and G. R. Mangun (Cambridge, MA: MIT Press), 911–920.

Lewis-Peacock, J. A., and Norman, K. A. (2014). Competition between items in working memory leads to forgetting. *Nat. Commun.* 5:5768. doi: 10.1038/ncomms6768

Liebe, S., Hoerzer, G. M., Logothetis, N. K., and Rainer, G. (2012). Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nat. Neurosci.* 15, 456–462. doi: 10.1038/nn.3038

Lutz, A., Brefczynski-Lewis, J., Johnstone, T., and Davidson, R. J. (2008). Regulation of the neural circuitry of emotion by compassion meditation: effects of meditative expertise. *PLoS One* 3:e1897. doi: 10.1371/journal.pone.0001897

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Petersen, S. E., and Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* 35, 73–89. doi: 10.1146/annurev-neuro-062111-150525

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63. doi: 10.1016/j.tics.2005.12.004

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage* 64, 240–256. doi: 10.1016/j.neuroimage.2012.08.052

Schmahmann, J. D., Doyon, J., McDonald, D., Holmes, C., Lavoie, K., Hurwitz, A. S., et al. (1999). Three-dimensional MRI atlas of the human cerebellum in proportional stereotaxic space. *Neuroimage* 10, 233–260. doi: 10.1006/nimg.1999.0459

Schmidt, S. A., Akrofi, K., Carpenter-Thompson, J. R., and Husain, F. T. (2013). Default mode, dorsal attention and auditory resting state networks exhibit differential functional connectivity in tinnitus and hearing loss. *PLoS One* 8:e76488. doi: 10.1371/journal.pone.0076488

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *Neuroimage* 49, 3110–3121. doi: 10.1016/j.neuroimage.2009.11.011

Tang, Y. Y., Hölzel, B. K., and Posner, M. I. (2015a). The neuroscience of mindfulness meditation. *Nat. Rev. Neurosci.* 16, 213–225. doi: 10.1038/nrn3916

Tang, Y. Y., Lu, Q., Feng, H., Tang, R., and Posner, M. I. (2015b). Short-term meditation increases blood flow in anterior cingulate cortex and insula. *Front. Psychol.* 6:212. doi: 10.3389/fpsyg.2015.00212

Tang, Y. Y., Lu, Q., Geng, X., Stein, E. A., Yang, Y., and Posner, M. I. (2010). Short-term meditation induces white matter changes in the anterior cingulate. *Proc. Natl. Acad. Sci. U S A* 107, 15649–15652. doi: 10.1073/pnas.1011043107

Tang, Y. Y., Ma, Y., Fan, Y., Feng, H., Wang, J., Feng, S., et al. (2009). Central and autonomic nervous system interaction is altered by short-term meditation. *Proc. Natl. Acad. Sci. U S A* 106, 8865–8870. doi: 10.1073/pnas.0904031106

Tang, Y.-Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., et al. (2007). Short-term meditation training improves attention and self-regulation. *Proc. Natl. Acad. Sci. U S A* 104, 17152–17156. doi: 10.1073/pnas.0707678104

Tang, Y.-Y., and Posner, M. I. (2014). Training brain networks and states. *Trends Cogn. Sci.* 18, 345–350. doi: 10.1016/j.tics.2014.04.002

Tang, Y. Y., Rothbart, M. K., and Posner, M. I. (2012). Neural correlates of establishing, maintaining and switching brain states. *Trends Cogn. Sci.* 16, 330–337. doi: 10.1016/j.tics.2012.05.001

Tang, Y. Y., Tang, R., and Posner, M. I. (2013). Brief meditation training induces smoking reduction. *Proc. Natl. Acad. Sci. U S A* 110, 13971–13975. doi: 10.1073/pnas.1311887110

Tong, F., and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63, 483–509. doi: 10.1146/annurev-psych-120710-100412

Vandewalle, G., Schwartz, S., Grandjean, D., Wuillaume, C., Balteau, E., Degueldre, C., et al. (2010). Spectral quality of light modulates emotional brain responses in humans. *Proc. Natl. Acad. Sci. U S A* 107, 19549–19554. doi: 10.1073/pnas.1010180107

Xu, J., Vik, A., Groote, I. R., Lagopoulos, J., Holen, A., Ellingsen, Ø., et al. (2014). Nondirective meditation activates default mode network and areas associated with memory retrieval and emotional processing. *Front. Hum. Neurosci.* 8:86. doi: 10.3389/fnhum.2014.00086

Zeng, L. L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., et al. (2012). Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain* 135, 1498–1507. doi: 10.1093/brain/aws059

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 28 November 2016
doi: 10.3389/fnsys.2016.00097

Regular Cycles of Forward and Backward Signal Propagation in Prefrontal Cortex and in Consciousness

Paul J. Werbos and Joshua J. J. Davis*

Department of Mathematical Sciences, Center for Large-Scale Optimization and Networks, University of Memphis, Memphis, TN, USA

This paper addresses two fundamental questions: (1) Is it possible to develop mathematical neural network models which can explain and replicate the way in which higher-order capabilities like intelligence, consciousness, optimization, and prediction emerge from the process of learning (Werbos, 1994, 2016a; National Science Foundation, 2008)? and (2) How can we use and test such models in a practical way, to track, to analyze and to model high-frequency ($\geq$ 500 hz) many-channel data from recording the brain, just as econometrics sometimes uses models grounded in the theory of efficient markets to track real-world time-series data (Werbos, 1990)? This paper first reviews some of the prior work addressing question (1), and then reports new work performed in MATLAB analyzing spike-sorted and burst-sorted data on the prefrontal cortex from the Buzsaki lab (Fujisawa et al., 2008, 2015) which is consistent with a regular clock cycle of about 153.4 ms and with regular alternation between a forward pass of network calculations and a backwards pass, as in the general form of the backpropagation algorithm which one of us first developed in the period 1968–1974 (Werbos, 1994, 2006; Anderson and Rosenfeld, 1998). In business and finance, it is well known that adjustments for cycles of the year are essential to accurate prediction of time-series data (Box and Jenkins, 1970); in a similar way, methods for identifying and using regular clock cycles offer large new opportunities in neural time-series analysis. This paper demonstrates a few initial footprints on the large "continent" of this type of neural time-series analysis, and discusses a few of the many further possibilities opened up by this new approach to "decoding" the neural code (Heller et al., 1995).

Keywords: backpropagation, synchronization, prefrontal cortex (PFC), consciousness, spike sorting, neural codes, bursts, alpha rhythm

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research, Inc., USA

Reviewed by:
Peter Erdi,
Kalamazoo College, USA
Zoltan Somogyvari,
Wigner Research Centre of the Hungarian Academy of Sciences, Hungary
Venkat Rao,
Parsons, USA

****Correspondence:***
Paul J. Werbos
werbos@ieee.org

Received: *11 July 2016*
Accepted: *08 November 2016*
Published: *28 November 2016*

Citation:
Werbos PJ and Davis JJJ (2016) Regular Cycles of Forward and Backward Signal Propagation in Prefrontal Cortex and in Consciousness. Front. Syst. Neurosci. 10:97. doi: 10.3389/fnsys.2016.00097

ALTERNATE NEURAL NETWORK MODELS TO EXPLAIN/REPLICATE CONSCIOUSNESS (QUESTION 1)

Mathematical neural network models actually fall into two categories: (1) models of mature (fixed) neural circuits, such as elaborate models by Grossberg articulating what was learned by neuroscientists like Van Essen in deciphering specific visual pathways as they appear in visual cortex of a mature adult; (2) models of the more fundamental and universal learning capabilities of the

brain, which aim to replicate competence in vision, decision-making, prediction, and other tasks as the emergent outcome of the learning process. This paper focuses exclusively on the second type of neural network model. That type of neural network model is itself a very large and diverse set. There have been efforts to combine the two types of neural network modeling (as in some efforts by Grossberg), but those are beyond the scope of this paper.

The effort to develop mathematical neural network models of intelligence and learning started from the seminal work of two groups: (1) the "cyberneticians" (Rav, 2002), such as Von Neumann, Wiener, and McCulloch, who developed the concept of neural networks as an approach to artificial intelligence; and (2) Donald Hebb, the neuropsychologist, whose book (Hebb, 1949) served as a manifesto to the new field of neural networks. Before Hebb, efforts to understand the dynamics of the cerebral cortex usually focused on very specialized attempts to understand the different functions of different Broca areas, in typical mature brains. Hebb called us to pay more attention to the experiments by Lashley on "mass action," showing how any one area of the cortex can take over functions which are usually found in another area, when the latter is destroyed and when the required connections still exist. Walter Freeman, one of the important followers of Lashley, played a pivotal role in expanding our understanding of mass action in the brain (Freeman, 1975/2004); Karl Pribram and Jerry Lettvin, among others, also performed important experiments on that topic. In effect, Hebb challenged us to try to answer question (1) above, and many of us have tried to rise to this challenge.

In his final great work, Walter Freeman (with Robert Kozma) challenged a group of experimental neuroscientists and relevant theorists to submit chapters to a book addressing a key question (Kozma and Freeman, 2016): are the mathematical models now used in computational neuroscience powerful enough to answer question (1), and, if not, what changes are needed?

As part of that book, Freeman and Kozma ask whether neural network models would have to be extended, to account for field effects over three dimensions or even over quantum mechanical effects (Werbos and Dolmatova, 2016), in order to explain or construct the highest levels of intelligence or consciousness. Even if we focus for now on trying to understand the level of general intelligence which we see in individual brains of mice or rats (Werbos, 2014), it is possible that field effects *within* neurons give them a level of computational power beyond what traditional neural network models allow (National Science Foundation, 2008). Those extensions are important topics for research, but this paper will focus on simpler extensions, already an important part of the neural network field.

In the 1960's, neural network models inspired by digital computers (Rav, 2002) generally assumed that the brain itself must be like a digital computer, and hence that the "neural code" would consist of ones and zeroes, encoded simply as the presence or absence of spikes. Even today, many of the models used in computational neuroscience continue that tradition, by assuming that the neural code consists of spikes or pulses propagating and integrated in an asynchronous way, without any kind of master clock of the sort one would find in a modern computer.

Unfortunately, it was very difficult to find learning models capable of training such networks to perform even very simple tasks, let alone the more complex tasks which mammal brains can handle (Minsky and Papert, 1969). The neural network field languished and became even disreputable within artificial intelligence and engineering, until the field learned to accept a new type of learning model which required a different kind of neural code. The simplest version of the new type of learning model was renamed "backpropagation" (originally the name of a different algorithm by Rosenblatt), and simplified and popularized very widely (Rumelhart et al., 1986). Backpropagation involved two major new elements: (1) use of a continuous-variable neural code, instead of 1's and 0's; (2) use in learning of the derivatives of some error measure, calculated by signals propagating backwards in the network, with or without scaling enroute, justified by the general chain rule for ordered derivatives proven in 1974 for feedforward networks with or without time-delayed recurrence (Werbos, 1994) and generalized in 1980 to all types of recurrent network (Werbos, 2006). This kind of adaptation requires alternating cycles of forward calculation and backward calculation, which in turn requires a kind of master clock.

When this concept was presented to Minsky himself circa 1970, he objected that people in the modeling field know that there are no clocks in the brain, and know that all neurons use a code which is strictly binary, strictly defined by presence or absence of a spike. In reply, he was shown patch clamp recordings from higher centers of the brain, taken from Rosenblith, which demonstrated a sequence of volleys or bursts (Rosenblith, 1961) with regular timing; the volleys can be viewed as a set of spikes "on top of each other," but the overall intensity of the bursts varied in a continuous way, from small bursts to large bursts. Thus, instead of viewing the data as a sequence of spikes at different times, one would view them as a continuous measure of intensity $x_k(t)$ for neuron k, where t takes on discrete integer values, relative to some kind of system-wide clock. Bursts continue to appear in the output of giant pyramid cells (Bear et al., 2007), cells which serve as the backbone and final output path of all parts of the cerebral cortex.

The rebirth of neural networks in the 1980s was based primarily on backpropagation, on learning models which assume a continuous neural code and an alternation of a forward pass to do computational work and a backwards pass for effective learning in the face of complex tasks. Also very important was a third class of neural network model (Grossberg, 1971), which we think of as the ODE type (ordinary differential equations), assuming a continuous neural code but, like the spiking models, asynchronous, and defined over continuous time.

It should be emphasized that the original, general form of "backpropagation" is a learning algorithm or a stream of local calculations implementing that algorithm. It is not the specific type of neural network topology, the Multilayer Perceptron (MLP), which was used most often in popularized books and simple applications. Backwards flows of calculation are needed for the efficient calculation of derivatives in general, whether scaled and modulated or unscaled (Werbos, 2006). The most powerful computational methods suitable for complex, general

nonlinear tasks do require the calculation and use of derivatives. The topology proposed in our theory (Werbos, 2009) is more complex and powerful than the simple MLP.

This gives us three general families of neural network learning model in use today: (1) spiking; (2) the backpropagation family as defined here; and (3) ODE. Spiking and ODE models are very popular in computational neuroscience, and have been used in the analysis of real-time data from brains. Models of the backpropagation family have been much more widespread in engineering and computer science, where they have led to major breakthroughs in intelligent control (Lewis and Derong, 2012; Werbos, 2014) and in pattern recognition with "deep learning" (National Science Foundation, 2008; Ng et al, 2008; Schmidhuber, 2015). The main purpose of this paper is to discuss and illustrate how models in the backpropagation family can also be engaged and tested on multielectrode array data, and to show that the data available so far do not rule them out.

In this paper, we do not argue that models of the backpropagation family are sufficient to answer question (1), or to replicate the full range of higher-level capabilities we see in the brains of rodents. Rather, we would envision a kind of hybrid model, in which giant pyramid cells of neocortex receive clock pulses from the nonspecific thalamus at a key junction on the apical dendrite (Werbos, 2009), and output bursts under the control of that clock, while a complex network of interneurons provide Supplementary capabilities like associative memory, influenced by their inputs from the pyramid cells but not directly governed by a global clock. This is part of a more general theory of intelligence in the mammal brain (Werbos, 2009), grounded in general mathematical principles derived from analyzing what is required to achieve functional brain-like capabilities in tasks like decision-making and prediction of the environment (Werbos, 2010).

Many modelers correctly observed years ago that models based on the simplified popularized versions of backpropagation (like the MLP) would not be plausible as models of biological neural networks (BNN). However, deeper work on systems neuroscience has already revealed flows of information and types of synaptic connection supporting the idea that backward passes (as in the more general family of backpropagation designs) do exist in the brain (Smirnova et al., 1993; Buzsáki et al., 2012). A thorough review of learning and rhythms in the hippocampus (Kahana et al., 2001) shows that the mechanisms of learning do appear to vary as a function of the time of stimuli within the theta clock cycle, even though the origins of the theta clock in the hippocampus remain controversial. This paper focuses on the cerebral cortex, in part because the fibers from the nonspecific thalamus to the apical dendrites of giant pyramid cells have been well-established for decades, and in part because of the intrinsic importance of the cerebral cortex. It would be possible to model the oscillations in the nonspecific thalamus with ODE, but it is not really necessary at this stage, because they are so regular, and because they are essentially a hard-wired feature of the brain, not the kind of feature which emerges in detail from learning.

There is also an important connection between the theory of brain functioning presented in (Werbos, 2009) and the "Global Workspace" theory of consciousness developed by Bernie Baars, one of the top leaders in consciousness research (Baars, 2016). Baars argues that the information in our "conscious awareness" is basically just the current image of reality reconstructed in the cerebral cortex, by the "working memory" mechanisms described in wet neuroscience work by Goldman-Rakic and Legothetis, among others. Those researchers have observed that recurrent neural networks, with the kind of reverberations necessary for short-term memory, play a central role in this kind of consciousness. From mathematical work on functional requirements and training of recurrent networks (White and Sofge, 1992), we understand that a different kind of recurrence and training is required in order to produce this kind of short-term memory or "nonlinear state estimation," compared with the kinds required for longer-term associative memory or "settling down" in image processing. Neither we nor Baars would say that recurrence in the brain is only of the time-delayed kind, but clocks and backwards passes turn out to be necessary for that kind, and for hybrid systems which include that kind of capability.

More concretely, the theory in Werbos (2009) proposes that the global workspace can be represented as the vector $\underline{R}(t)$ made up of the final axon burst outputs $R_k(t)$ of giant pyramid neurons k at clock time t, and that the cortico-thalamic system learns to build up this filtered image of reality and to predict inputs from the specific thalamus $\underline{X}(t)$ by a robust variation of the Stochastic Encoder/Decoder/Predictor (SEDP) model (White and Sofge, 1992). Simplified special cases of that model (like the Ford software for Time-Lagged Recurrent Networks) have won many recent time-series prediction competitions, but of course we expect the brain to have more powerful functional capabilities which include but surpass the simple TLRN. This is simply one way to translate the Baars theory into something we can test in a more fine-grained way on real-time brain data.

This theory of cortical function can also be seen as a way of implementing Llinas' theory of the brain as a prediction system (Llinas and Roy, 2009). Llinas' earlier work demonstrating highly precise synchronized clocks in the motor system of the brain is also relevant to the approach (Sugihara et al., 1993). In conversation at a workshop organized by Karl Pribram, Nicolelis reported that their important work on the cortico-thalamic system (Nicolelis et al., 1995) showed how cells in the thalamus which were initially good advance predictors of their neighbors [cells in $\underline{x}(t)$] would relearn this prediction ability after it was destroyed by a lesion.

Strictly speaking, the theory in Werbos (2009) asserts that giant pyramid cells are adapted based on backwards error signals which are the sum of signals based on prediction error in the cerebro-thalamic circuit and on signals based on error signals from the basal ganglia and the limbic system, reflecting additional ways in which the brain can assess the quality of the outputs produced by the cerebral cortex. It asserts that the limbic system implements some variant of reinforcement learning (Lewis and Derong, 2012) which requires a global clock cycle twice as long (θ) as the clock cycle (α) required for prediction. It does not specify what drives the theta rhythm in the limbic system, but it allows for the possibility that the primary clock is the alpha clock in the nonspecific thalamus and cortex, and that the theta rhythms are somehow synchronized with that one. In any case, this paper focuses more on the cerebral cortex.

SELECTION OF REAL-TIME MULTIELECTRODE DATA TO TEST FOR CLOCK CYCLES AND BACKWARDS PASSES

The new work reported in this paper was initially inspired by (unpublished) comments by Barry Richmond of NIH, enroute to a meeting at the Dana Foundation. In his data on the neural code, he said that he saw a regular alternation between a short quiet period (on the order of 10–20 ms), a kind of "normal window" of signals flowing in the usual expected direction from inputs to outputs, on the order of 40–50 ms, and then a puzzling backwards window of 40–50 ms in which information seemed to go in the opposite direction. "I am not sure what to make of that second window, but I would guess that it has something to do with adaptation, somehow." Given the prior work on neural network modeling, reviewed in the previous section, we found this to be very exciting, but we were unable to obtain more details, other than Richmond's published papers. The goal of this new work was essentially to reconstruct the details, by use of new data sets.

The new theory of cortical dynamics does not require the presence of a quiet period, but Richmond's observation suggested that it should be there. If so, it would be an excellent starting point for looking for a forward pass and a backwards pass. Thus, the first stage of our work was to look for that kind of regular quiet period.

Initially, we scanned the real-time Ecog data collected by Walter Freeman (Heck et al, 2016) to see whether it could be a basis for identifying quiet periods. Unfortunately, because this was data on field potentials at the outer surface of the cortex, the times of zero potential reflected a cancelation of positive and negative inputs to the neurons, rather than low activity as such. It seemed logical to expect that the "quiet periods" are best defined as periods when the outputs of the cortical pyramid cells (either zero or bursts, a monotonic output) were near to zero. Thus, we looked for real-time data from deeper in the brain, where they would reflect spikes or bursts output by neurons. (Unfortunately, we did not have access at that time to the spike-sorted parts of the Freeman data).

Note that simple Fourier analysis or wavelet analysis would not be a proper way to look for such regular quiet periods, because the activity in the brain at times which are not quiet depends a great deal on inputs which vary as a function of the experience of the rat or the mouse, and would show oscillations related to that experience (Heck et al, 2016; Kozma et al., 2012). The new theory does not question the existence of such important oscillations and activity, but it does require new methods of analysis in order to track the specific type of hardwired clock assumed here.

The next step was to thank Professor Jennie Si of Arizona State University for access to her extracellular data collected from deep in the brains of experimental rats (Yuan et al., 2015), data collected under NSF funding under a data management plan which promised public access to the data. Si warned us, however, that her real-time 16-channel data collected at 24 khz leaves open important and difficult questions about how to do spike sorting. In fact, when we looked for regular quiet periods in her raw data, we did not really find it. We found a mix of positive and negative signs as overwhelming as what we saw in the raw data from Freeman. There were a few hints of regular timing in plots in Excel of the high-pass filtered version of her data, at times of maximum activity in her experiments, but we decided to look for more monotonic data, more representative of the actual outputs of neurons, based on the current best state of the art in spike sorting, which we then studied in some detail (Harris et al., 2000; Buzsáki et al., 2012; Rossant et al., 2016).

All of the work reported here was performed using the database pfc-2 (Fujisawa et al., 2008, 2015) taken from the repository at crncs.org. All but some test and exploratory runs were based on two versions of the MATLAB file EE188_example, kindly emailed to us by Prof. Fujisawa of Riken. One version, about 3 megabytes in size, was identical to the file discussed in Fujisawa et al. (2008), underlying all its major reported results and Figures concerning local circuits and learning in prefrontal cortex. An expanded version, about 5 megabytes, included spike sorted data from an additional 32-channel silicon probe inserted into the CA1 region of hippocampus, from which real-time data were also collected on the same time scale (20 khz) in the same long sequence of sessions.

One of the great benefits of the pfc-2 database is that it includes a sorting of the pfc-neurons into three groups—confirmed pyramid cells, confirmed interneurons and confirmed others. This made it possible to estimate the location of quiet periods (start and end of each clock cycle) based on the spike sorted data from the pyramid cells only, and then use those estimates to analyze data from all of the pfc neurons. Because data was also available from CA1, we also performed a few analyses of CA1 data, but for reasons of time we used this data only for phase one of this work, the initial exploration of possible clock cycles. The greatest part of this work involved developing practical nonparametric analysis methods, and debugging and testing their use in MATLAB and in Octave.

The main part of these MATLAB files was a collection of six variables, of which we only used two:

(1) spiket(j), which gives the clock time at which spike number j was detected; and (2) spikeind(j), which contains a numerical ID (in the range from 1 to 400) for the neuron at which spike number j occurred. Of course, we also used the file containing the table of neuron types, identifying which was a pyramid cell and which was something else. We also used Excel to inspect the original spike sorted files, like EE188.res.1 and clu and fet, which contained the original version of the spike sorted data, and verified the simple exact mapping from those raw files to the more compact MATLAB files. Using that correspondence, it would be possible to repeat this analysis for all the sessions in the pfc-2 database, and evaluate the stability of the clock time over time, and with respect to sleep and wake states.

These two MATLAB files represented the best data we could find on the actual outputs of cortical neurons at the present time. However, the spike-sorting which was used to generate this data was all based on the concept of spike-based neural networks. It is only natural that the computational work on spike sorting has largely been inspired by the neural network models which are currently most popular, but this leaves open many questions about how well the spike-sorted data represents the actual output

of the neurons, and about how to test models in which pyramid cells output bursts more than spikes. From the review in Harris et al. (2000), it is clear that regular behavior in brains may be more visible when we focus on bursts rather than spikes, but the full methodology of burst-sorting given in Harris et al. (2000) was beyond the time and resources available for this initial work. In consequence, we used a very simple routine for burst-sorting, based on the easy half of the procedure described in Harris et al. (2000): we filtered all the spikes in the MATLAB file down to a smaller file, in which we simply threw out all spikes which were not accompanied by other spikes from the same neuron within 6 ms of the same time. We found that the burst-filtered version of the pfc2 database was only about 1/3 less than the size of the original database, and that all of the measures of pattern which we looked for were stronger on the burst-filtered version of the database.

We did consider trying to use the fet files in the pfc-2 database to perform the additional filtering used in Harris et al. (2000). However, some of the features in that file seemed to call for the use of distance measures based on distance, while others appeared to be more like measures of intensity, calling for the use of measures like inner product. Clearly it will be an important research task for the future to sort out these kinds of issues, to organize the development of burst sorting and measurement in a more systematic way, and to apply them to databases like pfc-2.

COMPUTATIONAL METHODS AND RESULTS USED TO PROBE FOR CLOCK CYCLES AND DIRECTION OF SIGNAL FLOW

Effective and robust dynamic modeling of complex systems like the cerebral cortex generally requires that we start with a phase of exploratory data analysis, in order to avoid missing major patterns and being limited by initial assumptions (Hoaglin et al., 1983). The measures used here were developed in order to be as simple and direct as possible, for this early stage, while—most importantly—articulating or estimating the key hypotheses under study (Werbos, 1990). The issue of robustness is very tricky, in a situation where the raw data includes only about 100 variables which are part of a very information rich highly nonlinear system containing billions of neurons evolved over millions of years to handle the maximum throughput of information (Macke et al., 2011). We strongly hope that future research will probe these theoretical issues in more detail. Here we will simply report what the exploratory measures are which we used, and leave the refinements to the future.

We developed and used four sets of computational measures here. All four required us (the user) to specify a candidate clock cycle time, and another parameter K, to be discussed further below. All four call out for some combination of simulation studies (like those used in Werbos, 1994) to assess the robustness of competing statistical methods) and mathematical analysis to develop more formal measures of statistical significance—though the main results in **Table 2** and **Figure 1** are large enough to be clear already.

Summary of Methods and Findings

For phase one of this work, we developed "quiet time" measures, to tell us whether there exists an interval of 10 ms, at the same phase of every clock cycle, over K clock cycles, which regularly experiences fewer spikes than other times in that clock cycle. We applied these measures to the four most active neurons in the entire database individually, to the pyramidal cells, and to larger sets of neurons, with or without burst filtering, for all possible clock times which were integer multiples of 0.1 ms between 100 and 200 ms. The four most active neurons included three cells in the hippocampus (neurons number 329, 349, and 373), and an interneuron in the cerebral cortex (120). The interneuron and neuron 349 did not show regular quiet periods, but neurons 329, 373 and the pyramidal cells in prefrontal cortex as a group all

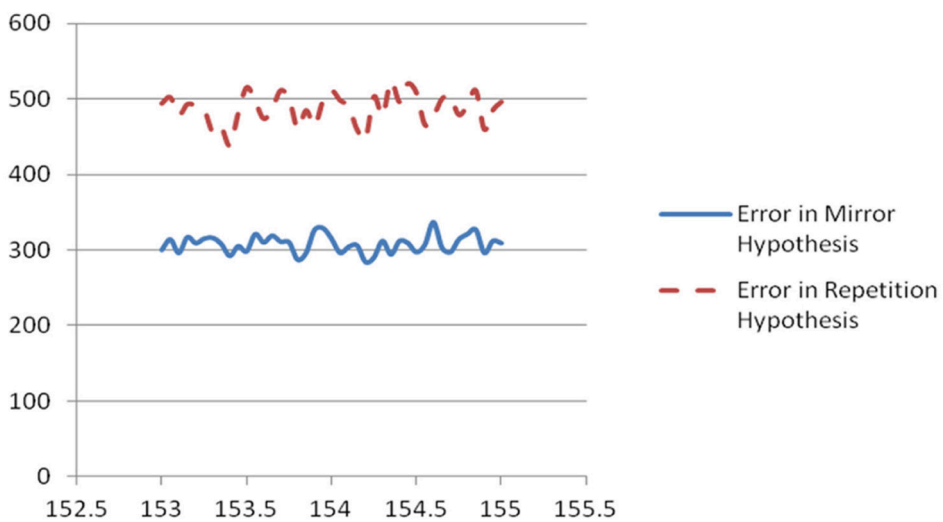


FIGURE 1 | Plot of scaled unweighted e↓ and e↑ vs. assumed clock time from Table 2.

showed regular quiet periods, at similar ranges of time intervals, with $K = 100$, $K = 1000$, and even $K = 10{,}000$. The strongest range of possible clock cycle times (the range with the quietest quiet periods) was 154 ± 1 ms, but 145 ± 1 and 134 ± 2 seemed plausible enough to warrant further investigation. It was striking that the same clock periods seemed to be best for all three data sets, with all three choices for K. However, it was also disappointing that we could not be sure what the best estimate of clock time would be, within those ranges, for the available data using this measure.

In phases two and three of this work, we mainly focused on using the quiet time results to identify clock cycles, and to test whether the sequence of firing in the later half of a clock cycle ("PM") is more like a repetition of the sequence in the first half ("AM") or like a reversed sequence, or mirror image. We also hoped that further analysis would give us more accuracy and certainty in knowing what the clock time is; that hope worked out for phase three, but not for the simpler work in phase two.

Before starting this work, we dreamed of studying neurons arranged in a network, such that we could actually see the "lights" (firing) moving from back to front in a forward pass ("AM") and then from front back in a backwards pass ("PM"). However, spike sorting provided only neuron IDs, not physical location. Fujisawa et al. (2008) provided what may be the best identification of neural networks from spike sorting available in the literature, but the identification did not cover most of the neurons in the dataset, and it was based on a cross-correlogram methodology which raises questions about robustness and possible systematic error. Thus, for phase two we looked at simple measures which describe the sequence of firing of individual neurons within a clock cycle, while for phase three we looked at which neurons fire in what order or sequence. The phase three results look more interesting, but for completeness we will also describe the phase two results. Both in phase two and phase three, the error in assuming mirror-image signal propagation was about 1/3 less than the error in assuming repetition of the same sequence, across all candidate clock cycle times, and K of 20, 100, and 1000 (the three choices we considered). In retrospect, we suspect that there might also be a useful way to exploit the information in the pfc-2 database about which neuron belongs to which of the twelve shanks, which does give some information about locations.

In phase 3, we also calculated three measures of "inertia," of the tendency for the same list of neurons to fire from one clock cycle to the next, in the same sequence in AM or PM. This generally sharpened and validated our estimate that 153.4 ms is the correct clock cycle time all across this data (session 188 for the rat identified as EE). We looked a bit for evidence of phase drift or cycle time drift from one span of data to another (where a "span" is K time cycles), but did not find any within this session.

Details of Phase One Methods and Results and Continued Quiet Time Analysis

For phase one, we developed and debugged a sequence of MATLAB functions to input recorded data from the brain, and report back how quiet the quietest phase of the proposed clock cycle time was. More precisely, we ultimately developed a function, Find_clock_in_spiket, for which the user would supply three input arguments: (1) delta, an integer, the proposed clock cycle time in the same units of time assumed in the spiket data; (2) K, number of clock cycles per span of data to be analyzed; and (3) spiket(j), an array simply containing the time at which a spike was observed, for all spikes j recorded (in order) in the dataset being analyzed. We also developed a simpler variation, find_clock_in_power, to analyze data of the form xpower(t), representing the time series of a nonnegative measure of signal power, tested on the Si data (Yuan et al., 2015).

To visualize the algorithm and the mathematical issues, it may help to consider a clock cycle of the brain by analogy to the 24-h cycle of a clock. If $K = 100$, we organize the spike data into the hours of 100 days. We calculate a histogram of what time of day the spikes occurred, in each span of 100 days. If activity was quietest, say, between 2 P.M. and 3 P.M. across all 100 days, then we measure "quiet power" for that interval as the sum of activity during that hour, summed over all days in the span, and we compare that later to the average activity across all hours. (Instead of an hour here, we actually looked for a quiet 10 ms interval, and considered all possible intervals starting from the beginning of the cycle, starting from 2 ms after the start of a cycle, and so on). The overall quiet time score for the entire dataset is simply the sum of the quiet time scores for each span of data in that dataset. (The function calculates the number of spans simply as the length of times in the database divided by the time length per span. The length of times in the database is simply the highest and last value of spiket(j), minus the starting time, spiket(1). "Left over" spikes, beyond the last whole span of time, are simply not used in the analysis. The final version of this function handled left-over spikes in a cleaner manner, and changed a few numbers slightly, but all qualitative results were similar to those with the early versions).

Now: what would happen if one applied such an algorithm with the wrong measure of the length of a day? For example, if one were to aggregate hourly electricity consumption data assuming a 25-h day, after about 30 "days" one would expect the measurements to be hopelessly out of synchronization, and the histograms would be flat. For this reason, we initially hoped that use of the quiet time measure would give a very sharp indication of what the clock time is, for those cells which actually are governed by a very regular clock. This would allow us to analyze issues like forward vs. backward signal propagation, in phases two and three, by simply using the precise clock estimates from phase one. A fuzzy estimate of clock time reduces the accuracy of any phase two and phase three analyses which depend on them.

However, the quiet time analysis by itself proved useful only as a screening method. As discussed in Section Summary of methods and findings, it identified four reasonable ranges of possible clock time, each 2 ms wide. Thus, for phases 2 and 3, we performed more intensive analysis of all possible clock times in those four ranges, for all multiples of 0.05 ms, measuring not only quiet power but additional measures. **Table 1** illustrates the final results of phase one, giving the quiet power (number of spikes across proposed quiet periods) for 21 possible choices of clock cycle time in the most promising of the four ranges considered, for the choices $K = 1000$ and $K = 10{,}000$. Notice

that the row of **Table 1** next to the bottom gives the total number of spikes counted in each analysis, and the bottom line gives the average number of spikes one would expect in a random 10 ms interval.

Note that the spikes in the quiet time are quite a bit less than what one would expect in a random "hour of the day," most notably for clock times of 153.4 and 154.5 ms, with $K = 1000$. The actual score of 127 is much less than the null expected score of 313. If the true clock time were, say, 153.35 ms, after 1000 cycles, one would expect missynchronization of $0.05*1000 = 50$ ms from the start of the span to the end; thus, when we only know the clock time to within 0.05 ms, it is quite remarkable to have such a degree of quiet power with K as high as 1000. (It is possible only because the actual quiet time interval may be a bit wider than 10 ms, as Richmond initially suggested, and because a $K = 1000$ implies that the cycles considered within each span are not more than 500 cycles away from the middle of the span). On the other hand, it is clear that these results do not tell us very clearly what the best candidate time is within this 2 ms window. Quiet power was of course higher, relative to the null expectation, in other time ranges.

The MATLAB function Find_quiet_time_in_spiket reports out the quiet power as described above, the total number of spikes actually considered in each quiet time analysis, and the number of complete spans found in the data. (If the user proposes a K too large for the dataset, the function was designed to consider the entire database as one span; however, we never actually tested that feature). It also provides an output array, wherebin, which can be used in debugging, in analysis, and in support for other MATLAB functions as in phase 2 and phase 3. For each of the spans identified in the data, it tells us "at what hour" the quietest period was, how many spikes were found in quiet periods in that span, and how many spikes in total were counted in that span. For example, with $K = 10000$, we only had seven spans in the EE188 data! If the location of the quiet period drifted systematically up or down in the wherebin data, this would suggest that a more refined estimate of the clock time would improve results; however, in our initial exploration of those diagnostics, we found no indications of such systematic drift. Note that we used underbars in the names of all of our MATLAB functions, simply because of how MATLAB works.

Finally, we note that the arithmetic of this analysis was simplified by the fact that the "spiket" variable in the pfc-2 database was based on a recording rate of 20,000 measurements per second (Fujisawa et al., 2008), such that the allowed values of "delta" represented multiples of 0.05 ms. It would have been possible to consider clock times with even more temporal detail, simply by multiplying the entire array "spiket" by 10, so that any clock time which is a multiple of 0.005 ms could be evaluated. That is one of the many variations and extensions which could be considered in future work.

Another extension would be to study whether the clock cycle time is or is not the same for the rat called EE in all the different sessions recorded in the original data (Fujisawa et al., 2008, 2015).

TABLE 1 | Quiet power vs. possible clock time in milliseconds.

$K = 1000$				$K = 10000$		
Proposed clock time in ms	**Neuron 329**	**Neuron 373**	**All pyramids in pfc**	**Neuron 329**	**Neuron 373**	**All pyramids in pfc**
153	4111	1038	141	3160	711	149
153.1	4225	991	133	3215	740	158
153.2	4231	983	145	3254	732	149
153.3	4398	1067	141	3340	762	165
153.4	4452	999	127	3307	741	147
153.5	4437	1007	143	3351	747	149
153.6	4507	1036	150	3372	753	155
153.7	4527	1031	154	3320	765	144
153.8	4551	1067	136	3421	751	159
153.9	4666	1072	129	3423	745	137
154	5323	1142	137	3404	754	149
154.1	4755	1084	143	2831	648	141
154.2	4953	1076	131	2934	695	155
154.3	4898	1101	135	2922	653	141
154.4	4938	1070	159	2911	615	141
154.5	5096	1128	127	2987	704	141
154.6	5059	1100	129	3056	697	137
154.7	5069	1096	140	2975	701	157
154.8	5182	1118	133	3146	716	159
154.9	5121	1129	145	3146	714	149
155	4145	983	153	3176	738	159
Nspikes	92117	22432	4820	54658	12524	2908
Null	5981.623	1456.623	312.987	3549.221	813.2468	188.8312

The underlying theory [Section Alternate neural network models to explain/replicate consciousness (question 1)] suggests that it might be, but experiments with other large shifts in global brain parameters (as with hormones or alcohol) suggest that brains may be able to learn to be robust with respect to them, and hence that natural selection may have permitted them.

Phase 2 Methods, Analysis and Results

Phase 2 and phase 3 both addressed the question: is the real-time data available here consistent with what Richmond said about alternating forward and backward passes in the cortex, as the backpropagation family of neural network models would predict?

The backpropagation family of models does not predict that the sequence of firing in the backwards pass is a perfect mirror image of the sequence in the forwards pass, with a cycle of brain operation. It may be more or less of a mirror image, depending on the degree of fast recurrence in the interneurons, the impact of long-term memories, and the structure of the tasks currently faced by the organism; the importance of current tasks in bringing out different aspects of network structure is illustrated very vividly in (Fujisawa et al., 2008). Nevertheless, the backpropagation models would predict that the backwards pass looks more like a mirror image of the forward task than like a repetition of the forward pass. If calculations were always running forward, both in the first half of a brain cycle and in the second half, one would expect the opposite result: the second half would look more like a repetition than like a mirror image.

The main goal of phases 2 and 3 was to find out which of these two possibilities better fits the data. Phase 2 took a minimal approach, trying to compare the two hypotheses (mirror vs. repetition) without making any assumptions at all about the relations and connections between different neurons. Phase 3 made use of the neuron ID information, and in our view, is much more conclusive and robust. Nevertheless, both types of measure strongly favored the mirror hypothesis over the repetition hypothesis.

Of course, to compare the events in the first half of a brain clock cycle with those in the second half, one must identify the time interval for all of the brain cycles to be analyzed. The phase 2 analysis was performed by a MATLAB function, Test_hypotheses, which started out by calling Find_clock_in_spiket (discussed in Section Details of phase one methods and results and continued quiet time analysis above) to identify the quiet intervals (10 ms wide) in each of the formal clock cycles.

The formal clock cycles which Find_clock_in_spiket starts from are different from the actual brain cycles it locates. In any span of data, Find_clock_in_spiket analyzes K intervals of time, formal clock cycles, and it calculates where the quiet interval is relative to the start of the formal cycle. In Test_hypotheses, a brain cycle is defined as the interval of time stretching from the middle of the quiet period in one formal cycle, to the next quiet period in the next formal cycle. Since there are K formal clock cycles in any span of data, this yields K-1 brain cycles. For each brain cycle, we may define t_- as the start time of the cycle, t_+ as the end time of the cycle, and t_0 as the exact mid-point between the two. We define the "AM" period of the brain cycle as the interval between t_- and t_0. We define the "PM" period as the interval between t_0 and t_+. Both in phase 2 and in phase 3, our goal was to answer the question: "Is the sequence of neurons firing in the PM period more like a mirror image of the sequence in the AM period, or like a repetition, over the entire dataset?"

In phase 2, we calculated two measures of error for each of the two hypotheses (mirror vs. repetition), for every neuron which fired at least once both in the AM part of the brain cycle and in the PM part of the brain cycle.. We calculated these four measures for each of the identified brain cycles, and simply added up total error over all brain cycles to generate the final error scores. We weighted the error by the number of spikes, because this reflects the greater importance of neurons and times of greater activity. As in phase one, we used the burst-filtered data from the pyramid cells to establish the clock intervals, but all of the remaining analysis used the entire burst-filtered dataset of all recorded neurons in the prefrontal cortex.

For each active neuron, in each clock cycle, we began by find out $t_{AM}{}^+$, the time of the last spike in the AM period, and $t_{AM}{}^-$, the time of the earliest spike in the AM period, and then $t_{PM}{}^-$ and $t_{PM}{}^+$. We also calculated N, the total number of spikes for that neuron in the AM and PM periods.

The first two measures of error tested whether the interval between first and last spike for that neuron in the PM matches a mirror image (through t_0) of the first and last spike in the AM, or whether it matches a repetition. We calculated the error in the mirror hypothesis as:

$$e\downarrow = N^*(|(t_0 - t_{AM}{}^+) - (t_{PM}{}^- - t_0)| + |(t_0 - t_{AM}{}^-) - (t_{PM}{}^+ - t_0)|) \quad (1)$$

We calculated the error in the repetition hypothesis as:

$$e\uparrow = N^*(|(t_0 - t_{AM}{}^+) - (t_+ - t_{PM}{}^+)| + |(t_0 - t_{AM}{}^-) - (t_+ - t_{PM}{}^-)|) \quad (2)$$

The next two measures were essentially the same, except that we compared the midpoints of the AM and PM intervals.

We ran the Test_hypotheses function using all four of the intervals discussed above for the possible clock time, with $K = 1000$. For the most plausible interval, from 153 to 155 ms, we also tried $K = 100$. We divided the error scores by 1,000,000 and rounded to the nearest integer, so as to provide a nice table, for each of the 41 clock times considered in each interval, giving a new estimate of quiet power, and results on each of the four error measures.

For the interval between 153 and 155 ms, every choice of clock cycle gave a score of 27 or 28 for e↓, and every choice of clock cycle gave a score of 36,37 or 38 for e↑, with $K = 100$ and with $K = 1000$. In other words, the mirror hypothesis was notably preferred over the repetition hypothesis for all choices of clock time and both choices of K. The mirror hypothesis was also preferred in a uniform way in the other three intervals: for 134–136 ms, e↓ was 23 or 24, vs. e↑ of 31, 32, or 33; for 144–146 ms, it was 25,26, or 27 vs. 33, 34, or 35; for 132–134 ms, it was 23 or 24 vs. 31 or 32. Likewise, for the midpoint error measure, in the 153–155 ms interval, it was 12 vs. 17 (with only two cases of 16 and 17). The corresponding quiet time error with $K = 1000$

was 134 at 153.4 ms, lower than the quiet time error at other candidates in that interval, and notably lower than the quiet time error in any of the other three intervals.

In summary, the phase two measures of hypothesis error did favor the mirror hypothesis over the repetition hypothesis, for all choices of possible clock time. 153.4 ms emerged a bit more clearly as the best estimate of the underlying clock time. We believe that the results from phase three are more robust and more convincing than those of phase two, but it is even more notable that two entirely different ways of evaluating the clock time and the mirror hypothesis led to the same conclusions.

Phase 3 Methods and Results

The phase 3 analysis was performed using a MATLAB function, Test_sequence_and_inertia, which provides three sets of statistical measures for each user-supplied choice of clock cycle time and K. First, it calls Find_clock_in_spiket, exactly as Test_hypotheses does, and outputs a quiet power measure, exactly the same as the quiet time measure calculated by Test_hypotheses. It also provides four new measures of error for the mirror hypothesis and the repetition hypothesis. Finally, it provides three useful measures of inertia or autocorrelation, which provide another way to evaluate whether the proposed clock cycle time is the correct one.

As in phase two, the four measures of error are calculated for each brain cycle, and added across all brain cycles and scaled, to get the total measures of error for the mirror hypothesis e↓ and for the repetition hypothesis e↑. Within each brain cycle, we first create a list of active neurons—neurons which fired both in the AM and in the PM. If there was only one active neuron, or none, this brain cycle is skipped, because there are no AM and PM sequences to be compared. Next, for each active neuron, we calculate the two simple averages, $(t_{AM}-+ t_{AM}+)/2$ and $(t_{PM}-+ t_{PM}+)/2$, which indicate when this neuron fired, both in AM and PM. We sort the neurons according to when they fired in the AM and when they fired in the PM. The unweighted measure used for e↑ is simply the inversion number comparing these two permutations; the inversion number (Foata, 1968) is a widely used standard measure for comparing the similarity of two permutations. The unweighted measure used for e↓ is the inversion number comparing the AM sequence and the reverse of the PM sequence. The weighted versions in each brain cycle are equal to the unweighted versions multiplied by the product of the total number of spikes in all active neurons in the AM and the total number in the PM.

The results of this analysis for the most important case are shown in **Table 2** and in **Figure 1**.

To understand the meaning of the e↓ and e↑ error measures, it may help consider two examples where 5 neurons (numbered N1–N5) fire in the following sequences within a clock cycle, and where t0 is the mid-point of the time cycle:

case A: N1, N2, N3, N4, N5, t0, N1, N2, N3, N4, N5

case B: N1, N2, N3, N4, N5, t0, N5, N4, N3, N2, N1

In case A, the unscaled value of e↑ is zero, because the sequence of firing after the mid-time is identical to the sequence before; however, the unscaled value of e↓ is 10 (4+3+2+1), because it takes 10 swaps to make the sequence after t0 match the mirror image of the sequence before. Case B is the opposite. The four columns on the right of **Table 2** are all sums of e↓ or e↑, unweighted or weighted by the level of neuron activity in the time cycle, scaled by the same factor for convenience in printing.

The first of the three new inertia measures in **Table 2** is simply the number of neurons which were added or dropped out from the list of active neurons, from one brain cycle in a span to the next. The second is the inversion number comparing the sequence of neurons firing in the AM, for those neurons which are active both in one brain cycle and the next. The third is the same, for the PM sequence.

DISCUSSION

Neural network models in the large family of backpropagation-based models have already performed well in challenging applications demanding an ability to replicate the kind of abilities brains have proven possible from pattern recognition to intelligent control, with a strong foundation in the technology disciplines which specialize in designs capable of addressing such tasks in a highly effective manner. There is every reason to believe that hybrid systems, effectively combing the capability of backpropagation networks and other types of network more common in computational neuroscience, could do still better in allowing us to replicate and understand the higher-order learning capabilities which drive mass action in the mammal brain, if we could make contact between the world of functional, mathematical neural network models and the world of empirical real-time data in neuroscience.

This work has done a quick initial evaluation of whether two key ideas in backpropagation might actually fit empirical real-time data from the brain, using a series of new quantitative measures which directly capture two of the most important predictions of that type of model—the prediction of a regular clock cycle, and of an alternation of forward and backward passes of calculation. We hope that these new measures inspire more work to address the many questions which flow from considering this new class of models of brain functioning.

A few of these questions and opportunities for future research were already discussed above, but there are many more. For example, it would be interesting to revisit the work of Fujisawa et al. (2008), and see what the cross-impacts and networks look like when the full database is partitioned into AM data and PM (and leftovers at the boundaries between spans). It would be interesting to revisit the work on models to predict neural signals over time, not only in burst-sorted and spike sorted data, but even in the original raw data, when we have the ability to partition that data into AM and PM, and to use the clock cycle here to use seasonal adjustment types of method as in time-series analysis (Box and Jenkins, 1970); see (Werbos, 1994) and (Werbos, 2010) for the extension of such time-series analysis methods to the multivariate and nonlinear cases, respectively. Because financial market data, like spike data, tend to involve discrete events and irregular kinds of statistics, it is quite possible that the approach used in (Werbos, 2010), drawing on Peters

TABLE 2 | Phase 3 results for 153–155 ms, with $K = 1000$.

Clock time in ms	Measures of clock accuracy				Measures of mirror vs. repetition			
	Quiet power	Change of neurons	Sequence change		Unweighted		Weighted	
			AM	PM	e↓	e↑	e↓	e↑
153	151	3689	13	11	300	494	20463	29868
153.05	149	3699	19	21	314	502	23197	32001
153.1	141	3671	20	19	296	479	20355	30066
153.15	154	3709	19	11	317	493	21875	28649
153.2	153	3669	19	21	309	491	21129	30163
153.25	135	3590	24	19	315	486	22258	31344
153.3	150	3643	25	18	316	456	22863	28318
153.35	136	3530	19	19	308	463	21739	30052
153.4	**134**	**3620**	**11**	**17**	**292**	**439**	**21967**	**27835**
153.45	137	3650	18	14	305	481	22080	30627
153.5	152	3659	16	22	298	515	23334	30109
153.55	143	3628	20	22	321	497	23171	31574
153.6	160	3611	23	20	310	474	23090	28241
153.65	140	3625	26	25	319	487	22364	29398
153.7	164	3673	22	26	311	511	22154	30875
153.75	154	3610	20	22	310	501	22823	30298
153.8	146	3573	14	19	287	461	21342	29518
153.85	153	3648	17	13	296	485	21126	32104
153.9	142	3541	18	19	327	465	23432	28995
153.95	150	3659	17	19	329	498	23760	31858
154	150	3659	22	20	315	511	22280	32133
154.05	143	3620	20	10	296	498	23683	30275
154.1	152	3696	18	19	304	490	23563	29190
154.15	138	3660	23	19	306	460	23271	28663
154.2	141	3567	15	17	284	448	21117	28161
154.25	135	3624	22	19	290	503	22837	29861
154.3	144	3660	27	25	312	478	22404	31269
154.35	151	3603	28	24	294	523	21108	32955
154.4	168	3619	21	22	312	496	20953	32755
154.45	150	3660	19	16	309	520	23933	31771
154.5	135	3635	21	23	297	508	22120	33437
154.55	137	3646	19	17	307	466	21698	29341
154.6	142	3582	18	19	337	477	24138	29713
154.65	156	3597	23	22	303	500	22021	30999
154.7	149	3633	16	16	297	503	21210	32023
154.75	143	3625	20	17	314	479	22250	29081
154.8	142	3665	25	18	321	492	23359	31363
154.85	151	3676	20	18	327	511	25244	30280
154.9	155	3611	14	18	296	460	21002	30199
154.95	160	3639	18	20	312	485	22848	30045
155	159	3602	21	16	309	496	22264	33578

(1996), could yield new insights in this context. And of course, these MATLAB functions, developed to be very general in nature, could be applied to other databases.

The effort to understand the mathematical and computational principles underlying intelligence in the mammal brain is perhaps one of the two most important and fundamental challenges to all of basic science for the coming century. (The other is the continuing quest to understand the fundamental laws of physics). It is also a key motivation for society as a whole to be interested. It is hoped that this work will inspire new work which fully rises to that grand challenge.

A reviewer of this paper has raised an interesting question: can we try to imagine new classes of model, beyond those discussed

in Section Alternate neural network models to explain/replicate consciousness (question 1), which would also be functional in information processing, but would treat time in a different way and fit our results from a very different basis? In fact, the work reported in Werbos and Dolmatova (2016); Werbos (2016a,b) does begin to suggest more radical types of model and technology which may or may not be relevant to understanding the basic rodent brain which we see in the laboratory.

Human cultures disagree violently at times about the nature of human consciousness, beyond the level of what we share with mice and rats. While we may hold different viewpoints (Werbos, 2012, 2015; Davis, 2016) on that larger question, beyond the reach of consensus science at present, we hope that we can agree that better understanding of what we share with mice and rats is an important steppingstone to understanding how to achieve the highest potential which we as humans can attain.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We express deep thanks to Dr. Ludmilla Dolmatova, who at many times in this work intervened to solve technical problems in juggling multiple computers which were beyond our capabilities. Without her help, the computer runs reported in **Tables 1, 2** would simply never have happened. As discussed in Selection of real-time multielectrode data to test for clock cycles and backwards passes, we are also grateful for help from Drs. Fujisawa and Si in providing the data, and to the wonderful leadership behind the website at http://crcns.org.

REFERENCES

Anderson, J., and Rosenfeld, E. (eds.). (1998). *Talking Nets*. Cambridge, MA: MIT Press.

Baars, B. J. (2016). "How does the cortex know? A walk through Freeman neuruodynamics," in *Cognitive Phase Transition in the Cerebral Cortex - Enhancing the Neuron Doctrine by Modeling Neural Fields*, eds R. Kozma and W. Freeman (Springer International Publishing), 117–125.

Bear, M. F., Connors, B. W., and Paradiso, M. A. (2007). *Neuroscience*: *Exploring the Brain*, *3rd Edn.* Baltimore, MD; Philadelphia, PA: Lippincott Williams and Wilkins.

Box, G. E. P., and Jenkins, G. M. (1970). *Time-Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. doi: 10.1038/nrn3241

Davis, J. J. J. (2016). A brief introduction to the neuro-genetics of spirituality towards a systemic peace propagation model. *Scientific GOD Journal* 7, 261–338.

Foata, D. (1968). On the Netto inversion number of a sequence. *Proc. Am. Math. Soc.* 19, 236–240. doi: 10.1090/S0002-9939-1968-0223256-9

Freeman, W. J. (1975/2004). *Mass Action in the Nervous System*. New York, NY: Academic.

Kozma, R., and Freeman, W. J. (eds.). (2016). "Introduction – on the languages of brains," in *Cognitive Phase Transition in the Cerebral Cortex - Enhancing the Neuron Doctrine by Modeling Neural Fields* (Springer International Publishing), 3–13.

Fujisawa, S., Amarasingham, A., Harrison, M. T., and Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. Neurosci.* 11, 823–833. doi: 10.1038/nn.2134

Fujisawa, S., Amarasingham, A., Harrison, M. T., Peyrache, A., and Buzsáki, G. (2015). *Simultaneous Electrophysiological Recordings of Ensembles of Isolated Neurons in Rat Medial Prefrontal Cortex and Intermediate CA1 Area of the Hippocampus During a Working Memory Task.* Available online at: https://crcns.org/data-sets/pfc/pfc-2

Grossberg, S. (1971). Pavlovian pattern learning by nonlinear neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 68, 828–831.

Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H., and Buzsáki, G. (2000). Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* 84, 401–414.

Hebb, D. O. (1949). *The Organization of Behavior*. New York, NY: Wiley.

Heck, D. H., McAfee, S. S., Liu, Y., Babajani-Feremi, A., Rezaie, R., and Freeman, W. J. (2016). Cortical rhythms are modulated by respiration. Available online at: http://biorxiv.org/content/early/2016/04/16/049007

Heller, J., Hertz, J. A., Kjaer, T. W., and Richmond, B. J. (1995). Information flow and temporal coding in primate pattern vision. *J. Comput. Neurosci.* 2, 175–193. doi: 10.1007/BF00961433

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (eds.). (1983). *Understanding Robust and Exploratory Data Analysis*, Vol. 3. New York, NY: Wiley.

Kahana, M. J., Seelig, D., and Madsen, J. R. (2001). Theta returns. *Curr. Opin. Neurobiol.* 11, 739–744. doi: 10.1016/S0959-4388(01)00278-1

Kozma, R., Davis, J. J., and Freeman, W. J. (2012). Synchronized minima in ECoG power at frequencies between beta-gamma oscillations disclose cortical singularities in cognition. *J. Neurosci. Neuroeng.* 1, 13–23. doi: 10.1166/jnsne.2012.1004

Lewis, F. L., and Derong, L. (2012). *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: Wiley.

Llinas, R. R., and Roy, S. (2009). The 'prediction imperative'as the basis for self-awareness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1301–1307. doi: 10.1098/rstb.2008.0309

Macke, J. H., Buesing, L., Cunningham, J. P., Byron, M. Y., Shenoy, K. V., and Sahani, M. (2011). "Empirical models of spiking in neural populations," in *Advances in Neural Information Processing Systems,* 1350–1358.

Minsky, M., and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

National Science Foundation (2008). *Emerging Frontiers in Research and Innovation 2008*. Available online at: http://www.nsf.gov/pubs/2007/nsf07579/nsf07579.pdf, section on COPN

Ng, A., Dan, Y., Boyden, E., and LeCun, Y. (2008). *EFRI-COPN Deep Learning in the Mammalian Visual Cortex*. Available online at: http://www.nsf.gov/awardsearch/howAward?AWD_ID=0835878&HistoricalAwards=false

Nicolelis, M. A., Baccala, L. A., Lin, R. C., and Chapin, J. K. (1995). Sensorimotor encoding by synchronous neural ensembly activity at multiple levels of the somatosensory system. *Science* 268, 1353.

Peters, E. E. (1996). *Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility,* Vol. 1. New York, NY: John Wiley and Sons.

Rav, Y. (2002). Perspectives on the history of the cybernetics movement: the path to current research through the contributions of norbert wiener, warren mcculloch, and john von neumann. *Cybern. Syst.* 33, 779–804. doi: 10.1080/01969720290040830

Rosenblith, W. A. (ed.). (1961). *Sensory Communication*. Cambridge, MA; New York, NY: MIT Press and Wiley.

Rossant, C., Kadir, S. N., Goodman, D. F. M., Schulman, J., Hunter, M. L. D., Saleem, A. B., et al. (2016). Spike sorting for large, dense electrode arrays. *Nat. Neurosci.* 19, 634–641. doi: 10.1038/nn.4268

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). "Learning internal representations by error propagation," in *Parallel Distributed Processing*, Vol. 1, eds D. Rumelhart and J. McClelland (Cambridge, MA: MIT Press), 318–362.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Smirnova, T., Laroche, S., Errington, M., Hicks, A., Bliss, T., and Mallet, J., et al. (1993). Transsynaptic expression of a presynaptic glutamate receptor during hippocampal long-term potentiation. *Science* 262, 430–436; Also see Stuart, G., Spruston, N., Sakmann, B., and Hausser, M. (1997). Action potential initiation and backpropagation in neurons of the mammalian central nervous system. *Trends Neurosci.* 20, 125–131.

Sugihara, I., Lang, E. J., and Llinas, R. (1993). Uniform olivocerebellar conduction time underlies Purkinje cell complex spike synchronicity in the rat cerebellum. *J. Physiol.* 470, 243. doi: 10.1113/jphysiol.1993.sp019857

Werbos, P. (2010). *Mathematical Foundations of Prediction under Complexity, Erdos Lectures/Conference 2010*. Available online at: http://www.werbos.com/Neural/Erdos_talk_Werbos_final.pdf

Werbos, P. (2012). Neural networks and the experience and cultivation of mind. *Neural Netw.* 32 86–95. doi: 10.1016/j.neunet.2012.02.026

Werbos, P. (2014). "From ADP to the brain: foundations, roadmap, challenges and research priorities," in *Proceedings of the International Joint Conference on Neural Networks*, (IEEE). Available online at: http://arxiv.org/abs/1404.0554

Werbos, P. (2015). *Links Between Consciousness and the Physics of Time, International IFNA -ANS Journal "Problems of nonlinear analysis in engineering systems"*. Available online at: http://www.kcn.ru/tat_en/science/ans/journals

Werbos, P. J. (1990). 2.1. Econometric techniques: t*heory versus practice.* Energy 15, 213–236. doi: 10.1016/0360-5442(90)90085-G

Werbos, P. J. (1994). *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting,* Vol. 1. New York, NY: John Wiley and Sons.

Werbos, P. J. (2006). "Backwards differentiation in AD and neural nets: past links and new opportunities," in *Automatic Differentiation: Applications, Theory, and Implementations*, eds M. Bücker, G. Corliss, U. Naumann, P. Hovland, and B. Norris (Berlin; Heidelberg: Springer), 15–34.

Werbos, P. J. (2009). Intelligence in the brain: a theory of how it works and how to build it. *Neural Netw.* 22, 200–212. doi: 10.1016/j.neunet.2009.03.012

Werbos, P. J. (2016a). "How can we ever understand how the brain works?" in *Cognitive Phase Transition in the Cerebral Cortex: Enhancing the Neuron Doctrine by Modeling Neural Fields,* eds R. Kozma and W. Freeman (Springer International Publishing), 217–228.

Werbos, P. J. (2016b). "New technology options and threats to detect and combat terrorism," in *Proceeding of NATO Workshop on Predetection of Terrorism, NATO/IOS*, eds Sharan, Gordon and Florescu (Amsterdam).

Werbos, P. J., and Dolmatova, L. (2016). Analog quantum computing (AQC) and the need for time-symmetric physics. *Quantum Inf. Process.* 15, 1273–1287. doi: 10.1007/s11128-015-1146-2

White, D., and Sofge, D. (eds.). (1992). *Handbook of Intelligent Control.* Chapters 10 and 13, ed Van Nostrand. Available online at: www.werbos.com/Mind.htm

Yuan, Y., Mao, H., and Si, J. (2015). Cortical neural responses to previous trial outcome during learning of a directional choice task. *J. Neurophysiol.* 113, 1963–1976. doi: 10.1152/jn.00238.2014

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 15 November 2016
doi: 10.3389/fnsys.2016.00084

Physics of the Mind

Leonid I. Perlovsky[1,2]*

[1] MGH/HST Martinos Center for Biomedical Imaging, Medical School, Harvard University, Cambridge, MA, USA,
[2] Psychology and Engineering Departments, Northeastern University, Boston, MA, USA

Is it possible to turn psychology into "hard science"? Physics of the mind follows the fundamental methodology of physics in all areas where physics have been developed. What is common among Newtonian mechanics, statistical physics, quantum physics, thermodynamics, theory of relativity, astrophysics... and a theory of superstrings? The common among all areas of physics is a methodology of physics discussed in the first few lines of the paper. Is physics of the mind possible? Is it possible to describe the mind based on the few first principles as physics does? The mind with its variabilities and uncertainties, the mind from perception and elementary cognition to emotions and abstract ideas, to high cognition. Is it possible to turn psychology and neuroscience into "hard" sciences? The paper discusses established first principles of the mind, their mathematical formulations, and a mathematical model of the mind derived from these first principles, mechanisms of concepts, emotions, instincts, behavior, language, cognition, intuitions, conscious and unconscious, abilities for symbols, functions of the beautiful and musical emotions in cognition and evolution. Some of the theoretical predictions have been experimentally confirmed. This research won national and international awards. In addition to summarizing existing results the paper describes new development theoretical and experimental. The paper discusses unsolved theoretical problems as well as experimental challenges for future research.

Keywords: physics of the mind, neuroscience, cognition, dynamic logic, knowledge instinct, aesthetic emotions, consciousness, beautiful

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research, Inc., USA

Reviewed by:
Angelo Cangelosi,
Plymouth University, UK
Robin W. Wilkins,
Joint School for Nanoscience and Nanoengineering Gateway MRI Center, USA

***Correspondence:**
Leonid I. Perlovsky
lperl@rcn.com

Received: *14 August 2016*
Accepted: *17 October 2016*
Published: *15 November 2016*

Citation:
Perlovsky LI (2016) Physics of the Mind. Front. Syst. Neurosci. 10:84. doi: 10.3389/fnsys.2016.00084

WHAT IS PHYSICS OF THE MIND?

The common to all areas of physics is a methodology that first, concentrates on finding few fundamental laws and their mathematical formulations; second, a mathematical theory developed from these few "first principles" that explains a vast area of knowledge without contradicting known facts; third, makes unexpected theoretical predictions, which could be verified experimentally, and actual experimental verifications which confirm or disconfirm the theory.

This paper briefly summarizes previously developed aspects of the physical theory of the mind and presents new developments. It discusses the first principles identified by a number of leading neuroscientists, mathematical methods suitable for their modeling, perception and cognition mechanisms based on these principles, the mechanisms of an approximate mental hierarchy. The physics of the mind and the related mathematical theory are extended toward the dual hierarchy of interactions between cognition and language, high cognition including emotions of the beautiful, a controversial idea of the meaning of life, as well as functions of these high principles in cognition.

It is further extended toward emotional prosody of speech as well as cognitive functions and the reasons for evolution of musical emotions from animal cries to Bach and Justin Bieber.

This theory does not contradict existing knowledge, explains psychological facts that have been poorly understood previously, and has made a number of unexpected experimentally verifiable predictions. The paper discusses theoretical predictions that have been experimentally confirmed (or tentatively confirmed). Among these confirmed predictions are mechanisms of perception and cognition, mathematical model of the mind that overcomes computational complexity, interaction between cognition and language, the nature and mechanisms of the beautiful and the meaning of life, as well as the cognitive functions and reasons for origin and evolution of musical emotions. Computational complexity interfered with modeling the mind, artificial intelligence and machine learning since the 1960s, a mathematical theory overcoming this difficulty is described. The paper presents physics of the mind that mathematically models psychological mechanisms at the functional level. Self-organization processes are related to thermodynamics and informational theories. The functional theory has been partly related to neural mechanisms, and some of these relations have been experimentally confirmed. The paper discusses predictions, which open vast areas for future research.

FUNDAMENTAL PRINCIPLES OF THE MIND-BRAIN

This section describes several fundamental principles of the mind-brain, later the paper describes mathematical models for some of them.

Concepts are a mechanism of understanding objects, events, and abstract ideas. Their contents are stored in neural representations. In the processes of perception concepts project their contents to the visual cortex to match sensory projections. Concepts are also called mental models of events and objects. In a simple case a concept is memory. The analogy with models is literal and neural representations are also called mental models. Mathematical models of concept mechanisms is discussed in (Perlovsky et al., 1997, 2011; Perlovsky, 2006a). Proof of detailed theoretical predictions of the mechanism in experimental neuro-imaging, including detailed descriptions of the brain regions involved was obtained by Bar et al. (2006) and Kveraga et al. (2007). The moment of perception is a match of these images.

Instincts are ancient mechanisms of survival. This paper follows Grossberg and Levine (1987) theory, which has been modeled mathematically and is appropriate for developing physics of the mind. This theory considers the instinct mechanism resembling "neural sensors that measure vital parameters important for functioning and survival" (Grossberg and Levine, 1987; Perlovsky, 2006a). For example, a low blood glucose level specifies an instinctual need for food. Measurements of glucose level sensors and the requirement to keep glucose level within bounds is a mechanism of instinct.

Emotions designate various mechanisms which are surveyed in a number of publications. Following Grossberg and Levine (1987) theory of drives and emotions the mechanism of emotions are neural signals connecting instinctual and conceptual brain regions. Emotions, emotional neural signals, related states and feelings communicate instinctual needs to conceptual recognition-understanding mechanisms. Their function is to motivate behavioral and conceptual representation-models, which correspond to objects or events that can potentially satisfy instinctual needs, so that these models receive preferential attention and processing resources within the brain. Thus emotions evaluate concepts for the purpose of instinct satisfaction. Emotional signals and related states of the mind are felt as emotional feelings.

Psychological research of emotions is usually limited to basic emotions, which are related to satisfaction of bodily instinctual needs, named by specific words, and limited in number to a few different emotions. There are only few basic emotions; they are a small part of our emotional abilities, the most ancient and noticeable ones. Our higher cognitive abilities involve many "continuous" emotions, which include aesthetic emotions discussed later, related to knowledge, including processes of learning, emotions in the voice prosody, emotions of cognitive dissonances, as well as musical emotions described later.

Behaviour is governed by several mechanisms. The most interesting for the initial development of physics of the mind is the mechanism of behavioral concepts; it is similar to the mechanism of cognitive concepts discussed above with the difference that behavioral concepts govern behavior. Most of human behavior occurs in the mind, it is directed at improving concepts, understanding, and knowledge.

Cognitive hierarchy is an approximately hierarchical structure (Kosslyn, 1980; Grossberg, 1988) of mental models and aesthetic emotions (discussed later) extending from sensory-motor representations at the bottom of the hierarchy, higher up to concepts of objects, contexts, situations, and many levels of abstract concepts, to the top of the hierarchy, which content will be elucidated in the paper. This description is not quite accurate, especially for neural mechanisms below objects; one can look for details e.g., in Grossberg publications, but it is adequate for higher levels. The hierarchy is functional, it is not organized from the bottom to the top along a specific geometric axis. Processes of understanding involve interactions among models at lower and higher levels. In these interactions higher level models are improved for better correspondence to lower level models; a higher level model unify lower level ones for creating a more abstract and general concept. The interaction is two-way: lower level models are also improved for better match to the details of the situation (lower level models) and for better matching the top level one. Neural signals involved in these interactions are called bottom-up, BU, and top-down, TD, signals.

The knowledge instinct, KI, is a special instinctual mechanism related to knowledge acquisition and improvement of concept-models (Perlovsky, 2001, 2006a, 2007b, 2008b; Perlovsky et al., 2011). Its model is an extension of Grossberg and Levine (1987) theory of bodily instincts to cognition. KI is similar to other instincts, it involves sensor-like neural mechanisms that

measures similarities between patterns in sensor data and mental models, or more generally between BU and TD signals. As discussed later, in humans and other higher animals mental models are vague, and matching them to objects and situations requires adapting them to BU signals. KI drives this adaptation. No perception or cognition would be possible without KI. For this reason KI is a most important instinct. KI is not related to bodily needs but to "higher" needs for cognition and in this sense it can be termed a higher instinct.

Aesthetic emotions are related to satisfaction of KI and they are modeled mathematically by changes in KI. This theoretical prediction have been experimentally confirmed (Perlovsky et al., 2010; Schoeller and Perlovsky, 2015, 2016). Relation of aesthetic emotions to knowledge was established by Kant (1790), although he could not formulate his thoughts with mathematical precision, the adequate mathematics did not exist at the time. His thoughts have not been understood by his followers. Aesthetic emotions are related to learning and understanding at every level of the mental hierarchy, understanding is pleasant (Perlovsky, 2001, 2006a; Perlovsky et al., 2011; Schoeller and Perlovsky, 2015, 2016). But we do not relate understanding at lower levels to the beautiful. Later this paper connects aesthetic emotions to the emotions of the beautiful.

Perception of objects refers to recognition of more or less familiar objects, and sometimes to noticing unfamiliar objects. Visual perception involves neural projections of retinal images to the visual cortex (BU). At the same time existing models (of expected objects) project TD signals to the visual cortex (TD). Driven by KI (or in other words, motivated by aesthetic emotions) TD and BU projections of models on the visual cortex are modified to match each other. When match is successful, perception occurs (Grossberg, 1988). This process is modeled mathematically (Perlovsky et al., 2011), and detailed predictions of this model are experimentally confirmed (Bar et al., 2006; Kveraga et al., 2007).

The above principles describe self-organization of conceptual and emotional mechanisms of perception and cognition. They encompass the mechanisms of imagination, intuition, planning, conscious, unconscious, and others, including higher abilities and aesthetic emotions (Perlovsky, 2001, 2006a, 2010d; Perlovsky et al., 2011). Most brain operations are unconscious, for example, individual neuronal firings usually can never be accessed by consciousness. This paper refers to the brain-mind neural processes that are not accessible to consciousness as being unconscious, and there are various degrees of unconsciousness. Some processes could never become conscious; others can be accessed by consciousness with significant mental effort, as in creative processes; still others become conscious under changing circumstances without special effort. Many theoretical predictions have been confirmed in experiments (Kosslyn, 1980; Grossberg, 1988; Perlovsky et al., 2010; Schoeller and Perlovsky, 2015, 2016).

Vague Representations. Mental models are not crisp like visual perceptions. A simple experiment can prove this. Look at an object in front of you and then recollect this object with closed eyes. This visual imagination is vague, one cannot recollect even a simple everyday object in all details with closed eyes. Imagination is a TD neural projection from memory to the visual cortex. Vagueness of imaginations is a consequence of vagueness of mental models (Perlovsky, 2016b).

This property of models is fundamental for perception. The reason is that an object would never appear exactly same as during previous perceptions; angles, lightings, surrounding objects would always be different. Therefore previously remembered objects would not match new object projections from retina to visual cortex. Attempts in artificial intelligence (AI) to recognize objects by matching sensory images to previous images took many years and resulted in failures. The number of possible modifications of previous images to match a new image, are combinatorially large. This number is larger than all interactions of all elementary particles in the Universe, therefore the resulting complexity is unsolvable. The problem is called combinatorial complexity, CC (Perlovsky, 1998). Vague models avoid a need to consider combinations. The vague-to-crisp process is fundamental for self-organization, perception and cognition; vague representations and processes are not conscious, possibly for this reason vagueness of representations has not been appreciated by psychologists and mathematicians modeling the mind, and this is the reason why mind processes have not been mathematically modeled and understood in artificial intelligence (Perlovsky, 2001; Russell and Norvig, 2010).

Dynamic logic, DL, is a mathematical technique modeling the brain-mind mechanism of matching vague models to crisp projections from the retina (Perlovsky, 2001, 2006a, 2013c; Perlovsky et al., 2011). Adequacy of DL vague-to-crisp process has been experimentally proven in (Bar et al., 2006; Kveraga et al., 2007). M. Bar and colleagues proved that the initial state of models is vague. The process "from vague to crisp" until models match retinal projections take approximately 150 ms. These includes many neuronal operations: about 10 ms per firing of a neuron, while tens of thousands of neurons are participating in parallel. The initial part of this process cannot be accessed by consciousness, vague models and processes are not accessible to consciousness. Conscious perceptions occur only at the moment of model-projections matching object-projections from the retina.

DL is a process-logic, it avoids logical states until the very end of the DL-process (Perlovsky, 2006a, 2013c). This is essential because CC has been shown to be equivalent to Gödelian incompleteness of logic when applied to finite systems, such as computers or brains (Perlovsky, 2013d). It is interesting to note that the founder of logic, Aristotle explained to his students that logic is needed to argue what has been already understood, but not for understanding of new phenomena, and logic should not be used for understanding working of the mind. Aristotelian theory of the mind is similar to DL, the mind understand the world by using forms that today we call models or representations. Initial states of the forms are potentialities, which are not logical. In the process of "the mind meeting matter", which today we call interactions of TD and BU signals, forms become actualities, which are logical states.

This section summarizes several fundamental principles of the mind: instincts, emotions, concepts, cognitive hierarchy,

the knowledge instinct, aesthetic emotions, perceptions, vague model-representations, dynamic logic. Not all of these principles are independent, e.g., KI and aesthetic emotions are extensions of general principles of instincts and emotions, vague representations are a part of dynamic logic; this repetition is justified by importance of the correspondence principles. Perception is a mechanism explained from fundamental principles. Mathematical foundations of these principles have been discussed, mathematical details will be presented later as well as few other fundamental principles. Identification of few fundamental principles is a first step toward developing physics of the mind.

THE BEAUTIFUL AND MEANING OF LIFE

The mind mechanisms are organized into an approximate hierarchy of concepts and aesthetic emotions. Cognitive hierarchy is illustrated in **Figure 1**.

The hierarchical organization of cognition and related brain structures are reviewed in (Badre, 2008). The hierarchy evolved for the purpose of developing more abstract and general concept-models (Perlovsky, 2006a). Consider a perception-cognition process of an everyday situation, e.g., a professor office. The knowledge instinct first drives the mind to perceive and understand objects in the office: chairs, computers desks, shelves, books... Animals also understand individual objects. Next, the knowledge instinct drives us to understand the concept "office" as a unity of objects. A mathematical model of this process was developed in (Ilin and Perlovsky, 2010; Perlovsky and Ilin, 2010a,b). The higher level abstract concepts we understand due to corresponding concept-models, such as "office." Similarly, we understand a "concert hall," and other situations by using higher-level concepts that our mind evolved for this purpose.

I will repeat the word purpose; every higher-level concept and its mechanisms evolved in individual learning as well as in genetic and cultural evolution with a purpose to be able to transform many lower-level concepts into a unified meaning. In this understanding lower-level concept-models acquire higher-level meanings and develop a more abstract understanding than lower-level meanings. This way our understanding of the surrounding world evolves from a "book" to an "office," to a "university", to an "educational system," and so on... to models near the top of our mental hierarchy. These "top" models "attempt to make sense, to understand the meaning of our entire experience. We understand-perceive-feel them as related to the meaning and purpose of our lives" (Perlovsky, 2006a, 2010c).

Models at every level unify lower level models, for example a situation-model symphony hall unifies lower level object-models: chairs, listeners, scene, etc. Continuing this argument to the top of the hierarchy, one concludes that models at the top unify the entire life experience. These top representations are understood as the meaning and purpose of life. As discussed, even lower level concepts are vague. Abstract concepts built on top of many levels of vague models are even vaguer (Perlovsky, 2011c), therefore the meaning and purpose of life are not finite exactly defined ideas like objects perceived with opened eyes. The next section discusses why sometimes it may seem that we can crisply formulate these ideas.

Learning of models at every level is driven by KI operated at that level, or in other words is motivated by aesthetic emotions at this level. At lower levels aesthetic emotions could be below consciousness. At the top of the hierarchy the highest aesthetic emotions are emotions of the beautiful, sometimes these emotions could be strong and even produce physiological effects, aesthetic chills (Perlovsky, 2002, 2006a, 2007c, 2008a, 2010b,c; Mayorga and Perlovsky, 2008; Schoeller and Perlovsky, 2015, 2016).

Let me emphasize that defining emotions of the beautiful as the highest aesthetic emotions (that is aesthetic emotions near the top of the mental hierarchy) corresponds to a well-accepted human intuition. Kant (1790) has been the first who related

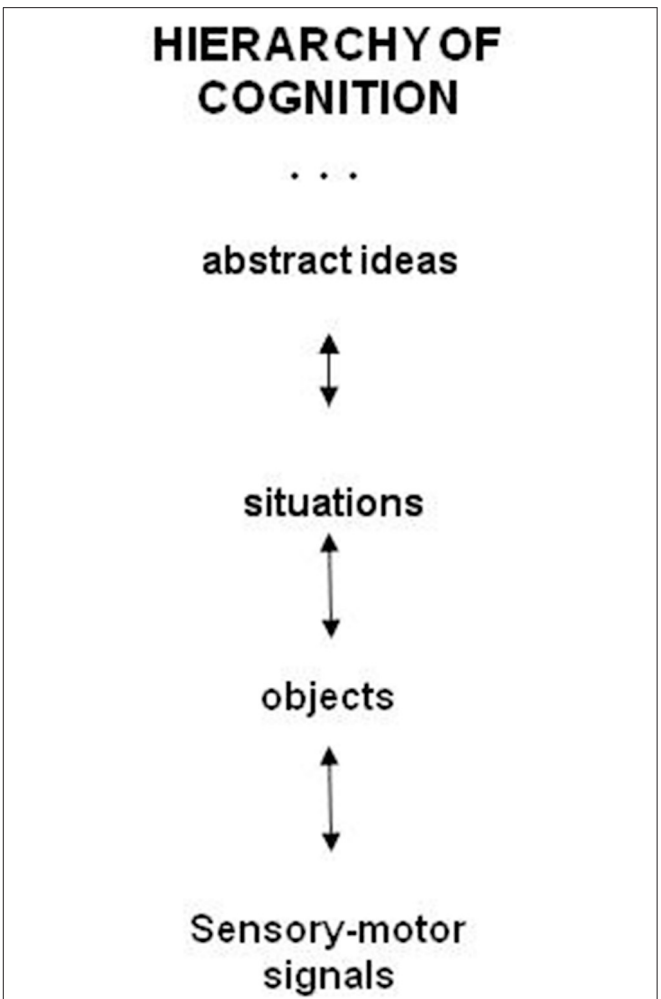


FIGURE 1 | The hierarchy of cognition. Cognition is organized into an approximate hierarchy of concepts from objects to abstract concepts and higher up to the highest cognitive concepts of the meaning and purpose. Aesthetic emotions function as motivators of learning concepts at every level, at the top of the hierarchy these are the emotions of the beautiful.

beautiful to the meaning and purpose of human life, yet his intuitions have been well ahead of understandings of his contemporaries. The only widely known Kantian idea about the beautiful is that it is "aimless purposiveness," often with emphasis on aimless, because the purposiveness of the beautiful is not understood even in contemporar y aesthetics. This is clearly seen when visiting museums of contemporary art.

While progress in understanding of the purpose of human life can be seen in art evolution from cave art to the 19th century, in the 20th century art the exploration of purposiveness has been disappearing. In rare pieces of contemporary art the meaning and purpose of life is explored. This ignoring Kantian intuitions in contemporary art is likely to be closely related to the fact that the idea of "science" become important in cultural life (without understanding of what science is). Existing science does not understand what is beautiful. For example, G. Dickie, an influential philosopher of art, a president of the American Society for Aesthetics, and author of popular textbooks developed an "institutional theory of aesthetics" which defines beautiful as what has been accepted as beautiful by respected art institutions; it is still widely accepted as a state of the art in understanding of the beauty. In wide culture beautiful is understood as more related to sex than to the meaning of life. In university courses on aesthetics beauty is related to shapes, colors, forms, and progressive social uses of art, rather than to the purposiveness of life. So, I would again emphasize that the theoretically predicted properties of the emotions of the beautiful, their relations to the meaning of life are unexpected in contemporary aesthetics and contradictory to accepted views. Nevertheless these theoretical predictions have been experimentally tested and confirmed (Schoeller and Perlovsky, 2015, 2016), which is the fundamental property of the science of physics of the mind.

This section gives an example of complicated cognitive mechanisms explained from the first principles, and making theoretical predictions that have been tested in experiments.

THE DUAL HIERARCHY, LANGUAGE, AND COGNITION

The recognition that language and cognition are not the same, that these abilities are served by different mechanisms of the mind, began a revolution in 20th century linguistics initiated by N. Chomsky (1957). Many psycholinguists and evolutionary linguists today disagree with Chomsky's complete separation of language from cognition (Cangelosi and Parisi, 2002; Christiansen and Kirby, 2003; Steels, 2011), yet many questions remained unanswered. What is the difference between cognition and language? Language is so important for thinking that it is difficult to comprehend what cognition would be without language. How does cognition interact with language? Do we think with words, or only use words as labels when a chunk of a thinking process is complete? There is virtually infinite number of possible associations between words and objects, so how is it possible that every child learns correct associations? Why children learn language by the age of 5 or 7, but do not think like adults until much later? What exactly are the changes in neural mechanisms? Do adults really understand what they say, and what does it mean to really understand? Some people are good at speaking language, while not equally good in discussing with other people, or understanding the real world. Opposite examples could be found. The science needs to understand the mechanisms of language and cognition interactions; why they are so interdependent, and so separate? What neural mechanisms animals need to learn language?

These questions and many more can be explored with adding one fundamental principle to those previously discussed. *Dual model* (Perlovsky, 2004, 2006a, 2007a,b, 2009a, 2013b) is a fundamental principle of the mind modeling interaction between language and cognition. According to the dual model, every mental model has a cognitive and language parts. Their initial states are vague. In a newborn brain most of cognitive and language models are placeholders without specific contents.

Adding dual models to the cognitive hierarchy in **Figure 1** leads to two parallel hierarchies of language and cognition shown in **Figure 2** (Perlovsky, 2006a, 2009a, 2013b). In childhood language representations are learned fast and become crisp and conscious. This is possible because language acquisition relies on language spoken around, in which contents of language models, words, phrases, are "ready-made" for learning. But many cognitive models remain vague until much later; cognitive learning is much more difficult, because cognitive models do not exist in the world "ready for learning." At an early age, everyone can talk about good guys and bad guys, but nobody even at 40

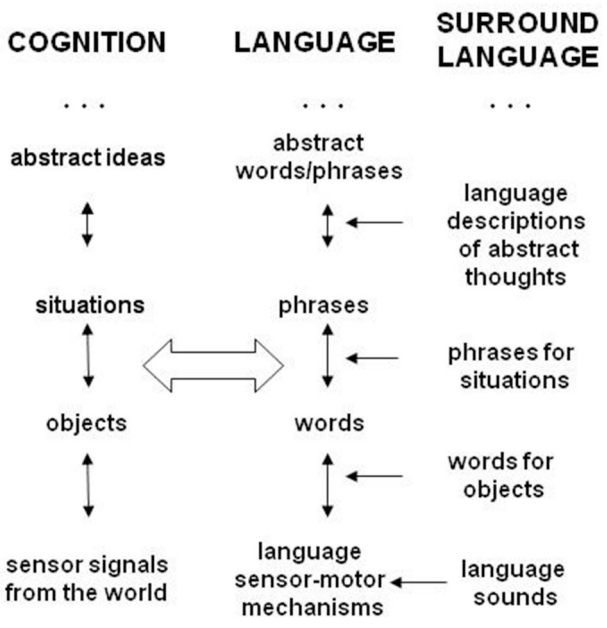


FIGURE 2 | The dual hierarchy model (see Perlovsky, 2006a). Language and cognition are organized into approximate dual hierarchy. Every representation has two parts, cognitive and language. Learning language is grounded in the surrounding language throughout the hierarchy. Cognitive hierarchy is grounded in experience only at the very "bottom"; the cognitive hierarchy is constructed from experience guided by language.

or 80, can use these concepts without errors in real life. Ideas of good and evil have been discussed for millennia.

Throughout the hierarchy, linguistic parts of representations are crisp and conscious in every mind at an early age, but equally crisp and conscious cognitive models my never be understood. Representations of objects are acquired early, alongside with language, because we see objects ready-made for cognitive learning. But contents of abstract concepts do not exist in the world "ready-made." Not every combination of objects or events is worth learning as a separate abstract concept, only few combinations are important for understanding. Therefore learning abstract concepts require experience guided by language. If a concept does not exist in language, it is likely that it does not exist in cognition, and corresponding events are not even noticed. Many people speak words without full cognitive understanding of what these words designate in real life. These aspects of language-cognition have not been duly noticed or explained; the mechanism of the dual model explains them. Language models refer to facts of language, and not directly to events in the world. Cognitive models combine language with experience and refer to events in the world.

The dual model hypothesis is tentatively supported by experimental data (Franklin et al., 2008). These authors describe that certain representations based in the right brain hemisphere in pre-linguistic infants are rewired to the left hemisphere as language is acquired. Brain modules and neural connections involved in the dual models and knowledge instinct were discussed in (Levine and Perlovsky, 2008a,b).

The dual model has been fundamentally important for the emergence of the hierarchy of the mind. Learning should be grounded in experience (Cangelosi and Riga, 2006; Tikhanoff et al., 2006; Coventry et al., 2010). But concept-models of cognition are grounded in experience only at the lower levels of concrete objects; at this level human abilities are not much different from that of pre-human animals' (Spelke and Kinzler, 2007). Understanding situations and abstract concepts cannot be based on experience alone. The referenced publications discuss in detail why this is mathematically impossible: there are simply too many combinations of objects and events (more then all elementary interactions in the life of the Universe). No life's experience would ever be sufficient to learn which combinations are important for noticing them and learning as separate abstract concepts.

The dual hierarchy model offers a resolution of an old problem of sign and symbol (Perlovsky, 2007a). "Symbol is the most misused word in our culture" (Deacon, 1989). "Symbol" is used in simple cases referring to traffic signs, or axiomatic mathematical notations, and in the most profound cases of cultural and religious symbols. The dual model explains that a sign corresponds to a language part of the dual model (even if it is a part of a special sign system, such as notations in chess or mathematics). Symbols can be used profoundly to denote processes of sign interpretation, connecting language and cognitive parts of the model in DL processes from vague to crisp.

The dual model answers questions formulated at the beginning of this section; it explains how cognition interacts with language: a cognitive process proceeds using both cognitive and language representations. This enables thinking to be grounded in the real world to the extent available to the thinker, and still proceed using language whenever understanding of the real world is insufficient. Similarly both representations are used when speaking, depending on one's abilities and preferences; language or cognitive models receive preference in the speaker's mind. People of "speaking type" can shift between cognitive and language models automatically and without notice, while "cognitive types" may concentrate on cognition.

Associations between words and objects or events are learned among virtual infinity of possible association by every child through the process "from vague to crisp" modeled by DL. The reason language is learned first and cognition is learned later because language representations exist "ready for learning" in the surrounding language, while learning cognitive representations requires experience and guidance by language. Adults are different from children in that a larger percentage of their cognitive models are crisp. Still a significant part of what most people are saying is understood only in terms of language, but not necessarily in terms of real world entities. Animals cannot learn language no cognitive hierarchy because they are missing a neural mechanism of the dual language-cognitive model.

To summarize, cognitive models at higher levels are learned based on both, life experience and language models. In this process language guides cognition: language identifies for cognition, which combinations of lower level concepts are meaningful for learning as a higher level concept. Language hierarchy is learned "ready-made" from the surrounding language at an early age. During the rest of an individual's life the knowledge instinct drives the mind to learn the cognitive hierarchy from life experience in correspondence with the language hierarchy. If a certain idea does not exist in a language, this idea does not exist in cognition, and corresponding events would not even be noticed. Cognitive models are grounded in language. Many theoretical predictions of the dual model hierarchy in this section have not yet been experimentally proven, they should be considered as hypotheses and their experimental verifications are topics for future research.

This section, as the previous one, gives an example of a vast field of complicated cognitive mechanisms explained from the first principles, and making theoretical predictions that can be tested in experiments; few of these predictions have been tested experimentally. It answers many questions that could not even have been formulated previously. Detailed mathematical models are discussed in given references. The theory discussed in this section is a part of physics of the mind, it is a step toward making psychology a "hard" science.

EMOTIONALITY OF LANGUAGES AND CULTURES

In non-human animals voice muscles are controlled from ancient involuntary emotional centers. For this reason animal voicing is mostly inseparable from emotions. "Voluntary control of vocalization is limited" (Deacon, 1989; Schulz et al., 2005; Perlovsky, 2009b; Simonyan and Horwitz, 2011). Evolution

of language, semantically loaded voice, required to free voice mechanisms from involuntary emotional control. For this purpose in the course of language evolution human brain evolve recent laryngeal control centers in cortex, which make possible voluntary control of voice muscles.

Involuntary emotionality of voice has been significantly reduced. With evolution of language an ability for strongly emotional voice has mostly evolved into a separate ability for song and music (Perlovsky, 2012c). But unconscious emotionality of voice could not completely disappear. This everyday low emotional prosody performs a highly important cognitive function: it motivates connecting sounds of words with their cognitive meanings (Perlovsky, 2011b, 2012a). Let me emphasize this possibly non-obvious point. Language and its main way of functioning, speech, can only function if sounds of words are perceived emotionally. If a word sounds produce no emotions and no motivations, this word has no meaning. I dwell on this point because it contradicts accepted understanding. "Emotional speech" often is used as a synonym of meaningless or at least devoid of deep meaning, which could be true, especially if emotionality is high and emotions overtake the reason. Here I emphasize the opposite point: no emotionality indicates absence of any interest, and therefore convey no meaning. Proper emotionality is essential (Perlovsky, 2009b, 2011a). Low emotional prosody, which is below the level of consciousness has not been studied. This is an important area for future research.

The following part of this section makes theoretical predictions that follow from the few basic principles formulated above, mathematical models have been presented in given references, and the theoretical predictions are experimentally testable and will be tested in the near future. Emotional prosody of human voice, even if unnoticed, affects the entire psyche and even culture. In pre-human animals conceptual and emotional systems (understanding and evaluation) are less differentiated than in humans. Animal cries engage their psyche as a whole, rather than conceptual and emotional mechanisms separately. For example consider calls of vervet monkeys (Seyfarth and Cheney, 2003). The calls designate types of predators around; still "understanding of a situation (concept of danger), evaluation (emotion of fear), and behavior (cry and jump on a tree) are not differentiated, each call is a part of a single concept-emotion-behavior-vocalization psychic state with very little differentiated voluntary control" (Perlovsky, 2006a).

Humans on the opposite have separate mechanisms of emotions, concepts, and behavior. Differentiation of psychological states with voluntary control over each part must have evolved contemporaneously with evolution of language and rewiring of the brain.

It follows that language, while contributing to developing detailed ability for concepts, also contributed to separating and perfecting functions of concepts, emotions, and behavior. This differentiation destroyed the unity of psyche inherited from the pre-human past. Language evolution also led to losing unity of psyche, started losing wholeness. While in pre-human animals every element of knowledge is tightly connected to emotional evaluation of a situation, and to appropriate behavior satisfying instinctual needs, this is not so for humans. A significant part of cultural knowledge formulated in language is not emotionally connected to human instinctual needs. This is tremendously advantageous for development of conceptual culture, for science, and technology. Humans can deliberately discuss ideas.

But this freedom of deliberate conversation and clear conceptual thinking exerts a price on human psyche. Human psyche is not necessarily unified. Language is not directly linked to instinctual mechanisms. Often knowledge developed in culture does not fit with instinctual requirements that remain our inseparable part. In addition some elements of knowledge often contradict other elements. Human psyche must be unified by the highest models of the meaning and purpose evolved for this purpose at the top of the hierarchy of the mind (Perlovsky, 2007b, 2009b, 2013b). Therefore contradictions in the system of knowledge, a disconnect between knowledge and instincts, the lost synthesis, lead to internal crises and may cause clinical depressions. When psychic states missing synthesis preoccupy majority of population, knowledge loses its value, including knowledge of social organization, cultural calamities occur, wars and destructions (Diamond, 1997; Perlovsky, 2006a, 2007b, 2012d). Evolution of culture requires a balance between differentiation and synthesis. Differentiation is the very essence of cultural evolution. But it may lead to emotional disconnect between conceptual knowledge and instinctual needs, to the lost feeling of the meaning and purpose, including the purpose of any cultural knowledge, and to cultural destruction.

There is much evidence that languages differ in their emotional and conceptual contents (Guttfreund, 1990; Buchanan et al., 2000; Harris et al., 2003; Perlovsky, 2007a,b, 2009b, 2012d). While all contemporary languages lost involuntary connections between sounds and emotions characterizing animal vocalizations, this "hardwired" connections between voice and emotions has been replaced by habitual connections. As long as sounds of a language remain unchanged, the language maintains historical connections between word sounds and associated emotions. But if sounds of a language change fast, this historical connections might be lost.

A significant mechanism affecting a speed of language sound change and therefore emotionality of the language is word morphology, such as inflections expressing grammatical cases, voices, aspects, genders, numbers, tenses, and other constructs. A strongly inflected language may have dozens or even many dozens of inflections expressed by affixes and other grammatical devices. Every child hears these affixes every day, therefore knows how to pronounce them, even if does not know which grammatical category it expresses and when it should be used. In inflectional languages, like most European languages, pronunciation of affixes is to some extent fused with pronunciation of the word roots. Therefore positions of laryngeal muscles for pronouncing word roots should be concordant with pronouncing affixes. Sounds of affixes control to some extent sounds of roots. Affixes are "tale that wag the dog," like anchors keeping the word sounds, and therefore historical emotions. Languages with many affixes tend to keep their sounds changes slow.

For example, Middle English, similar to other Germanic languages, had a number of inflections. About 500 years ago

during transition to Modern English most of inflections have been lost (remaining inflections include "ed" for the past tense and "s" for plurals). English lost anchors for its sounds, and sounds of English started changing fast (Lerer, 2007). English sounds significantly change in each generation. e.g., a well know change is the Great Vowel Shift. Much less research have been devoted to losing historical connections between word semantic meanings and corresponding emotions. This has led to low emotionality of contemporary English.

This low emotionality makes English a powerful tool of semantic thinking. Ancient emotions determined by language sounds and unrelated to semantic contents do not interfere with the thought train. English is very good for science and engineering (Perlovsky, 2013c, 2016a). The other side of low emotionality is that English is losing historical connections of value words to cultural values evolving over millennia. Recent generations change cultural values according to current fads; a lot of people think that this is possible because today people are smarter than in the old days, and therefore are not bound by meaningless traditions. It is not appreciated that this freedom from traditional values (good or bad) is due to the fact that English language sound are changing fast and for this reason English is literally "losing anchors," which is not a guarantee that current fads are better than millennial traditions.

On the "other side" of language emotionality is Arabic language. It is a fusional language, in which inflections are strongly fused with word roots. It follows that sounds of Arabic change slowly (if at all). Semantic meanings of Arabic words are strongly connected to historically ancient emotions. Arabic may not be flexible for scientific thinking. But Arabic moral values are strongly rooted in history. Many Arabic people therefore are sure about their moral values. It is important to appreciate that current contradictions between Arabic and English speaking cultures do not depend on specific political leaders, but are rooted in the very sounds of Arabic and English languages (Perlovsky, 2009b, 2011a, 2012d).

Again, this section gives an example of a vast field of complicated cognitive and language mechanisms as well as their affects on cultures. A vast field of knowledge is explained from the first principles; a theory makes predictions that can be tested in experiments; few of these predictions have been tested experimentally, these are directions for future research. It answers many questions that could not even have been formulated previously. Mathematical models are discussed in given references. The theory discussed in this section is a part of physics of the mind, it is a step toward making psychology a "hard" science.

COGNITIVE FUNCTIONS OF MUSIC

Cognitive functions of music, the reasons for its evolution from pre-human vocalizations to Beethoven, Chopin, and Justin Bieber could not have been understood. Aristotle (1995) asked "why music being just sounds reminds the states of soul?" Kant could not understand the role of music in cognition Kant (1790). Darwin (1871) thought that music is the "greatest mystery." And contemporary musicologists could not find an answer (Editorial, 2008; Honing et al., 2015).

An explanation of music cognitive functions have been derived from the dual model (Perlovsky, 2006a, 2010a, 2012b,c, 2013a, 2014, 2015a,b). Evolution of language led to explosion of knowledge and a number of concepts. Concepts contradict other concepts to some extent. These contradictions among concepts dissatisfy the knowledge instinct and produce unpleasant emotions, cognitive dissonances (Festinger, 1957). Cognitive dissonances are immediately resolved: the new contradictory knowledge is discarded fast and usually without reaching consciousness (Jarcho et al., 2011). It follows that evolution of language, cognition, and culture required a cognitive mechanism for overcoming cognitive dissonances without rejecting knowledge.

This mechanism had to act fast and to be related to language. And this mechanism existed since the beginning of language evolution, it is language prosody. Low-emotional prosody is overcoming minute cognitive dissonances present in everyday choices. Overcoming stronger cognitive dissonances, which appeared with evolution of language and culture, dissonances related to unrequited love, betrayals by friends and loved one, required stronger emotions. These stronger emotions appeared in songs, an ability which eventually evolved into music (Perlovsky, 2006c, 2016a,b,c).

This theory which relates music evolution to cognitive dissonances explains why many people enjoy listening to sad music. Some music is so sad it cannot be listened without tears. Listeners of the BBC's Today program in 2004 voted Barber's Adagio for Strings the "saddest classical" work ever. It is among the highest-selling classical music piece. The physical theory of the mind described here explains a mysterious power of music over us as well as Biblical statement: "in much wisdom is much grief" (Ecc. 1:18). For we leave in the sea of cognitive dissonances, in the sea of grief. Music helps us overcome the grief of knowledge and to continue developing the culture.

The theoretical prediction, resolving the millennial mystery of music by relating it to cognitive dissonances have been confirmed experimentally (Masataka and Perlovsky, 2012a,b, 2013; Cabanac et al., 2013; Perlovsky et al., 2013). This theory opens a vast field for future research, including experimental measurements of musical emotion, e.g., what is the emotional distance between a musical phrase from Beethoven and another musical phrase from Chopin. How many musical emotions exist?

DYNAMIC LOGIC, DL

DL is a mathematical technique modeling the knowledge instinct, or more specifically, the brain-mind mechanism of matching vague top-down signals to bottom-up signals without computational complexity (Perlovsky and McManus, 1991; Perlovsky, 2001, 2006a,b; Perlovsky et al., 2011; Vityaev et al., 2011; Kovalerchuk et al., 2012; Perlovsky and Shevchenko, 2014). It is a mathematical foundation of the physics of the mind and all

results discussed in this paper. It is a fundamental principle of the mind describing the process from vague to crisp representations.

The mathematical description, following (Perlovsky et al., 2011) is given below. An index m numbers top representations; an index n numbers bottom representations; an index i numbers BU signals making up the n-th representation. Parameters x_{ni} measure the strength of association of the BU signal i with bottom representation n, and p_{mi} measure the strength of association of the BU signal i with top representation m. Values of these parameter are limited between 0 and 1. Associations between top and bottom representations are modeled by

$$f(m|n) = r(m)\,\ell(n|m) / \sum_{m' \in M} r(m')\,\ell(n|m'). \quad (1)$$

$$\ell(n|m) = \prod_{i=1} p_{mi}^{xni}(1 - p_{mi})^{(1-xni)} \quad (2)$$

Here $\ell(n|m)$ are pdf-like measures, and f(m|n) are probabilities-like measures, similar to a posteriori Bayes probabilities. Under certain conditions, these variables indeed can be interpreted as probabilistic measures. For preserving these probabilistic interpretations $\ell(n|m)$ is defined so that integration over x yields 1. And parameters r(m) are used to model the proportion of signals m in top-down representations. These representations model a single level in the hierarchical mental structure; at the lowest level of the hierarchy x_{ni} represent sensor signals: if a feature i is present in object or event n, $x_{ni} = 1$, otherwise 0.

Learning in DL processes constitutes adapting parameters p_{mi} and r(m) so that top representations m correspond to patterns in bottom representations x_{ni}. This process maximizes a total similarity measure between all bottom patterns and top representations,

$$L(\{n\},\{m\}) = \prod_{n \in N} \sum_{m \in M} r(m)\ell(n|m). \quad (3)$$

Maximizing this similarity is a model of KI.

The learning process maximizing KI (Perlovsky et al., 2011) can be specified iteratively,

$$p_{mi}^{it+1} = p_{mi}^{it} + dt \sum_{n} f(m|n)[\partial \ln \ell(n|m) / \partial p_{mi}]^{it}, \quad (4)$$

$$f^{it+1}(m|n) = [r(m)\,\ell(n|m) / \sum_{m' \in M} r(m')\,\ell(n|m')]^{it}, \quad (5)$$

$$r^{it+1}(m) = [(1/N) / \sum_{n} f(m|n)]^{it}, \quad (6)$$

In equation (4) a parameter dt is an increment of the internal time t of the DL iterations. A fundamental aspect of the DL learning is an initial vague state, which is achieved by specifying the unknown parameter values p_{mi} near 0.5. This value of p_{mi} corresponds to maximal variances of $\ell(n|m)$ and vague representations f(m|n). This state corresponds to the Aristotelian potentiality. In the process of perception, "mind meets matter,"

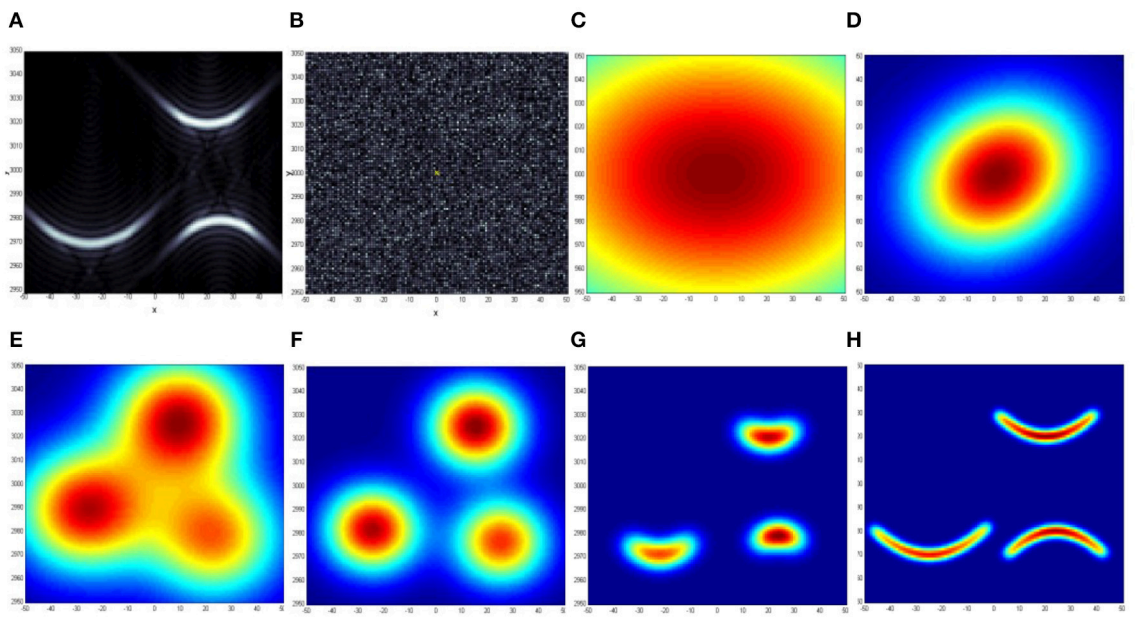


FIGURE 3 | Perception of "smile" and "frown" patterns in noise, an example of dynamic logic "from vague-to-crisp" process: (A) true "smile" and "frown" patterns are shown without noise; **(B)** actual image available for recognition (signals are below noise, signal-to-noise ratio is about 1/3); **(C)** an initial vague concept-model; **(D)** through **(H)** show improved concept-models at various iteration stages (total of 21 iterations). Between stages **(D)** and **(E)** DL tries to fit the data with more than one model and decided, that it needs three models to "understand" the content of the data. Until stage **(G)** the DL "thought" in terms of simple blob models, at **(G)** and beyond, the algorithm decided that it needs more complex parabolic models to describe the data. Iterations stopped at **(H)**, when similarity (3) stopped increasing.

TD and BU signals interact and representations reach crisp states corresponding to the Aristotelian actualities. We show in an example below that this process converges fast.

In this example, **Figure 3**. illustrates the DL perception of "smile" and "frown" patterns in noise. Patterns without noise are shown in A; with noise, as actually measured they are shown in B.

When models come close to the true shape, iteration 17, **Figure 3G**, there is sufficient sensitivity to determine that parabolic shapes better match signals, three parabolic shapes are activated. At iteration 21, **Figure 3H**, iterations stop, because similarity (3) stopped increasing with iterations. The number of computer operations in this example was about 10^9. Thus, a problem that was not solvable due to CC becomes solvable using DL.

To summarize this example, during DL learning initial vague and uncertain models (Aristotelian potentialities) are associated with structures in the input signals (Aristotelian forms interact with matter), and vague models become more definite and crisp with successive iterations. In the image available for recognition, **Figure 3B**, signal is below noise, signal-to-noise ratio is about 0.3. This is a significant improvement over other state-of-the-art practically working algorithms; a standard required signal-to-noise ratio is more than 30. The achieved improvement is about 100 times.

The above formulation of DL includes dual models as well as the dual hierarchy. Some x_{ni} and p_{mi} correspond to cognitive representations and other correspond to language representations. Language representations exist in surrounding language and are learned early in life. Cognitive representations are learned from experience under guidance of existing language representations. Existing preliminary simulations of systems with cognitive and language data indicate that interactions of cognition and language can self-organize by association of both types of representations in a single model. This process could be speeded up if certain associations among vague models are inborn. Understanding inborn associations is a future research direction.

Adequacy of DL for modeling neural mechanisms of perception has been experimentally proven in (Bar et al., 2006; Kveraga et al., 2007).

CONCLUSION

This paper establishes a new area of science, physics of the mind. Physics of the mind, let's repeat is methodologically similar to all areas of physics in identifying few fundamental principles and their mathematical models, a general mathematical model built from these few principles, describing a vast area of knowledge, and making experimentally testable predictions. Experimental tests of these predictions confirm or disconfirm the theory.

Physicists know that the very first test of a scientific theory is its elegance and beauty; these include Einstein (see McAllister, 1999), Poincare (2001), Dirac (1982). The beauty of a scientific theory is its ability to describe a vast area of knowledge from few basic principles, and to make experimentally testable predictions. The actual experimental tests are the final proof of the theory. Currently a number of theoretical prediction have been experimentally confirmed, even so they are unexpected and go against accepted views.

Still a number of predictions remain to be confirmed, a vast area of theoretical and experimental development is opened for future research. Traditional psychology is a "soft" science that does not develop mathematical models of the mind self-organization based on few principles, describing vast areas of knowledge, and making experimentally verifiable predictions. A new area of science physics of the mind extends psychology toward "hard" sciences.

Opportunities for unified areas of research arise in place of former misunderstandings and contradictions.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Aristotle (1995). "The complete works," in *The Revised Oxford Translation*, ed J. Barnes (Princeton, NJ: Princeton Univ. Press. Original work VI BCE).

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends Cogn. Sci.* 12, 193–200. doi: 10.1016/j.tics.2008.02.004

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al., (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U.S.A.* 103, 449–454. doi: 10.1073/pnas.0507062103

Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., et al. (2000). Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cogn. Brain Res.* 9, 227–238. doi: 10.1016/S0926-6410(99)00060-9

Cabanac, A., Perlovsky, L., Bonniot-Cabanac, M.-C., and Cabanac, M. (2013). Music and academic performance. *Behav. Brain Res.* 256, 257–260. doi: 10.1016/j.bbr.2013.08.023

Cangelosi, A., and Parisi, D. (eds). (2002). *Simulating the Evolution of Language*. London: Springer.

Cangelosi, A., and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: experiments with epigenetic robots. *Cogn. Sci.* 30, 673–689. doi: 10.1207/s15516709cog0000_72

Chomsky, N. (1957). *Syntactic Structures*. Haag: Mouton.

Christiansen, M., and Kirby, S. (2003). *Language Evolution*. Oxford, UK: Oxford University Press.

Coventry, K. R., Lynott, L., Cangelosi, A., Monrouxe, L., Joyce, D., and Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. *Brain Lang.* 112, 202–213. doi: 10.1016/j.bandl.2009.06.001

Darwin, C. R. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Deacon, T. (1989). The neural circuitry underlying primate calls and human language. *Hum. Evol. J.* 4, 367–401. doi: 10.1007/BF02436435

Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies*. New York, NY: W.W. Norton and Co.

Dirac, P. A. M. (1982). *The Principles of Quantum Mechanics*. Oxford: Oxford University Press.

Editorial (2008). Bountiful noise. *Nature* 453, 134. doi: 10.1038/453134a

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston, IL: Row Peterson.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P., and Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3221–3225. doi: 10.1073/pnas.0712286105

Grossberg, S. (1988). *Neural Networks and Natural Intelligence*. Cambridge, MA: MIT Press.

Grossberg, S., and Levine, D. S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: blocking, inter-stimulus interval, and secondary reinforcement. *Appl. Opt.* 26, 5015–5030. doi: 10.1364/AO.26.005015

Guttfreund, D. G. (1990). Effects of language usage on the emotional experience of Spanish-English and English-Spanish bilinguals. *J. Consult. Clin. Psychol.* 58, 604–607. doi: 10.1037/0022-006X.58.5.604

Harris, C. L., Ayçiçegi, A., and Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Appl. Psycholinguist.* 24, 561–579. doi: 10.1017/S0142716403000286

Honing, H., ten Cate, C., Peretz, I., and Trehub, S. E. (2015). Without it no music: cognition, biology and evolution of musicality. *Philos. Trans. R. Soc. B.* 370, 1664. doi: 10.1098/rstb.2014.0088

Ilin, R., and Perlovsky, L. I. (2010). Cognitively inspired neural network for recognition of situations. *Int. J. Nat. Comput. Res.* 11, 36–55. doi: 10.4018/jncr.2010010102

Jarcho, J. M., Berkman, E. T., and Lieberman, M. D. (2011). The neural basis of rationalization: cognitive dissonance reduction during decision-making. *Soc. Cogn. Affect. Neurosci.* 6, 460–467. doi: 10.1093/scan/nsq054

Kant, I. (1790). *The Critique of Judgment* (Trans. J. H. Bernard). Amherst, NY: Prometheus Books.

Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.

Kovalerchuk, B., Perlovsky, L., and Wheeler, G. (2012). Modeling of phenomena and dynamic logic of phenomena. *J. Appl. Non-classical Logics*, 22, 51–82.

Kveraga, K., Ghuman, A. S., and Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain Cogn.* 65, 145–168. doi: 10.1016/j.bandc.2007.06.007

Lerer, S. (2007). *Inventing English*. Chichester, NY: Columbia University Press.

Levine, D. S., and Perlovsky, L. I. (2008a). Neuroscientific insights on Biblical myths. Simplifying heuristics versus careful thinking: scientific analysis of millennial spiritual issues. *Zygon* 43, 797–821.

Levine, D. S., and Perlovsky, L. I. (2008b). "A network model of rational versus irrational choices on a probability maximization task," in *World Congress on Computational Intelligence WCCI* (Hong Kong).

Masataka, N., and Perlovsky, L. I. (2012a). *Music Can Reduce Cognitive Dissonance in Nature Precedings*. Available online at: http://precedings.nature.com/documents/7080/version/1

Masataka, N., and Perlovsky, L. I. (2012b). The efficacy of musical emotions provoked by Mozart's music for the reconciliation of cognitive dissonance. *Sci. Rep.* 2:694. doi: 10.1038/srep00694

Masataka, N., and Perlovsky, L. I. (2013). Cognitive interference can be mitigated by consonant music and facilitated by dissonant music. *Sci. Rep.* 3:2028. doi: 10.1038/srep02028

Mayorga, R., and Perlovsky, L. I. (eds). (2008). *Sapient Systems*. London: Springer.

McAllister, J. W. (1999). *Beauty and Revolution in Science*. Ithaka, NY: Cornell Univ Press.

Perlovsky, L. (2016a) Cognitive function of music and meaning-making. *J. Biomusic Eng.* S1:004.

Perlovsky, L. I. (1998). Conundrum of combinatorial complexity. *IEEE Trans. PAMI* 20, 666–670. doi: 10.1109/34.683784

Perlovsky, L. I. (2001). *Neural Networks and Intellect: Using Model-Based Concepts*. New York, NY: Oxford University Press.

Perlovsky, L. I. (2002). Aesthetics and mathematical theories of intellect. *Iskusstvoznanie* 2/02, 558–594. (Russian).

Perlovsky, L. I. (2004). Integrating language and cognition. *IEEE Connections Feature Article* 2, 8–12.

Perlovsky, L. I. (2006a). Toward physics of the mind: concepts, emotions, consciousness, and symbols. *Phys. Life Rev.* 3, 22–55. doi: 10.1016/j.plrev.2005.11.003

Perlovsky, L. I. (2006b). Fuzzy dynamic logic. *New Math. Nat. Comput.* 2, 43–55. doi: 10.1142/S1793005706000300

Perlovsky, L. I. (2006c). *Music–the First Principles, Musical Theater*. Available online at: http://www.ceo.spb.ru/libretto/kon_lan/ogl.shtml (Accessed December 14, 2001).

Perlovsky, L. I. (2007a). "Symbols: integrated cognition and language," in *Chapter in Semiotics and Intelligent Systems Development,* eds R. Gudwin and J. Queiroz. (Hershey, PA: Idea Group), 121–151.

Perlovsky, L. I. (2007b). Evolution of languages, consciousness, and cultures. *IEEE Comput. Intell. Mag.* 2, 25–39. doi: 10.1109/MCI.2007.385364

Perlovsky, L. I. (2007c). The mind vs. logic: Aristotle and Zadeh. *Soc. Math. Uncertain. Crit. Rev.* 1, 30–33.

Perlovsky, L. I. (2008a). Music and Consciousness, Leonardo. *J. Arts Sci. Technol.* 41, 420–421.

Perlovsky, L. I. (2008b). "Sapience, consciousness, and the knowledge instinct. (Prolegomena to a physical theory)," in *Sapient Systems*, eds R. Mayorga and L. I Perlovsky (London: Springer), 33–60.

Perlovsky, L.I. (2009a). Language and cognition. *Neural Netw.* 22, 247–257.

Perlovsky, L. I. (2009b). Language and emotions: emotional Sapir-Whorf Hypothesis. *Neural Netw.* 22, 518–526. doi: 10.1016/j.neunet.2009.06.034

Perlovsky, L. I. (2010a). Musical emotions: Functions, origin, evolution. *Phys. Life Rev.* 7, 2–27. doi: 10.1016/j.plrev.2009.11.001

Perlovsky, L. I. (2010b). Neural mechanisms of the mind, Aristotle, Zadeh, and fMRI. *IEEE Trans. Neural Netw.* 21, 718–733. doi: 10.1109/TNN.2010.2041250

Perlovsky, L. I. (2010c). Intersections of mathematical, cognitive, and aesthetic theories of mind. *Psychol. Aesthetics Creat. Arts* 4, 11–17. doi: 10.1037/a0018147

Perlovsky, L. I. (2010d). The mind is not a kludge. *Skeptic* 15, 51–55.

Perlovsky, L. I. (2011a). Language and cognition interaction neural mechanisms. *Comput. Intell. Neurosci.* 2011:454587. doi: 10.1155/2011/454587

Perlovsky, L. I. (2011b). "High" Cognitive emotions in language prosody, commentary on "emotional voices in context: a neurobiological model of multimodal affective information processing" by C. Brück, B. Kreifelts, and D. Wildgruber. *Phys. Life Rev.* 8, 408–409. doi: 10.1016/j.plrev.2011.10.007

Perlovsky, L. I. (2011c). Abstract concepts in language and cognition, commentary on "Modeling the Cultural Evolution of Language" by Luc Steels. *Phys. Life Rev.* 8, 375–376. doi: 10.1016/j.plrev.2011.10.006

Perlovsky, L. I. (2012a). Emotions of "higher" cognition, comment to Lindquist at al "The brain basis of emotion: a meta-analytic review." *Brain Behav. Sci.* 35, 157–158. doi: 10.1017/S0140525X11001555

Perlovsky, L. I. (2012b). Cognitive function, origin, and evolution of musical emotions. *Music. Sci.* 16, 185–199. doi: 10.1177/1029864912448327

Perlovsky, L. I. (2012c). Cognitive function of music, Part I. *Interdiscip. Sci. Rev.* 37, 129–142. doi: 10.1179/0308018812Z.00000000010

Perlovsky, L. I. (2012d). Emotionality of languages affects evolution of cultures. *Rev. Psychol. Front.* 1, 1–13.

Perlovsky, L. I. (2013a). A challenge to human evolution – cognitive dissonance. *Front. Psychol.* 4:179. doi: 10.3389/fpsyg.2013.00179

Perlovsky, L. I. (2013b). Mirror neurons, language, and embodied cognition. *Neural Netw.* 41, 15–22. doi: 10.1016/j.neunet.2013.01.003

Perlovsky, L. I. (2013c). Language and cognition – joint acquisition, dual hierarchy, and emotional prosody. *Front. Behav. Neurosci.* 7:123. doi: 10.3389/fnbeh.2013.00123

Perlovsky, L. I. (2013d). Learning in brain and machine - complexity, Gödel, Aristotle. *Front. Neurorobot.* 7:23. doi: 10.3389/fnbot.2013.00023

Perlovsky, L. I. (2014). The cognitive function of music, part II. *Interdiscipl. Sci. Rev.* 39, 162–186. doi: 10.1179/0308018813Z.00000000041

Perlovsky, L. I. (2015a). How music helps resolve our deepest inner conflicts. *Conversation*. Available online at: https://theconversation.com/how-music-helps-resolve-our-deepest-inner-conflicts-38531

Perlovsky, L. I. (2015b). Origin of music and the embodied cognition. *Front. Psychol.* 6:538. doi: 10.3389/fpsyg.2015.00538

Perlovsky, L. I. (2016b). Beauty, music, the meaning of life, and vague representations. *Front.*

Perlovsky, L. I. (2016c). *Music: Passions and Cognitive Functions*. San Diego, CA: Elsevier.

Perlovsky, L. I., Bonniot-Cabanac, M.-C., and Cabanac, M. (2010). Curiosity and pleasure. *Psychology* 1:WMC001275. doi: 10.9754/journal.wmc.2010.001275

Perlovsky, L. I., Cabanac, A., Bonniot-Cabanac, M.-C., and Cabanac, M. (2013). Mozart effect, cognitive dissonance, and the pleasure of music. *Behav. Brain Res.* 244, 9–14. doi: 10.1016/j.bbr.2013.01.036

Perlovsky, L. I., Deming, R. W., and Ilin, R. (2011). *Emotional Cognitive Neural Algorithms with Engineering Applications; Dynamic Logic: From Vague to Crisp*. Heidelberg: Springer.

Perlovsky, L. I., and Ilin, R. (2010a). Neurally and mathematically motivated architecture for language and thought. special issue "brain and language architectures: where we are now?" *Open Neuroimaging J.* 4, 70–80. doi: 10.2174/1874440001004020070

Perlovsky, L. I., and Ilin, R. (2010b). Grounded symbols in the brain, computational foundations for perceptual symbol system. *WebmedCentral Psychol.* 1:WMC001357.

Perlovsky, L. I., and McManus, M. M. (1991). Maximum likelihood neural networks for sensor fusion and adaptive classification. *Neural Netw.* 4, 89–102. doi: 10.1016/0893-6080(91)90035-4

Perlovsky, L. I., Plum, C. P., Franchi, P. R., Tichovolsky, E. J., Choi, D. S., and Weijers, B. (1997). Einsteinian neural network for spectrum estimation. *Neural Netw.* 10, 1541–1546. doi: 10.1016/S0893-6080(97)00081-6

Perlovsky, L. I., and Shevchenko, O. (2014). "Dynamic logic machine learning for cybersecurity," in *IEEE and INNS, International Joint Conference on Neural Networks* (Beijing: IJCNN14).

Poincare, H. (2001). *The Value of Science: Essential Writings of Henri Poincare*. Modern Library.

Russell, S., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach, 3rd Edn*. Pearson.

Schoeller, F., and Perlovsky, L. I. (2015). Great expectations - narratives and the elicitation of chills. *Psychology* 6, 2098–2102. doi: 10.4236/psych.2015.616205

Schoeller, F., and Perlovsky, L. I. (2016). Aesthetic chills: knowledge-acquisition, meaning-making and aesthetic emotions. *Front. Psychol.* 7:1093. doi: 10.3389/fpsyg.2016.01093

Schulz, G. M., Varga, M., Jeffires, K., Ludlow, C. L., and Braun, A. R. (2005). Functional neuroanatomy of human vocalization: an H215O PET study. *Cereb. Cortex* 15, 1835–1847. doi: 10.1093/cercor/bhi061

Seyfarth, R. M., and Cheney, D. L. (2003). Meaning and emotion in animal vocalizations. *Ann. N. Y. Acad. Sci.* 1000, 32–55. doi: 10.1196/annals.1280.004

Simonyan, K., and Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *Neuroscientist* 17, 197–208. doi: 10.1177/1073858410386727

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Dev. Sci.* 10, 89–96. doi: 10.1111/j.1467-7687.2007.00569.x

Steels, L. (2011). Modeling the cultural evolution of language. *Phys. Life Rev.* 8, 339–356. doi: 10.1016/j.plrev.2011.10.014

Tikhanoff, V., Fontanari, J. F., Cangelosi, A., and Perlovsky, L. I. (2006). "Language and cognition integration through modelling field theory: category formation for symbol grounding," in *Book Series in Computer Science*, Vol. 4131, eds S. D. Kollias, A. Stafylopatis, W. Duch, and E. Oja (Heidelberg: Springer), 376–385.

Vityaev, E. E., Perlovsky, L. I., Kovalerchuk, B. Y., and Speransky, S. O. (2011). Probabilistic dynamic logic of the mind and cognition. *Neuroinformatics* 5, 1–20.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 02 March 2017
doi: 10.3389/fnsys.2017.00008

Understanding and Self-Organization

Natika W. Newton *

Department of Philosophy, Nassau Community College, State University of New York (SUNY), Garden City, NY, USA

How do we manage to understand a completely novel state of affairs, such as the sudden effects of an unexpected earthquake, or the arrival of a total stranger instead of the sister we were waiting for? In each case, for a moment we might be stunned, but we are able quite quickly to fit these events into our overall framework for understanding the world. However, terrified and despairing we feel, we know what earthquakes are and this event fits that schema; in the case of the stranger we know that this kind of thing happens, and that we must ask the stranger "Who are you, and where is my sister?" This paper asks about the mechanisms by which we rapidly achieve an understanding of our world, both the unexpected changes we may experience, and the ongoing comfortable familiarity we normally have with our surroundings. We attempt a solution by means of examining fundamental questions:

- What is it to understand something?
- What sorts of things do we try to understand?
- Is there a conscious EXPERIENCE of understanding?
- Does understanding involve conscious mental images?
- What is self-organization?

I will argue that these questions revolve around the need of a living organism to take action, and that understanding anything involves knowing how we might act relative to that thing in our environment. The experience of understanding is a feeling that the action affordances of a situation are clear and available. Action (as opposed to reaction) includes imagery, particularly motor imagery, which can be used in the guidance of action. Understanding requires a conscious process involving motor imagery of action affordances, and action can be understood only in self-organizational terms. I explain how self-organization can ground the kinds of action affordance experience needed for conscious understanding. The paper concludes that our day-to-day understanding of our environment is the result of a self-organizing process.

Keywords: understanding, self-organization, consciousness, representation, recursion, emergence

OPEN ACCESS

Edited by:
Yan Mark Yufik,
Virtual Structures Research, Inc., USA

Reviewed by:
Ralph D. Ellis,
Clark Atlanta University, USA
Alianna JeanAnn Maren,
Northwestern University, USA

***Correspondence:**
Natika W. Newton
natika.newton@gmail.com

Received: *18 October 2016*
Accepted: *13 February 2017*
Published: *02 March 2017*

Citation:
Newton NW (2017) Understanding and Self-Organization. Front. Syst. Neurosci. 11:8. doi: 10.3389/fnsys.2017.00008

WHAT IS IT TO UNDERSTAND SOMETHING?

Answering this question requires distinguishing two ways it can be taken. One way is as a "success" term, with the assumption that there can be correct and incorrect understanding. Thus, we can say of someone that she thinks she understands the term "water" if she views water as any clear liquid that quenches thirst, but that she does not really understand it, being ignorant of the molecular composition of water. The other way of taking the concept of understanding is purely

as a mental state of a subject, in which she finds for herself a way of interpreting something (word, sentence, object, or event) that allows her to use her interpretation to think about or act upon her interpretation in a way that satisfies her. We can call this "having AN understanding" of something. Having an understanding in this sense does not involve success or failure, as long as the subject herself is satisfied.

In this paper we are concerned with the second sense of understanding. The former sense is the subject of philosophy of language, while the latter sense concerns only what is going on in the subject's head—i.e., her nervous system. Reaching a state of having an understanding is independent of "correct" definitions, or of knowing the correct meaning in a given language, and is concerned solely with the subject's experience. While the interpretation does not necessarily map onto the objective world, it allows the subject to use her interpretation to incorporate it into existing schemas and to create new models consistent with it. If what she understands in this way is incompatible with the objective world, she will sooner or later discover this, and will have to revise her understanding or abandon it for one more compatible in objective terms. She still has an understanding of the object or situation, which is again not necessarily accurate, but which allows her to act on this revised interpretation.

For example, suppose Susan is rude to her friend Tom, who consequently is hurt and interprets Susan's behavior as evidence that she no longer likes him. His understanding of her behavior satisfies him intellectually, although of course he is disturbed, and his interpretation allows him to decide to snub her at their next encounter. But when he does so she bursts into tears, while his friends tell him that she has been under a great strain and is seriously depressed. Then Tom will no longer interpret her rudeness as a sigh of rejection of him personally. Once again, his new interpretation may be objectively inaccurate, but it provides *an* understanding that allows him to interact with Susan in a way that is intellectually comfortable.

Finally, we can be even more precise by looking at a clear *lack* of understanding, in the well-known example of the Chinese Room (Searle, 1984, discussed at greater length below). Searle imagines himself going through all the motions of a computer with a translation function, which receives questions in Chinese and delivers answers, still in Chinese. Someone on the outside might well-believe that the computer understands Chinese. But Searle, as the computer inside the room, knows that he does not understand, and we can certainly take his word for it.

WHAT SORTS OF THINGS DO WE UNDERSTAND?

Most commonly, we think of understanding *language*. But in fact we must have an understanding of every aspect of our lives—every object we encounter, every event we are involved in or witness, everything that is part of our environment. Without this understanding, we are at a loss as to what to do next. It may sound strange to speak of understanding an *object*. But in the case of any object we need to know what it is for, how we may use it or avoid it, what actions it *affords.* Even a meaningless rock lying in a field can be picked up, thrown, taken home as an ornament, etc. We might be mistaken in our particular understanding of an object if we think a rock is a mushroom and try to bite into it; in that sort of situation we search for a different, we may say more successful understanding. Leaving some aspect of our environment not understood is a worry; we have to figure it out so as to know what to do with it or expect from it. We seek an understanding of events, such as two people whispering at the faculty meeting. What can they be whispering about? We are then able to interpret their whispering as an attempt to locate the memo referred to by the speaker, and we then can go on to the next issue in our general attempt to understanding the meeting as a whole. The unspoken premise here is that being alive is for us a process in which we are always acting in relation to our environment, however minimally (Ellis and Newton, 2010). Lack of understanding obstructs the process of acting, and hence is felt as a problem.

IS THERE A CONSCIOUS EXPERIENCE OF UNDERSTANDING?

There must be conscious experience, at least in the initial encounters with a thing, because having an understanding puts us at ease, enabling us to feel that we can make use of what is understood, or incorporate it into our global experience. In other words, the above is *what it is like* to have an understanding. Lacking any understanding leads to puzzlement and a feeling of insecurity or discomfort: what are we supposed to do in this situation? If we are in an apathetic state because of depression or if we are sufficiently distracted by something else, we may not try to understand (This experience can also be common when we do not care whether we understand or not, as during a murky movie when we have given up and stay only because of politeness). But if we are called upon to DO something about the thing, or perform some actions in light of it, we must attempt to arrive at an understanding, to incorporate the thing into the rest of our current situation.

The discomfort of lacking an understanding—"What's going on?"—leads to attempts to arrive at an understanding, and when we are successful there is the well-known "Ah-ha!" experience. This is normally a positive experience—the discomfort is eased—unless understanding reveals the object to be scary, dangerous or sad. Even then, we are better off for knowing how to react. As we grow accustomed to our familiar environment we take the understanding for granted; if someone brings me coffee while I am working I need not go through a moment of puzzlement, unless this offer of coffee is not at all typical of the normal behavior of this person. Then I might ask some questions. The upshot is that normally, when we have an understanding, we are comfortable in our surroundings, without much thought, and we are unpleasantly aware when we lack it. So we may speak of the "A-hah!" experience, and the "What's going on?" experience, as *what it is like* to have or lack an understanding.

It should be clearly noted that we have access to our state of understanding—we may say "privileged access"—we know when we understand something and when we do not. It can be argued

that we have no special access to our mental states, which are determined in part by the environment. But this objection applies to the first sense of understanding discussed above, which we are not concerned with in this paper. Here, having *an* understanding is a state of which the subject is fully aware. "Accuracy" of understanding does not apply. In some situations I might pretend to understand what's going on, when I really do not. No one else may notice, but I know the difference. This situation can lead to social complexities and confusion, but I myself am usually aware of my role in the awkwardness.

John Searle's "The Chinese Room" is rich in examples. Searle argues that the computational theory of mind—the theory that thinking is manipulating symbols, meaningless themselves, that we have learned correspond to objects and events in the environment—is not accurate. His main argument, that syntax does not yield semantics, uses the well-known example of a person inside a closed room, manipulating Chinese symbols in response to input, matching input with output by following rules in a book. According to computationalism, Searle says, correct manipulation—syntax—should be equivalent to knowing Chinese. But in that case, Searle said, the person handling the symbols should understand Chinese. Searle, imagining himself as the person in the room KNOWS that he does not understand Chinese. So the computational theory is false.

Not only does he know that he doesn't understand Chinese, he knows how to read and obey the instructions in the English manual. His knowledge is made clear, to us and to him, by the fact that he obeys the instructions with no difficulty. To provide a personal anecdote: I was once playing a word game with another person, who for a while seemed to be keeping up with the rules. But she claimed not to understand what she was doing, but was just lucky; when I explained the rules to her, she repeated that she had not understood how the game was played, but that now she did. She clearly described states of not understanding, and of understanding; it seems to follow that those states yielded conscious, reportable experiences (This example is discussed in Newton, 1996).

DOES UNDERSTANDING INVOLVE CONSCIOUS MENTAL IMAGERY?

Mental images have traditionally been thought of as primarily visual—a picture of something is often called an image while a 3-D representation of a structure is a model; an annoying, persistent memory of a tune is "a tune in one's head." But tunes in one's head are no different from pictures in one's head in their central function: they are imagined reproductions of past audio-visual stimuli that we have experienced and now remember. We can recall flavors—taste images; pains—pain images; extreme temperatures—heat or cold images. There seem to be no sensory experiences that cannot be reproduced as mental images: they are like the actual experiences, but are no longer objectively present (Pearson and Kosslyn, 2015).

We can also have proprioceptive images, images of events in our bodies, such as hunger pangs, or motor images, which reproduce the sensations of moving parts of our bodies. Motor images have received much attention in recent years for their role in generating overt bodily movements (Sacks, 1984). According to Jeannerod (1988), movements of our limbs begin with imagery of the movement in the motor cortex; this imagery is activated by allowed to proceed to the sending of nerve impulses to the muscles, which execute the movement. The action can also be prevented prior execution by inhibitory signals in the cortex. Proponents of free will argue that while arm movements can be predicted prior to execution by activation of the motor image, there is still time for the action to be inhibited at the last minute; during this short time the subject can choose inhibition or not (Libet, 1985, p. 143). Whether or not this account is successful will not be discussed in this short paper.

We have seen that having an understanding of something means knowing how one might use or interact with the object or event to be understood. If, as we have argued, there is a feeling of understanding that is conscious, then this feeling must consist in the experience of representations or imagery of some of the possible interactions. For example, suppose you enter an unfamiliar gym and see a novel type of equipment. You can ask how one uses it, or you can simply look at it and try to figure out where the feet go, where the arms go, what types of motion the equipment allows. You are trying to understand the equipment, and the attempt entails sensorimotor imagery of interacting with the machine, imaging what the motions will feel like to execute, etc. You can do all this while passively observing the machine.

Is there any other way of coming to understand the machine? Suppose you ask the attendant to explain it; can't you understand and apply his verbal information ("you step on the foot pedals and hold on to those bars") without producing motor images in your head? No; the images let you know if you can do what he says. Don't we do that kind of thing all the time? That question takes us back to the discussion of conscious understanding. It was argued that having an understanding of something means knowing how one might interact with it, use it, participate with it in some way. Often this kind of knowing is non-verbal. For example, you know how to keep your balance on a bicycle while riding it, and you can imagine doing so. But if you are asked to verbalize this knowledge you may be at a loss. When riding, you tighten and flex muscles, and shift your body weight around, in ways that feel automatic, that you can image clearly through proprioceptive and motor imagery. Confronting a bicycle with understanding, we might say, means generating an image of how you would ride it (And if you cannot generate such an image you will feel, or protest, that you don't know how to ride it). Thus of the two apparently possible ways of having an understanding of riding a bicycle, only the way involving imagery will be satisfying to you. Hearing the explanation without knowing what it would be like for you to ride it does not help you understand, if you do not understand the elements of the explanation. And knowing what it would be like for you to ride is being able to generate imagery of your body in the act of riding. Thus, having an understanding of something involves conscious mental imagery. The only exception will be in cases where you have performed the act so often that you feel sure, without mental rehearsal, that you are familiar with the object or event. But even in those cases, your feeling of confidence can be an "image," in an attenuated sense of

image, of traces of the "Aha" feeling that you achieved when you originally developed an understanding.

Try an experiment: suppose you are asked if you are able to reach the vase on top of the bookcase. How do you decide? If you aren't sure, try introspection. Do you not imagine standing in front of the bookcase and reaching up? Perhaps you have a motor image of standing on your toes and straining to touch the vase. If that image leaves you undecided, you can walk over and try to reach the vase, and get the right answer.

It can be objected that if you know quantitatively the height of the bookcase, and your own height with your arm length added, you could find the answer with no imagery necessary. And the objection is correct, in that you can now find the answer by simple addition. But that fact leads to another case of conscious or semi-conscious motor imagery: the dependence of arithmetic on representation of basic action patterns. Let us look in detail at the most abstract way we use action imagery. I have been defending the claim that understanding, in the sense that once has AN understanding of something, is a process that maps novel stimuli or experiences onto an original structure that is already understood. The novel material, thus mapped, is understood as well as the original material, in that it has become part of the subject's repertoire of usable structures, or mental models.

For example, suppose we call an original structure "reach, grasp, pull." This structure emerged, let us say, when the infant first saw a desirable toy and grabbed it. The structure of that act is the act itself; it is understood because it is created by the infant in response to her own desire. It is the means of satisfying that desire. Now that she has that movement pattern in her repertoire she can use it at will to obtain other desirable things. She is also now ready to use that pattern in other circumstances, to interpret concepts or environmental input that goes beyond immediate satisfaction of a desire for an object. Suppose she is now older and is told she will be taken to a store to buy new shoes. She can understand that prospect easily once she can see it as another instance of "reach, grasp, pull;" going to the store is reaching, selecting, holding, and buying the shoes is grasping, and wearing the shoes home is pulling.

What about the novel qualitative and quantitative properties that are structured by the familiar pattern? How are they understood? It seems correct to say that they are experienced as properties of the fundamental pattern, and are not viewed as distinct from that pattern, but as features of experience that merge with the pattern. If they are purely qualitative, they can be experienced but not described except in terms of a pattern. In describing quantitative properties of the shoes, other recursive patterns will be applied to the details of buying the shoes, the trip, the transaction, etc. E.g., the car is a thing that you get into, and that moves you from point A–B. The red color of the shoes, on the other hand, cannot be described but only experienced and named. Sensory qualities like the smell of the leather, the smoothness, the heft of the shoe are unanalyzable, but as properties of the object are subject to the object's handling and do not need sensorimotor patterns of their own. In general, qualia are presented to us as bound to objects and events we understand through motor imagery, and do not present philosophical difficulties in ordinary circumstances.

We can abandon sensory qualities in purely abstract contexts like mathematics. Now imagine that a child is learning how to add three digit numbers, and must understand what it is to "carry" a numeral. The "reach, grasp, pull" pattern can be applied in an attenuated form here: when we add 123–789, for example, we must "carry" the one from the rightmost column to the middle one, where it is incorporated into the sum of 8 and 2. The term "carry" is clearly metaphorical and derived from physical operations: "add" can be understood in terms of placing one object into a collection of others, "divide" in terms of separating *n* sets of objects from a larger collection, etc. Note that the "reach, grasp, pull" structure is only one of many metaphors based on bodily movements in space: *into* and *out of* are derived from experience with containers in which we can put things or find ourselves inside of; and they can structure metaphors in vastly different contexts—e.g., "the voters put him *in* office"; or "three *goes into* nine three times."

The basic point here is very simple: we understand our bodies in being able to move them and use them to satisfy our wants and needs. Understanding anything has its roots in this ability. Our bodies are our tools for achieving our goals, and there is nothing more to having an understanding of them than intentionally moving them. Because our essence as agents is expressed in conscious, voluntary bodily activity, then having an understanding of our own voluntary actions is itself nothing more than being intentionally able to engage in bodily activity or imaging it (knowing what it is like). We understand ourselves, moreover, *as* voluntary agents. Being self-aware is being aware of our bodies, not as something we, as disembodied subjects, *have,* but as what we are.

Understanding anything, in the sense we are using it, involves situating it into contexts or structures with which we are already familiar. As a person grows and acquires new experiences, the original sensorimotor structures of early life stretch to accommodate these experiences in ways that he can "make sense" of, in light of the earlier experiential structures. It is hard to imagine how one could develop any understanding of novelties except by connecting them with prior experience. An example from academia is the teaching of Plato's Theory of the Forms. The instructor is helpless to explain what a Form is unless she can find something in students' experience to relate it to. She can try beginning, for example, by explaining that Plato's realm of Forms is to the world of concrete objects as, in Christianity, Heaven is to Earth. The success of this move depends upon the students' understanding of Christianity, and that understanding depends, in turn, in part on spatial metaphors such as "above," for Heaven, and "below" for Earth. In other words, trying to explain a complex new concept in terms pertaining only to that concept, with no terms from the hearer's experience to ground the explanation, is useless. The hearer can learn which new words to use with which other new words, but, like Searle in the Chinese Room, will have no grounded understanding of the concept that she can use to think about it in a satisfying and possibly creative way.

Yufik and Friston propose a theory of understanding (this issue) that is highly compatible with that of this paper, except that their theory focuses on neuronal events underlying

understanding, while this one is more concerned with mental acts on a conscious level. One might say that their theory is more "bottom up," mine more "top down." Importantly, both theories exemplify enactivism—the view that cognition is a mode of human activity, and both theories emphasize the role of action representations, or mental models:

> Notice the two key themes of this formulation are an emphasis on active inference or volitional sampling of the world—of the sort that characterizes enactivist or situation approaches to cognition. Second, the progressive elaboration of internalized ("as if") stimulus-response links induces conditional dependencies between the sensory input and internal models of how those predictions were caused—through active sampling (Yufik and Friston, 2016).

In summary: Understanding is tightly coupled with the need of a living organism to take action. Understanding involves knowing how we might perform goal-directed actions relative to the environment. The experience of understanding is a feeling that the action affordances of a situation are not entirely unclear. Action (as opposed to reaction) requires imagery, including motor imagery, that can be used in the guidance of action.

SELF-ORGANIZATION

With this sketch of a theory of understanding, we will attempt to see it as a self-organizing process. To do that requires that we first consider it as a recursive process. Not all cases of recursion are biological, like our theory of understanding. Just as many writers view intentionality as a property of natural language, logic, mathematics, and language provide instances of recursion: the application of a function to its own values to generate an infinite sequence of values. "***Recursion*** occurs when a thing is defined in terms of itself or of its type" (Wikipedia, *Recursion*).

In what follows we examine the concept of self-organization particularly as it takes place in an organism—self-organization as a biological property. The much broader research project originated in applications to dynamical systems theory (Ashby, 1947), followed by physics, chemistry, computer science, and more recently to human behavior. A major influence was the work of I. Prigogine on self-organization in irreversible thermodynamic systems, for which he won the Nobel Prize for Chemistry in 1977 (Prigogine and Nicolis, 1977). Among other central thinkers is Hermann Haken who finds highly important examples of self-organization in brain function (Haken, 2008; Karsenti, 2008).

Below we look at two properties central to self-organization in any type of system, including the brain and human cognitive functions in general.

RECURSION

Language is *recursive* when a type of clause in a sentence is used to make a new sentence. For example "Bobby went to the store" can become a new sentence "Bobby went to the store and to the pharmacy," and "Bobby went to the store and to the pharmacy, and to the movies." The prepositional phrase "to the store" is a grammatical function within the sentence, and that function can be infinitely repeated to make new sentences. Mathematics is recursive when a number series n is extended by "n+1" and "n+1 +1."

Recursion is part of a complete definition of language and mathematics, and the concept has been used to deny language ability to intelligent non-human animals. Recursion also occurs naturally in non-living entities such as crystals (e.g., snowflakes and blocks of quartz), and it frequently results in emergent properties, such as the symmetry of crystals, the shapes of flocks of flying birds, or traffic patterns, which maintain their overall shapes as their sizes change.

Metaphors are essential tools in understanding, both linguistic and otherwise. We understand something by being able to see it in terms of something we already understand. That fact may help to explain our inability to articulate a definition of sensory properties like color; red itself has no articulable components that can be compared to or mapped onto anything else (Lakoff and Johnson, 1987).

The examples of recursion we have been examining include both inanimate, fixed structures like language and mathematics, events like presidential elections, and animate biological processes. We have argued that understanding is a biological process, based on recursive iteration of action structures, or action images. The importance of recursion in our theory of understanding is that, through recursive processes in which structures are extended to new data, the new material is understood simply by being incorporated into a wider context already fully understood.

We are now ready to understand how forming an understanding of elements in our environment involves *emergent properties.*

EMERGENT PROPERTIES

Emergent properties are properties of the patterns resulting from self-organizing processes that are not present in, or predictable from, the individual components that have been organized. A well-known example is traffic patterns: "In the case of a traffic jam, what appears is an entity whose properties need not have anything in common with the properties of its constituent units (cars). In particular, one may have a stationery or even moving back traffic jam while all cars are moving forward. This higher level structure, whose equations of motion are not easily derivable from those of cars, emerges from the interactions between the cars" (Bonabeau et al., 1995). Because the properties are not properties of the individual "agents" but only of the organized whole, these properties are "emergent;" meaning that they did not exist previously in nature, and could not have been predicted from the properties of the "agents." They are more than, and different from, a mere aggregate of the agents that make it up (Anderson, 2011).

Another important feature of emergent properties is that they have causal powers not found in the individual entities making them up. Traffic patterns, when they cause tie-ups in the traffic

flow, cause not only traffic jams but also extreme irritation on the part of the drivers involved. But the traffic jams and the irritation are not caused by individual cars, because even if a single car is driving too slowly for the comfort of other drivers, they are not forced to stay behind it, but can move around it freely, as they could not in the case of a traffic jam (Kerner, 1998). Other properties of collections of individuals, such as electorates that exist at a higher level of organization than the parts, are not in themselves emergent properties, since electoral powers are as true of a mere aggregate of the individuals as of an overall electorate, whose causal properties are reducible to the aggregate of causal powers of individual voters (Ellis, 2012). One might say that the "emergent" causal powers exist because the individuals are arranged in a particular pattern, and that property is true of the individuals as aggregated. It is true that a particular arrangement of individuals has led to the "emergent" properties. These properties, however, are previously unseen in nature, and were not predictable from knowledge of possible aggregates of individuals. They appear, moreover, spontaneously out of chaotic states of individuals, and are not composed by external intelligent agents. Thus, these patterns with their novel properties can be said to emerge from chaotic states because of causal powers of their own.

The concept of emergent properties is now so well-established that it has tempted some to apply the term in cases where the "agents" are not well-understood: for example, consciousness has been called an emergent property in humans and some non-human animals. This would be an appealing way to solve the hard problem of consciousness, if the "agents" that would self-organize to produce it are known. They are presumably states of the brain, but unlike individual cars in the traffic pattern case they have never been observed to create the emergent property of consciousness in which they could be found. Emergent properties are properties of groups of entities that can be observed as part of the final form, not losing their material nature. In the case of consciousness, it seems that no individual entities, or agents, are in any way observable or detectable in conscious experience, which manifests a unity for the conscious agent.

Understanding, however, as we have been using the term, *is* an emergent property of biological processes. The entities that lead to forming an understanding of something are detectable, and can be analyzed out of the experience of understanding. An experience of understanding contains, at least, motor imagery of familiar action patterns, a mental state of puzzlement or tension followed by a relaxing of the mind into the structure and affordances of what is understood, and confidence in the planning of future actions related to the entities or situations now understood. The state of understanding as a whole, moreover, has causal powers (in a given situation) that its components, representations of action patterns, do not have. A satisfying feeling of understanding, such as the "Aha" experience is accompanied by images or representations of action patterns, which the subject uses metaphorically to interpret the novel state of affairs—object or event. These representations may not be the most prominent aspects of the experience, in which attention would be focused on the newly-understood state of affairs. But they are introspectively available. In themselves, these action representations do not constitute understanding of the novel situation, but together with representations of the current stimuli do combine to form a whole scenario that is understood. Without the emergence of understanding, as it is presented here, normal human life would be impossible; one would literally never know what to do. The components of understanding must unite for any functioning in the world to occur.

One source of evidence for the role of action representations is the work of McNeill (1992), who studies the role of gesture in expressing such representations:

> For example, consider a speaker who says, "I was holding a big box" and produces a gesture that mimes holding a big box. In this case, both modalities express the same idea, so the degree of redundancy between gesture and speech is high. The gesture also expresses additional nuances of meaning, such as information about the position of the hands as they hold the box, but the semantic information expressed in the two modalities is largely overlapping. At the other end of the continuum are cases in which there is little or no semantic overlap between the two modalities. In one often-cited example, a speaker describing a scene from a Sylvester and Tweety cartoon said, "she chases him out again" while swinging her arm as if wielding a weapon. In fact, the speaker was describing a scene in which Granny chases Sylvester while swinging an umbrella. In this example, the speaker expresses an aspect of the scene in gesture (swinging the umbrella) that she does not express at all in speech. Thus, in this case, the degree of redundancy between gesture and speech is low (Alibalia et al., 2009).

On this account the speakers are expressing in gestures the motor imagery they are using to describe a scene. The use of gestures indicate that the speaker is thinking of action patterns that she uses to understand and describe the scene. In the second example, the speaker is thinking of a component of the scene that she does not express in words, indicating that she understands the scene herself in terms of her prior experience of swinging an object in her hand.

To summarize this section: understanding is a state of mind that emerges from the blending of other mental states, driven ultimately by the emotion of wanting to be comfortable in one's environment. The mental states of experiencing motor imagery of action patterns and sensory input from the environment, and relating these metaphorically to basic action patterns learned by experience in infancy, create an emergent state of confident action planning which would be impossible if these components were not united by the emotional drive to be "at home" in the world.

SELF-ORGANIZED UNDERSTANDING

What elements are self-organized in the case of understanding? The basic constituents are the simple bodily movements themselves. After mastering the reach-grasp-pull pattern, an infant soon finds that the pattern can be extended in various ways, such as grasping two small things at once between reach and pull. We can say that the larger pattern (reach-grasp-grasp-pull) is self-organized if a) the change was not conceived by the infant in advance, but was a spontaneous extra-grasp addition to the original pattern—in other words, the repeated grasp creates

a higher order pattern reach-GRASP-pull, with the intervening GRASP now encompassing two smaller iterations.; and b) the process is recursive, in that components of the original pattern are used to construct an emergent pattern within the same structure. We assume here that the infant is not thinking out this plan in advance, but is responding to a motivation—to obtain the toys—in a somewhat automatic way.

Let us look at higher-level processes of understanding. Take Searle's example of a person in the Chinese Room. His instructions are to take the input, consisting of Chinese characters, look them up in a book, and return the prescribed different set of characters through the output slot. Certainly he doesn't understand the characters. But he does have a clear understanding of what he is supposed to do. Not only can he express them in English (as he does in his article), he understands them in terms of our basic pattern of reach-grasp-pull. He reaches for them as they come in the input slot, grasps them, and pulls them to him and looks at them. This case of understanding is, of course, very simple, and probably minimally conscious. The notion that reach-grasp-pull clearly applies here, and can be clearly extended to apply to any abstract cognitive tasks in the "grasp" mode.

I have described the application of pre-learned sensorimotor patterns to examples of simple tasks to make clear the recursive activities involved in a range of cases of emergent understanding. One more aspect of understanding, seen as a recursive activity built upon basic patterns, is the motivation that leads the understanding subject to apply the patterns, with growing sophistication, to the constantly arriving new states of affairs that must be incorporated into the subject's world view. How do we know to keep applying the same basic patterns to novel input? We need a concept to express the growing facility with which we incorporate novel states of affairs into our world-view. Why do we not struggle for understanding in the case of radically novel input, not to mention the constantly changing environment with which we are confronted moment by moment? Not only is there normally no struggle, but the basis for understanding a novel state of affairs is in place *before* we can puzzle over the scene. The general schema for understanding our world allows an even flow from one scene to another, seamlessly, because we are motivated to "look for" such a framework before the event of new sensory input.

The recursive building activity that lets us feel at home in the familiar but changing world is known as *stigmergy*. As Camazine et al. (2001) explains, referring to stigmergy:

> [a] process of decentralized coordination ... where individuals respond to stimuli provided by the emerging structure itself can be a rich source of information for the individual. In other words, information from the local environment and work-in-progress can guide [and motivate through positive feedback] further activity. As a structure such as a termite mound develops, the state of the building process continually provide[s] new information for the builders (p. 23).

The preceding quotation applied in the original to termites, but it can apply equally well to cognitive understanding in an intelligent individual. As we grow in the world, novel conditions inspire more use of sensorimotor patterns to understand them. Like a growing termite mound, our growing understanding of the world, via *stigmergy*, supplies a conscious sense of satisfaction and guards against confusion at first encountering the new situation. The individual termites have innate instructions that lead each one to coordinate its activities with the others, while being unaware of the activities of the others and focusing on its own tasks. In the case of us humans, we need more conscious motivators; our pattern-use is not a result of blind innate drives. Our motivators, as mentioned earlier, are conscious feelings of satisfaction and discomfort. We are normally uncomfortable when confused, and seek understanding. When we have achieved that, we are satisfied; our work at building our extended world is completed. We seek to be *at home* in the world, not at a loss. When at home, we know what we can do next. We can simply say that the more we understand the better we feel.

So far I have discussed the use of sensorimotor patterns to incorporate novel situations into a given cognitive framework. Sometimes, more rarely, there appear cases of true novelty, structure-breaking events that require almost complete re-evaluation in terms of the subject's previous system of understanding. All of a sudden all the lights go out. It is pitch-dark outside my open window, and surrounding me inside. Nothing has prepared me for this. It is as though I have gone blind in an instant. There are no clues as to what I should, or even can, do. What next?

My own physical body is my sole remaining anchor. I can take stock of where my limbs are and what objects I can touch. Two new interpretations are available to me: I am totally blind, or something has happened to the light sources outside me. The latter seems a more hopeful option; I try to construct plausible scenarios, and finally find one involving unnoticed growing lateness of the hour (the outside darkness), and a major electrical fuse blowing (the inside darkness). That works.

In true novelty, when I would not have even my body to anchor me, there might be no plausible scenario. In such an extreme case, with no familiar structures to turn to, I could only search my cognitive repertoire for some logically possible explanation, in terms of what I am acquainted with. If that search failed, as it would with no proprioceptive input whatever, I can suppose that only blind panic would take over. That state is unimaginable to a normally embodied subject (like this author). A brave attempt to imagine it can be found in the novel "Zero K" (DeLillo, 2016, p.155ff) when a newly disembodied subject first awakens to her situation). The conclusion I draw is that the body is foundational for any self-aware cognition with which to construct some degree of understanding. With the body I can tell some sort of story; without the body, there is nothing—even memories of embodiment would not locate *me now*. Truly novel situations are possible only against a background of minimal familiarity; take that away, and subjective cognitive activity must cease. But with some element of familiarity, which must include some degree of embodiment, a possible world might be constructed to fit my experience. Understanding requires embodiment, and thus any understanding will be structured by actual or possible bodily actions (Boden, 1990).

Returning to the subject of self-organization, one might argue that the preceding is not an explanation of self-organization, since we, the subjects, consciously monitor the construction of understanding, which means that the successive states of applying motor patterns to our environment are not independent of "intervention by external directing influences." And it is true that the subject of the understanding is consciously trying to understand, motivated by a need, and that she therefore accepts or rejects candidate patterns with which to organize the new data. Nevertheless, she selects or rejects by means of her conscious feelings only, and not by an independent standard of fitness of the pattern with what has gone before. She need not know that the successful pattern is a metaphor for the basic patterns of movement that she has carried with her from infancy (and as I have discovered, may well deny it when it is suggested to her!). Here is a personal example: I was trying to convince a colleague that his memory of a bad experience at a party was structured by a sensorimotor pattern from earlier experiences, and that his memory of the party was composed of representations of those experiences, and not of purely linguistic representations such as "I had a bad time at the party." He denied it, so I asked "when you think of the party, is there any bodily reaction that you are aware of?" and he answered "I wince." That convinced him (at least partially) that such memories are not independent of personal patterns of reactions to unpleasantness.

The point is that the unpleasant experiences used to structure the memory of the party, and motivate the wincing, were in place before my colleague reacted to them. The feelings of satisfaction that arise when we find a pattern with which to interpret anything new are automatic, and we can then proceed to use our understanding with confidence, knowing that we have successfully expanded our experienced world. And, as we see with the example of sudden, complete sensory deprivation, motivation to construct a pattern would be baseless, and could not begin. To be a self is, necessarily, to be located with respect to an environment. If that is gone, nothing remains.

CONCLUSION: UNDERSTANDING AS A SELF-ORGANIZING PROCESS

Imagine what it would be like if all minute-to-minute attempts at understanding of our experienced environment were fully conscious and deliberate. There would be no comfortable feeling of being "at home" in the world. Instead, there would be constant confusion and uncertainty, at worst a deep fear of the immediate future as something we cannot, but must, prepare for. In Being and Time, Heidegger describes the state of humans, Dasein, as Being-in-the-World:

> In the projecting of understanding, entities are disclosed in their possibility. The character of the possibility corresponds, on each occasion, with the kind of Being of the entity which is understood. Entities within the world generally are projected upon the world—that is, upon a whole of significance, to whose reference relations concern, as Being-in-the-world, has been tied up in advance. When entities within-the-world are discovered along with the Being of Dasein—that is, when they have come to be understood—we say that they have *meaning [Sinn]*. But that which is understood, taken strictly, is not the meaning but the entity, or alternatively, Being. Meaning is that wherein the intelligibility of something maintains itself (Heidegger, 1927).

In Heidegger's terms, lacking an understanding of the kind of Being of an encountered entity is lacking a knowledge of one's possibility in this novel situation. But awareness of one's possibilities of acting is precisely what makes a situation comfortable; we are "at home" in the world when we know what we can do next. Not knowing that would be a condition of fear and hopelessness; our environment would be meaningless. As the quoted passage implies, the significance of our world has been "set up in advance," meaning that as we move through time we bring with us the structures through which we can understand novel entities. Thus, finding a structure for interpreting newly-encountered entities is not a constant anxiety-ridden necessity but, we may say, a self-organized process that can guarantee our unbroken comfort in our world. This means that the process must be self-organizing, for otherwise we would have no time for acting upon possibilities, but would be in constant destabilizing fear. The conditions for understanding, the "significance of our world" are essential for the existence of possibilities. If dependent upon our conscious organizing powers, each new present moment would be a new cause for alienation and anxiety. That we are not, as a rule, constantly in that state of extreme anxiety, seems to be strong evidence that understanding is a self-organizing process.

Many of the arguments in Sections (a) through (c) are given fuller treatment in Newton (1996).

AUTHOR CONTRIBUTIONS

NN is the sole author and responsible for all research in the paper.

REFERENCES

Alibalia, M., Evans, J., Hostetterc, A., and Ryana, K. (2009). Gesture-speech integration in narrative: are children less redundant than adults? *Gesture* 9, 290–311. doi: 10.1075/gest.9.3.02ali

Anderson, P. W. (2011). *More and Different: Notes from a Thoughtful Curmudgeon*. Hackensack, NJ: World Scientific.

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *J. Gen. Psychol.* 37, 125–128. doi: 10.1080/00221309.1947.9918144

Boden, M. (1990). "Implications of language studies for human nature," in *Language, Mind and Brain*, eds T. W. Simon and R. J. Scholes (Hillsdale, NJ: Lawrence Erlbaum), 129–143.

Bonabeau, E., Dessalles, J.-L., and Grumback, A. (1995). Characterizing emergent phenomena: a critical review. *Rev. Int. Syst.* 9, 327–346.

Camazine, S., Deneubourg, J.-L., Franks, N., Sneyd, J., Theraulaz, G., and Bonabeau, E. (2001). *Self-Organization in Biological Systems*. Princeton, NJ: Princeton University Press.

DeLillo, D. (2016). *Zero K*. New York, NY: Scribner Publishers.

Ellis, R. (2012). "Reduction versus emergence," in *The Encyclopedia of Clinical Psychology*, eds R. Cautin and S. Lilianfeld (Wiley Online Library).

Ellis, R., and Newton, N. (2010). *How the Mind Uses the Brain*. Chicago, IL: Open Court Press.

Haken, H. (2008). Self-organization of brain function. *Scholarpedia* 3:2555. doi: 10.4249/scholarpedia.2555

Heidegger, M. (1927). *Being and Time*. New York, NY: Harper and Brothers. Transl. by J. Macquarrie, and E. Robinson (1962).

Jeannerod, M. (1988). *The Neural and Behavioral Organization of Goal-Directed Movements*. Oxford: Clarendon Press.

Karsenti, E. (2008). Self-organization in cell biology: a brief history. *Nat. Rev. Mol. Cell Biol.* 9, 255–262. doi: 10.1038/nrm2357

Kerner, B. S. (1998). Experimental features of self-organization in traffic flow. *Phys. Rev. Lett.* 81, 3797–3800. doi: 10.1103/PhysRevLett.81.3797

Lakoff, G., and Johnson, M. (1987). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav. Brain Sci.* 8, 529–566. doi: 10.1017/S0140525X00044903

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Newton, N. (1996). *Foundations of Understanding*. Amsterdam: John Benjamins Publishing Company.

Pearson, J., and Kosslyn, S. (2015). The heterogeneity of mental representation: ending the imagery debate. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10089–10092. doi: 10.1073/pnas.1504933112

Prigogine, I., and Nicolis, G. (1977). *Self-Organization in Non-Equilibrium Systems*. New York, NY: John Wiley and Sons.

Sacks, O. (1984). *A Leg to Stand on*. New York, NY: HarperCollins Publishers, Inc.; Simon and Schuster, Inc.

Searle, J. (1984). *Minds, Brains and Science*. Cambridge: Harvard University Press.

Yufik, Y. M., and Friston, K. (2016). Life and Understanding: the origins of "understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 09 December 2016
doi: 10.3389/fnsys.2016.00098

Life and Understanding: The Origins of "Understanding" in Self-Organizing Nervous Systems

Yan M. Yufik[1]* *and Karl Friston*[2]

[1] Virtual Structures Research, Inc., Potomac, MD, USA, [2] Wellcome Trust Centre for Neuroimaging at UCL, London, UK

This article is motivated by a formulation of biotic self-organization in Friston (2013), where the emergence of "life" in coupled material entities (e.g., macromolecules) was predicated on bounded subsets that maintain a degree of statistical independence from the rest of the network. Boundary elements in such systems constitute a *Markov blanket*; separating the internal states of a system from its surrounding states. In this article, we ask whether Markov blankets operate in the nervous system and underlie the development of intelligence, enabling a progression from the ability to sense the environment to the ability to understand it. Markov blankets have been previously hypothesized to form in neuronal networks as a result of phase transitions that cause network subsets to fold into bounded assemblies, or *packets* (Yufik and Sheridan, 1997; Yufik, 1998a). The ensuing neuronal packets hypothesis builds on the notion of neuronal assemblies (Hebb, 1949, 1980), treating such assemblies as flexible but stable biophysical structures capable of withstanding entropic erosion. In other words, structures that maintain their integrity under changing conditions. In this treatment, neuronal packets give rise to perception of "objects"; i.e., quasi-stable (stimulus bound) feature groupings that are conserved over multiple presentations (e.g., the experience of perceiving "apple" can be interrupted and resumed many times). Monitoring the variations in such groups enables the apprehension of behavior; i.e., attributing to objects the ability to undergo changes without loss of self-identity. Ultimately, "understanding" involves self-directed composition and manipulation of the ensuing "mental models" that are constituted by neuronal packets, whose dynamics capture relationships among objects: that is, dependencies in the behavior of objects under varying conditions. For example, movement is known to involve rotation of population vectors in the motor cortex (Georgopoulos et al., 1988, 1993). The neuronal packet hypothesis associates "understanding" with the ability to detect and generate coordinated rotation of population vectors—in neuronal packets—in associative cortex and other regions in the brain. The ability to coordinate vector representations in this way is assumed to have developed in conjunction with the ability to postpone overt motor expression of implicit movement, thus creating a mechanism for prediction and behavioral optimization via mental modeling that is unique to higher species. This article advances the notion that Markov blankets—necessary for

OPEN ACCESS

Edited by:
Jonathan B. Fritz, University of Maryland, College Park, USA

Reviewed by:
Hal S. Greenwald, The MITRE Corporation, USA
Steven L. Bressler, Florida Atlantic University, USA
Robinson E. Pino, Air Force Research Laboratory, USA
Simon Berkovich, George Washington University, USA
Alessandro Sarti, CNRS-EHESS, France

***Correspondence:**
Yan M. Yufik
imc.yufik@att.net

Received: *20 April 2016*
Accepted: *08 November 2016*
Published: *09 December 2016*

Citation:
Yufik YM and Friston K (2016) Life and Understanding: The Origins of "Understanding" in Self-Organizing Nervous Systems. Front. Syst. Neurosci. 10:98. doi: 10.3389/fnsys.2016.00098

the emergence of life—have been subsequently exploited by evolution and thus ground the ways that living organisms adapt to their environment, culminating in their ability to understand it.

Keywords: understanding, consciousness, neuronal packets, variational free energy, thermodynamic free energy

INTRODUCTION

This article offers a synthesis of recent developments in theoretical neurobiology and systems neuroscience that may frame a *theory of understanding*. We suggest that cognitive capacities, in particular understanding, are an emergent property of neuronal systems that possess conditional independencies. In this view, cognition is predicated on associative neuronal groups—or assemblies—that form bounded structures (*neuronal packets*) whose Markov blankets maintain a degree of statistical independence from each other. Such quasi-stable, quasi-independent structures capture regularities in the sensorium, giving rise to the perception of "objects"; namely, the external causes of sensations. These neuronal packets are context-sensitive but maintain their structural integrity. They are composed to form mental (generative) models that reflect the coordinated dynamics of "objects" in the world that cause sensory inputs.

Our basic thesis is that conditional independencies in the causal structure of the world necessarily induce neuronal packets with a similar statistical structure. In effect, the brain "carves nature at its joints" using statistics—to capture the interaction among the factors or causes of sensory data. The implicit factorization of probabilistic representations provides an incredibly efficient process to infer states of the world (and respond adaptively). In physics, this carving into marginal probability distributions (i.e., factors) is known as a *mean field assumption*. Here, we suggest that many aspects of the brain can be understood in terms of a mean field assumption; from the principle of functional segregation, through to the dynamic and context-sensitive maintenance of neuronal packets, groups or cell assemblies. The ensuing theory casts the interaction between the brain and the environment as an allocation of (representational) resources; serving to minimize free energy and thereby maintain homoeostasis (and allostasis).

Variational free energy will figure recurrently in our arguments. Variational free energy is a statistical construct that provides a mathematical bound on surprise or self information (i.e., the improbability of some sensory data, under a generative model of those data). Crucially, free energy is a functional of a (posterior) probability distribution or "belief" about the causes of sensory data—as opposed to a (surprise) function of sensory data *per se*. This means that when a system minimizes its free energy, it is implicitly optimizing its "belief" about the objects that are causing sensory input—based upon an internal or generative model of how that input was caused. Free energy is the difference between accuracy and complexity. This means that minimizing free energy provides an accurate explanation for input that is as simple as possible (where complexity can be construed as a cost function). This complexity reducing aspect of free energy minimization will be important in what follows.

From the point of view of a phenotype, success rests on a deep "understanding" or modeling of the environment. In other words, phenotypes that anticipate and avoid surprising (high free energy) exchanges with their environment possess a generalized form of homoeostasis and implicitly minimize surprise and uncertainty. "Understanding" can therefore be construed as a resolution of surprise and uncertainty about causal structure and relationships in the environment—and in particular the relationship of self to the environment (and others). Differences in adaptive efficiency—between humans and other species—may be determined by formal differences in the generative models used to predict and understand environmental changes over different temporal scales: for example, deep models with hierarchically organized representations vs. shallow models that preclude context-sensitive repertoires of behavior.

This article starts with an overview, followed by four sections: section I reviews theories of understanding in the literature, section II outlines our theoretical proposal, section III presents some empirical findings and examines the correspondence, or absence of such, between our theory and other proposals, section IV re-visits our main suggestions, placing them at the intersection of thermodynamics, information and control theories in systems neuroscience. Our focus in this section is on reconciling the variational (free energy) principles (based upon statistical formulations) with the thermodynamic and homoeostatic imperatives of living organisms—and how these imperatives may furnish a theory of understanding.

Overview

We pose the following questions:

1. What is "understanding"?
2. What does "understanding" contribute to the overall function performed by the nervous system?
3. What are the underlying mechanisms?
4. How do mechanisms—that can be described in terms of physical processes or information processes (abstracted from physics)—reconcile in a theory of understanding?
5. How does the theory reconcile current views concerning the anatomy and functional architecture of the nervous system?
6. How can one express the theory in a tractable formalism?
7. What is the difference between learning (without understanding) and (learning with) understanding?
8. If the formalism is tractable, what would it entail?
9. What is the key proposal that follows from these considerations?

The article claims no complete answers but suggests where useful answers could be sought. Our framework is system-theoretic, focusing on the general principles of operation in the nervous system. We call on eleven notions: Markov blankets, neuronal packets, self-adaptive optimization, folding, enfolding, unfolding, virtual associative networks, mental modeling, negentropy generation, surface tension and cognitive effort. These and other notions have been elaborated previously (Yufik, 1998a, 2002, 2013; Friston, 2013). For convenience, they are rehearsed briefly in a glossary (please see "Glossary of Terms" below) and will be unpacked as necessary throughout the article.

Glossary of Terms

A Markov blanket is a set of nodes in a network forming an interface between the nodes that are external and internal to the blanket. The conditional dependencies among the nodes endow internal and external nodes a degree of statistical independence within the network: i.e., they are conditionally independent given the states of the nodes in the Markov blanket.

Neuronal packets are bounded assemblies (subnetworks) forming spontaneously in associative networks and possessing boundary energy barriers that separate them from their surrounds. Neuronal packets are physical instantiations of Hebbian assemblies, as opposed to information processing abstractions, leading to the conclusion that free energy barriers must exist at the assembly boundary (Yufik, 1998a). This notion predates recent formulations of memory systems as physical devices, as opposed to circuit theory abstractions, and suggests that free energy barriers must exist to "protect" memory states from dissipation (dubbed "stochastic catastrophe"; Di Ventra and Pershin, 2013). Hebbian assemblies devoid of protective energy barriers are subject to "stochastic catastrophe" and dissipate quickly: hence, neuronal packets.

Self-adaptive resource optimization is taken to be a principle of operation in the nervous system: the neuronal packet hypothesis views cognitive processes and cognitive development as an optimization of neuronal resources, and considers spontaneous aggregation of neurons into packets as the key mechanism. Thermodynamic energy efficiency is the optimization criteria: the system seeks to maximize extraction of free energy from the environment while minimizing internal energy costs incurred in mobilizing and firing neurons (Yufik, 2002). Resource optimization implies adaptation to changes in the environment as well as to those occurring inside the system (hence, the *self-adaptation*). The notion that spontaneous aggregations (assemblies) of neurons constitute functional units in the nervous system was originated by Hebb, and continues to play a prominent role in theories of neuronal dynamics that focus on the mechanisms of coordination, segregation and integration (e.g., Bressler and Kelso, 2001; Razi and Friston, 2016).

Folding denotes the spontaneous formation of regions in networks of interacting units acquiring a degree of statistical independence from their surrounds (i.e., formation of Markov blankets at the boundary). We assume that life emerges in networks that are amenable to folding; thereby regulating material and energy flows across the boundary. This article offers a unifying theoretical framework and explanatory principle for life (and intelligence) that rests on the formation of Markov blankets. The synthesis may reconcile thermodynamic and information-theoretic accounts of intelligence.

Enfolding and *unfolding* denote cognitive (deliberate, self-directed) operations on packets: unfolding operates on the internal states of a packet while enfolding treats packets as functional units. Mathematically, enfolding involves computing packet response vectors (the sum of neuronal response vectors), while unfolding reverts to the constituent response vectors. Cognitive processes alternate between enfolding and unfolding; namely, alternating between integrative and focused processing modes. For example, alternations between groups of units ("situations" comprising interacting "objects") and a focus on particular features of such units ("objects") and their changes as the situation unfolds. Computationally, the process alternates between matching packet response vector to the input and matching neuronal response vectors. Perceptually, the process manifests, e.g., in grouping visual targets into units, or "virtual objects" and tracking the units, alternating with focusing on and tracking individual targets (Yantis, 1992).

Virtual associative networks denote associative networks undergoing self-partitioning (folding) into packets. Mathematically, packets are obtained as minimum-weight cutsets (Luccio-Sami, or LS-cutsets) in networks where nodes are neurons and link weights are determined by the relative frequency of their co-firing (Hebb's co-firing rule). LS-cutsets "carve out" subsets (packets), such that internal nodes are connected more strongly to each other than to external nodes. In this way, self-partitioning into packets produces a coarse representation of statistical regularities in the environment. Statistically, the nodes of a packet—from which the LS links emanate—constitute its Markov blanket. In other words, they form a boundary, engendering a degree of statistical independence between the packet and its surrounds. Physically, the independence is maintained by energy barriers. The process is similar to structure acquisition in unsupervised learning, except that the quality of learning is adjudicated by thermodynamic constraints. Figuratively, neuronal packets can be viewed as Hebbian assemblies "wrapped" in Markov blankets.

Mental modeling denotes self-directed (deliberative, attentive) composition of packets into groups (*mental models*) such that mutual constraints in the packets' responses can be explored in search of a best fit between implicit models of stimuli. Attaining a good enough fit underlies the experience of reaching, grasp, or understanding. The process improves on and fine-tunes the results of spontaneous packet formation. Mental modeling allows anticipation and simulation of future conditions, and initiating preparations before their onset (anticipatory mobilization of neuronal resources), thus providing a mechanism of neuronal resource optimization.

Understanding is a form (component) of intelligence. Intelligence denotes the ability of a living organism to vary its responses to external conditions (stimuli) in a manner that underwrites its survival; e.g., a sunflower following the sun is a manifestation of "plant intelligence" (Trevawas, 2002). Learning is a form of intelligence involving memory and subsequent reproduction of condition-response associations. On the present theory, understanding denotes the ability to compose and manipulate mental models representing persistent stimuli groupings, or "objects", their behavior under varying conditions, and different forms of behavior coordination (i.e., relations between objects). Understanding overcomes the inertia of prior learning and enables construction of adequate responses under novel and unfamiliar circumstances.

Negentropy generation denotes production of information and increases in the order of a system as a result of internal processes. The distribution of weights in associative networks is the result of information intake from the environment (negentropy extraction). Self-directed composition of packets into models increases internal order, without further information intake and without impacting the weights; hence, negentropy generation. Mental modeling amounts to endogenous production of information requiring energy expenditure, the payoff is an increase in adaptive efficiency; i.e., the ability to extract energy from the environment under an expanding range of itinerant conditions. This mechanism enables productive thinking that is sustained by information inflows but is not limited by them.

Surface tension is a general thermodynamic parameter defining the thermodynamically favored direction of self-organization in a system. Surface tension corresponds to the amount of free energy in the surface. The neuronal packet hypothesis attributes formation of packets in virtual associative networks to phase transitions (Haken, 1983, 1993; Fuchs et al., 1992; Freeman and Holmes, 2005; Kozma et al., 2005) and accumulation of thermodynamic free energy across boundaries. Boundary free energy barriers are responsible for a packet's resilience; i.e., the ability to persist as cohesive units—resisting dissipation under fluctuating conditions and entropic erosion.

Cognitive efficiency denotes the ratio of free energy extraction (from the environment) and internal energy costs incurred in sustaining energy inflows. The higher the ratio, the higher the efficiency. Mental modeling involves expending free energy to increase internal order (generate negentropy), which entails a more efficient (robust under a wide range of circumstances) energy extraction.

Cognitive effort denotes expenditure of thermodynamic free energy incurred in mental modeling. Our theory of understanding associates consciousness with the process—and subjective experience—of exerting cognitive effort. Exerting effort alternates with (relatively) effortless release of genetically supplied and/or experientially acquired (learned) automatisms. Consciousness accompanies the work of suppressing the inertia of prior learning, adjusting learned responses to the current conditions, and composing new responses to anticipate environmental fluctuations. In short, the experience of consciousness is rooted in a high-level mechanism of self-organization and self-adaptive resource optimization in the nervous system. This article focuses on the mechanisms of understanding, postponing a detailed discussion of consciousness for the future.

With these notions in place, the answers to the questions above can be framed as follows:

1. Understanding rests on mental (generative) models representing objects, their behavior and behavioral coordination (i.e., mutual constraints on the behavior of objects).
2. Generative models serve to optimize an organism's control of its own behavior in a changing environment in the interests of survival (i.e., enduring preservation of structural integrity). The advent of the capacity to understand offered a quantum leap in control efficiency.
3. Control optimization in a changing environment requires anticipatory mobilization of neuronal resources; i.e., progressively improving the ability to select and arrange neuronal representations before the onset of stimuli. Conditioning is the most basic anticipatory mechanism that is shared by all species. The evolution of conditioning to understanding may have proceeded in three stages, predicated on the packet mechanism: Packets capture recurring stimuli groupings. As a result, control efficiency (as compared to conditioning) improved in two ways—by increasing the probability of successful representation and by reducing the cost (i.e., complexity) of internal processing. The formation of packets underlies the perception of *objects*; i.e., bounded stimulus-bound groupings distinct from the sensory background. In the next evolutionary step, the ability to optimize packet allocations (selectively inhibit/amplify neuronal activity within packets) emerged. This ability underlies the apprehension of *behavior*; i.e., changes that objects can sustain without losing their self-identity. Finally, the ability to orchestrate the allocation of packets emerges, giving rise to the apprehension of *relations*; i.e., different forms of behavioral coordination among groups of objects. Apprehending relations requires abstraction from the sensory contents (enfolding): e.g., the relationship of the type "*A rests on B*" defines how the behavior of A coordinates with the behavior of B and vice versa, regardless of how A and B look, smell, sound, etc. Inducing coordinated variations in packet arrangements constitutes *mental modeling*. This capacity supports anticipation into the indefinite future, accounting for large (perhaps, indefinitely large) sets of environmental contingencies.
4. Neuronal firing expends energy. Survival (free energy minimization) is predicated on minimizing the computational cost or complexity of adaptive processing that enables accurate matching of neuronal representations to objects in the environment. In other words, thermodynamic and informational imperatives cannot rely on transitory fluctuations in the system. Instead, a mechanism is needed which produces neuronal structures that withstand entropic erosion and are implicitly available for reuse. It has been suggested previously that neuronal packets are produced

by phase transitions in associative networks—and are maintained by "tension" in the surface separating the phases. From an information-theoretic standpoint, *mobilizing* a packet corresponds to inducing a neuronal hypothesis that a particular neuronal packet will provide the best explanation for upcoming sensory input. Accordingly, thermodynamic and information-theoretic approaches converge: the principle of thermodynamic free energy minimization on the packet surface corresponds to the principle of variational free energy minimization in probabilistic inference (Friston et al., 2006; Friston, 2010), both principles referring to the same neuronal mechanism that transcends thermodynamic and variational principles.

5. In what follows, packet variations (selective inhibition/amplification) will be represented as rotation of (population) vectors computed over the internal neuronal states of a packet. On that notion, mental modeling involves the coordinated rotation of packet vectors. For example, motor control is known to entail coordinated rotation of population vectors in the motor cortex. It is not unreasonable to assume that rapid evolution of intelligence in humans expanded the elaborate apparatus of sensorimotor coordination in hominids—to allow packet coordination in the associative cortex and other regions in the brain.
6. The formalism of packet vector coordination for control optimization (self-adaptive allocation of neuronal resources) appears to be tractable.
7. Learning without understanding confines performance to situational envelopes narrowly constrained by past exposures. Understanding expands the envelope indefinitely, enabling counterfactual ("what if") modeling, simulation of the future—and an implicit ability to "anticipate" the consequences of action.
8. Developing the formalism may help design artifacts to progressively improve their ability to carry out complex tasks, under unfamiliar conditions and unforeseen circumstances.
9. A formal theory appears to be within reach, centered on the notion of Markov blankets, offering a parsimonious account of intelligence that encompasses the transition from inanimate matter to organismal self organization—and from simply sensing the environment to understanding it.

In summarizing, an example may help bring together the perspective on offer: one learns to play chess by first learning to recognize pieces. Learning proceeds by associating different behavioral rules with chess pieces and culminates in the ability to apprehend behavioral constraints (e.g., this black pawn blocks diagonal movement of that white Bishop). Understanding chess involves the ability to apprehend constraints across a composition of pieces—and to determine the possibilities for coordinated maneuvers the composition affords (e.g., "attack on the left flank"). Apprehending behavior coordination requires abstraction (e.g., pin is a form of coordination where the pinned piece shields a more valuable piece behind it). The variety of positions affording this type of coordination is practically infinite. "Chess intuition" collapses its combinatorial space into "lines of play" (Beim, 2012), thus enabling analysis (e.g., 15 moves look-ahead analysis by chess masters (Kasparov, 2007) can be compared to tracing a hair-thin line in combinatorial Pacific Ocean).

THEORIES OF UNDERSTANDING

Aristotle's Metaphysics (350 BC) opens with a statement traditionally translated as "All men by nature desire to know." Contrary to traditional interpretations, recent analysis (Lear, 1988) suggests that the statement permits a dual interpretation—"to know" and "to understand"; with the latter interpretation being closer to the original intention. Cognition grows out of the capacity to experience puzzlement, accompanied by the feeling of discontentment and desire to resolve it. This capacity to resolve uncertainty is shared by many animals. But only in humans is the desire to resolve uncertainty not fully discharged until a complete understanding is attained (Lear, 1988). Aristotle observed that "animals other than man live by appearances and memories but little of connected experience. . ." and attributed to men the ability to form connections, i.e., organize disparate data into connected structures. "Wisdom" is attained when such structures reveal causes:

> *". . .men of experience know that the thing is so but do not know the why, while the others know the "why" and the cause"*
> *—(Metaphysics, book 1).*

What progress has been made since Aristotle in uncovering the inner workings of understanding? The problem remained largely unaddressed for over two millennia but became prominent in philosophical discourse in the XVIII—XIX centuries (Hume, Spinoza, Berkeley, Kant, Descartes, et al). However, it was not until the middle of the last century that the scope of discourse was radically expanded; largely in response to challenges faced in scientific enquiry, where rapidly accumulating data resisted traditional modes of understanding and explanation (e.g., Bunge, 1979; Cushing, 1994; Sloman, 2005). Philosophy was joined by psychology and cognitive science and, more recently, by what could be defined as *physics of the mind*—an emergent discipline combining statistical physics, information theory and neuroscience to elucidate neuronal underpinnings of cognition (Penrose, 1989, 1994, 1997; Friston et al., 2006; Friston, 2010, 2013). The *physics of mind* framework is consistent with the "enactive" view, deriving cognition from an interplay between external conditions and self-organization in the nervous system. In other words, (non-radical) forms of enactivism enable prediction to guide action on the environment that ensures survival (e.g., Thompson and Varela, 2001). Self-organization places the nervous system in the domain of dissipative systems that are thermodynamically open to the environment. Our proposal for a theory of understanding is thus formulated within *the physics of the mind* framework.

Research areas relevant for understanding include the study of language, consciousness, intentionality, explanation, causality and prediction, logic and reasoning, inference, attention, etc. A detailed review of the relevant research is impossible and is not intended here. What follows is a summary of findings that address some key aspects of the function of "understanding".

Webster's Ninth New Collegiate Dictionary defines understanding as comprehension or "mental grasp, the capacity to apprehend general relations of particulars". This suggests that "understanding" requires a (generative) model that embodies general relationships of particulars; i.e., model that can generate particular consequences from general causes (Craik, 1943; Gentner and Stevens, 1983; Johnson-Laird, 1983, 1989, 2003; Sanford, 1987). Theories of understanding can be roughly organized in five groups, focusing on the different roles of generative models in understanding: (a) volitional (self-directed, deliberate) activity; (b) simulation; (c) need satisfaction and optimization; (d) unification, explanation and prediction; and (e) problem solving. We will reference exemplar theories in each of these groups,—and attempt to relate them to the *physics in the mind* approach.

Understanding Results from Volitional Operations Targeting Inputs from the Outside and Representations on the Inside

The "foundational theory of understanding" (Newton, 1996) asserts that understanding results from volitional (deliberate, self-guided) actions that involve directing one's attention to sensory inflows and reconciling current sensations with memory structures in a manner consistent with the current intentions, or goals.

The volitional aspect of cognition is emphasized in the theory of mind-body relationships in Humphrey (2000, 2006). This theory traces volitional activities to their evolutionary origins, as follows. A primitive organism senses physical conditions, or stimuli occurring at its boundary surface and generates commands targeting locations on the surface where the *conditions* were sensed. Commands are said to generate "wiggles" on the surface, the substrates of sensing are not the conditions but the type of "wiggles" produced by the organism adapting to those conditions (e.g., sensing "red" is produced by "wiggling redly," sensing "salt" is produced by "wriggling saltily"; i.e., selecting and emitting a response appropriate for the occasion of salt arriving at the surface. Gradually, evolution shifted "response targeting" from surface sites to the efferent, or "sensory nerves" emanating from sites along the surface. Shifting response targets further upstream culminated in the emergence of mechanisms confining responses to internal loops—comprised of efferent and afferent links. In such loops, afferent signals become "as-if commands" (i.e., models): they would have produced appropriate behavior had they been carried all the way to the sensorimotor periphery (Humphrey, 2000, p. 17).

Central to this formulation is the notion of "targeting"; i.e., self-directed mobilization (or recruitment, Shastri, 2001) and focused allocation of neuronal resources. On that notion, an organism is not just registering the flow of sensory impressions but engages in targeted probing and composition of responses fine-tuned to the data returned by sensory samples (consistent with Noe, 2004; Friston et al., 2014). The notion resonates with the sensorimotor contingency, or "action-in-perception" theory (Noe, 2004) and other theories centered on the idea of the "volitional brain" (Libet et al., 2000; Nunez and Freeman, 2014).

Notice the two key themes of this formulation are an emphasis on active inference or volitional sampling of the world—of the sort that characterizes enactivist or situated approaches to cognition. Second, the progressive elaboration of internalized ("as if") stimulus-response links induces conditional dependencies between the sensory input and internal models of how those predictions were caused—through active sampling.

Understanding Involves Simulation which is Effortful (Work-Consuming)

Two key characteristics are generally attributed to generative models: models are "structural analogs of the world" (Johnson-Laird, 1983), and models allow simulation of processes and events in the world (Chart, 2000). These characteristics are mutually supportive: if two systems (the world and the model) are formally homologous, one can manipulate and observe the behavior of one system (an internal model) in order to predict and postdict the behavior of the other (an external world). In Chart (2000), simulation is taken to be the essence of understanding, enabling one to both anticipate events and to cope with the unanticipated outcomes. Simulation engages "mutors" i.e., physical mechanisms effecting transformations in the models. The simulation system is hierarchical, including "effectors" responsible for combining "mutors" into groups and attributing meaning and values to the groups, and "simulors" responsible for grouping "effectors." Crucially, all stages of grouping involve work. An important insight here is that understanding requires the investment of work performed on or by internal representations.

The notion of understanding via simulation can be traced to Craik (1943), who hypothesized the existence of physical mechanisms in the brain functioning as (generative) models of the environment. The theory of understanding in Chart (2000) substantiates this early hypothesis, bringing to the fore a crucial aspect of mental modeling—the necessity to invest work. This was investigated in detail in Kauffman (2000), who postulated that the ability to perform work is the determining factor in perpetuating life and developing capacities that enable an organism to sustain life in a changing environment, while maintaining relative autonomy from it (the emphasis on performing work in the course of mental operations resonates with Freeman et al. (2012) using generalized Carnot cycle to describe process in the cortex). As formulated in Kauffman (2000).

> *"...an autonomous agent is a self-reproducing system able to perform at least one thermodynamic work cycle...work itself is*

often used to construct constraints on the release of energy that then constitutes further work. Work constructs constraints, yet constraints on the release of energy are required for work to be done"

—(Kauffman, 2000, p. 4.).

We see here a close connection between (variational) free energy formulations of the imperatives for life that we will return to in the next section. In brief, having a formal physics of mind provides a clear link between understanding (minimization of surprise or variational free energy), a concomitant minimization of thermodynamic free energy and the implicit exchange of work and entropy of a system's internal representations (by physical states) and the external world to which it is thermodynamically open.

Understanding Entails Optimization

Generative models improve one's ability to satisfy homoeostatic needs, when navigating an inconstant and capricious environment—and facing predictable changes as well as the unpredicted (Chart, 2000). Adaptive exchange with the environment is thought of as a measure of need satisfaction (Margenau, 1959; Werbos, 1994, 1998; MacLennan, 1998; Pribram, 1998). Under all circumstances, the activity an agent is engaged in is *the best attempt at the time* to satisfy the current need (hence, the optimization; Glasser, 1984; Werbos, 1998).

The key insight afforded by this perspective is that one can cast all adaptive or intelligent behavior as a process of optimizing some value or need function. In physics, this function is variously known as the *Lyapunov function* or Lagrangian. The existence of this function means that intelligent behavior or understanding can be reduced to "approximate constrained optimization" (Werbos, 1994, p. 40). Again, we see a convergence on optimization or minimization imperatives offered by a physics of mind. In the present context, the objective function is (variational) free energy, where biological imperatives or needs are encoded in prior beliefs about the states a particular agent should occupy. These prior beliefs constrain active sampling of the environment to minimize surprise—and thereby search out preferred states.

Interestingly, the minimization of variational free energy in machine learning is also known as approximate Bayesian inference. In other words, the form of internal modeling that we engage in is quintessentially approximate by virtue of minimizing free energy, as opposed to surprise *per se*. This approximate aspect will become particularly important when we appeal to another ubiquitous device in statistical physics; namely the mean field approximation that provides a clear example of partitioning and functional specialization that may be a crucial aspect of generative models in the brain. We will later suggest that the mental modeling—with mean field approximations in humans—obtains a degree of optimization unavailable to other species.

Understanding Entails Explanation

According to the Deductive–Nomological (DN) theory of understanding, phenomenon B is understood if particular conditions A are identified along with some appropriate laws such that, given A, the occurrence of phenomenon B is to be expected (Hempel, 1962, 1965). The DN theory was subsequently augmented to account for unification (rendering phenomenon B dependent on phenomenon A must take place in a broader framework, where the number of independent phenomena is reduced), simplification (Kitcher, 1981) or compression (comprehension is compression) and representation of causality (explanation, von Wright, 1971). Establishing causality involves partitioning of A and re-formulating the question "why B?", as follows:

"Why does this x which is a member of A have the property B?" The answer to such a question consists of a partition of the reference class A into a number of subclasses, all of which are homogeneous with respect to B, along with the probabilities of B within each of these subclasses. In addition, we must say which of the members of the partition contains our particular x"

—(Salmon, 1970, p. 76).

This account of explanation entails an explicit Bayesian formalism (subclasses are hypotheses, encountering B provides evidence) but adds a crucial insight: Explanation is predicated on partitioning heterogeneous A into homogeneous groups, or subclasses. That is, A is a mixed bag, before using the contents for explaining B (and submitting them to Bayesian procedure), they must be sorted into groups that are different (have some features by which they can be told apart) and, at the same time, homogeneous with respect to B. Crucially, partitioning heterogeneous A into homogeneous subclasses is accompanied by production of information and thus requires work. In general, A can admit multiple partitions. Following Carnap (1962), Salmon (1970, 1984, 1989) suggests that the quality of a partition is determined by some utility maximization function imposed at the outset and motivating the investment of work. In this way, Salmon (1970) reveals intimate connections between inference, causality and goal satisfaction.

Establishing causality involves deep inference, or reduction to deeper representation levels (as in seeking the neuronal underpinnings of psychological conditions) as well as determination of intra-level relations (e.g., relating psychological conditions to psychologically traumatic events). Descent to deeper levels in constructing a model (theory) serves to expand the range of surface-level phenomena explained by the model (Dieks and de Regt, 1998). The interplay of the reduction, compression and expansion criteria in constructing models was succinctly defined by Einstein:

"conceptual systems. . .are bound by the aim to permit the most nearly possible certain (intuitive) and complete co-ordination with the totality of sense-experiences; secondly they aim at greatest possible sparsity of their logically independent elements. . ."

—(Einstein, 1949, p. 13).

From the perspective of minimizing variational free energy, the implicit many to one mapping between consequences and causes is captured in the notion of minimizing complexity

(simplification). Complexity corresponds to the degrees of freedom used to explain data accurately (technically, it is the Kullback-Leibler divergence between a posterior and prior belief). This means that an explanation (to the best inference) is one that maximizes model evidence and minimizes complexity by accounting for a diversity of outcomes (consequences) with the smallest number of plausible explanations (partition of causes).

Understanding Enables Problem Solving

Arguably, the most extensive and influential body of psychological research on the role of understanding in problem solving was accumulated by Piaget and his school (Piaget, 1950, 1954, 1976, 1977, 1978, Piaget and Inhelder, 1969). Experiments were conducted with young children, which rendered their findings particularly revealing: the problems studied were elementary and their solutions were uncontaminated by prior experience and associations. The main conclusions boil down to the following: problem solving requires establishing relations between "all the multifarious data and successive data" bringing the relations into "*co-instantaneous mental co-ordination*" within a simultaneous whole (i.e., generative model; Piaget, 1978, p. 219).

The notion that problem solving involves "*co-instantaneous co-ordination*" in generative models, thereby imposing simple explanations for "all the multifarious data and successive data" extends from elementary problems solved by children to the highest reaches of theoretical abstraction:

> *"The general theory of relativity proceeds from the following principle: Natural laws are to be expressed by equations which are co-variant under the group of continuous co-ordinate transformations. . . .The eminent heuristic significance of the general principles of relativity lies in the fact that it leads to us to the search for those systems of equations which are in their general covariant formulation the simplest ones possible. . ."*
>
> *—(Einstein, 1949, p. 69).*

Mathematical equations are expressions of relations between variables; similarly, systems of equations express co-ordination between groups of such relations (Sierpinska, 1994). Accordingly, understanding mathematical formalisms boils down to grasping the relations they entail:

> *". . .if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we "understand" the equation"*
>
> *—(Feynman et al., 1964, cited in Dieks and de Regt, 1998, p. 52).*

Visualization plays a role in problem solving and scientific understanding (van Fraasen, 1980) albeit a limited one. According to self-reports by a number of prominent scientists, the role of verbalization is even less significant (Einstein, 1949; Poincare, 1952; Hadamard, 1954; Penrose, 1989). For example, in his often quoted letter from to Hadamard, Einstein asserts that words hardly participate in his thinking, which consists of "combinatorial play with entities of visual and muscular type. . .words have to be sought for laboriously only in the secondary stage" (Hadamard, 1954, p. 148). Such self-reports are consistent with experimental findings indicating that verbalization does not facilitate problem solving and can, in fact, interfere with the process (Schooler et al., 1993). They also accord with the analysis of causality placing strong emphasis on the notion that mind establishes causal relations based on mental events, as opposed to verbal accounts that are subsequently formulated (Davidson, 1970, 1993).

Summary

If not through words and images, then what is the medium of understanding? The perspectives reviewed in this section implicate complexity reduction through factorization and partitioning to explain heterogeneous data. Accordingly, the cardinal aspects of understanding can be formally summarized in terms of minimizing surprise (or free energy) that necessarily entails a generative model of coordination and relations—a model that provides an accurate (unsurprising) and minimally complex explanation for past sensory inputs and predicts forthcoming experiences, including the likely consequences of one's own actions. We now turn to the mechanisms responsible for such modeling.

TOWARDS A THEORY OF UNDERSTANDING

Following Johnson-Laird (1983), one can distinguish three cognitive mechanisms—symbol processing, image processing and mental modeling: with the latter denoting connected representations and operations on these representations. Our theory is confined to internal modeling, and refers to the process and outcome of such modeling as situational understanding (or *situated cognition*). Cognitive operations underlying the development and exercise of understanding are different from—and do not reduce to—those involved in learning via pattern recognition. The following examples help to appreciate the distinctions.

Fishes can be trained to recognize geometric shapes; e.g., circles (Siebeck et al., 2009). Humans can recognize shapes, name them and, ultimately, define them (e.g., circle is a set of all points in a plane equidistant from the center), which does not yet amount to understanding. A true generative model of a circle comprises representations and operations that enable one to create or manipulate a circle—in practice or "in mind" and at will. For example, the model should account for experiences like handling a circular object, following a circular path, performing circular movements, etc. Having examined a circular object with the eyes closed (e.g., passing a hoop between the palms), one can conjure up an image of a circle; situational understanding manifests, for example, in expecting (not being surprised by) the sensation of a circular edge on palpitating a coin, visually or haptically. These abilities require a generative model; they are distinct from simply recognizing objects or associating symbol strings (names, formulae, descriptions, definitions, etc.)

with such objects. In short, understanding is quintessentially enactive and "embodied" (Lakoff, 2003), requiring one to actively engage with the causes of sensations. In the setting of enactive cognition, this means that understanding requires generative models that define affordances for action offered by sensory cues.

Generative models produce meaning; the meaning of "circle" rests on a model that enables one to do "circling" in the mind (stated differently, the meaning resides in the ability to "wiggle roundly" as the meaning of "red" resides in the ability to "wiggle redly" (Humphrey, 2000)). When fishes are trained to recognize shapes, these shapes acquire significance (predict feeding) but not meaning, fishes form connections but make no sense of them. To appreciate the distinction, note that the definition of "circle" resists visualization (the set of *all* points in a plane equidistant from one point), while the image in your mind is by no means suggestive of the definition. What is then the connection between the definition and the image, what is holding them together? Consider the problem in **Figure 1**.

Group A_1 is not a "circle-like" pattern that can be "recognized" in A, nor group A_2 can be "recognized" as a "point-like" pattern in A, and neither group would be likely to emerge in A had the task been different. Grouping is imputed to A, as opposed to being recognized in—or somehow extracted from—it. The emergence of groups is concomitant with their "co-instantaneous co-variation." Groups A_1 and A_2 are homogeneous with respect to the "go round" variation; the activities of grouping and co-variation in the context of the task yield understanding and determine visualization and verbalization of the solution they produce. To summarize, understanding is yielded by generative models representing objects, behaviors and behavioral constraints. How do such models form and operate in the nervous system?

Representing Objects

Within the theory of neuronal packets, distinct and bounded entities or objects are recovered from sensory streams as a result of folding in associative networks producing bounded subnetworks (neuronal packets). Associative links form between co-firing neurons, where firing is orchestrated by optimization (free energy minimizing) processes allocating neuronal activity to the stream of stimuli. In this view, free energy is the underlying universal currency in the organism-environment exchange: neuronal firing expends and dissipates energy, while successful neuronal activity extracts energy from the environment. The expending-extracting cycle in the formation of links is illustrated in **Figure 2**.

Note the dual nature of the process in **Figure 2**: on the one hand, the process is a thermodynamic cycle, where energy is received and expended in performing work. On the other hand, mobilizing x_j amounts to forming a hypothesis—entailed by x_i—about the identity of the stimulus, with subsequent validation. The two thermodynamic and information-theoretic perspectives are united by the fact that validation comes in the form of a thermodynamic reward and invalidation entails unrecoverable energy consumption. Associative links decay but are reinforced with every subsequent co-firing of linked neurons. Due to response field overlap, across the neuronal system, a connected associative network gradually forms with the distribution of link weights reflecting statistical regularities in the sensory stream (i.e., repetitive co-occurrence of the stimuli). It has been hypothesized (drawing on the principles of Synergetics (Haken, 1983, 1993)) that the development of the network is punctuated by phase transitions, occurring in tightly coupled subnetworks and causing their folding into bounded aggregations (neuronal packets; Yufik, 1998a,b) Packets are internally cohesive and weakly coupled to (have a degree of statistical independence from) the rest of the network. That is, folding induces Markov blankets in the neuronal pool, as illustrated in **Figure 3**.

Again, firing of any neuron within a packet mobilizes the entire packet, amounting to the neuronal hypothesis that subsequent stimuli are likely to come in a cluster represented by the neuronal group within the packet. Packet boundaries

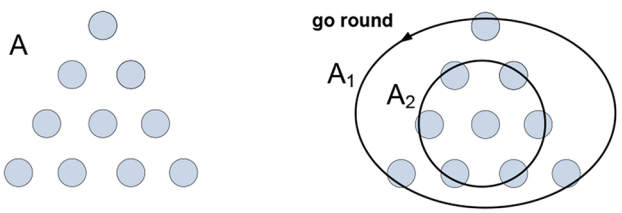


FIGURE 1 | The arrangement of coins A needs to be inverted in a minimal number of moves. The solution is obtained by partitioning arrangement A into groups A_1 and A_2 allowing their co-instantaneous co-variation: A_1 "goes round" A_2. The mental act of "going round" is the medium and the gist of the concept "circle."

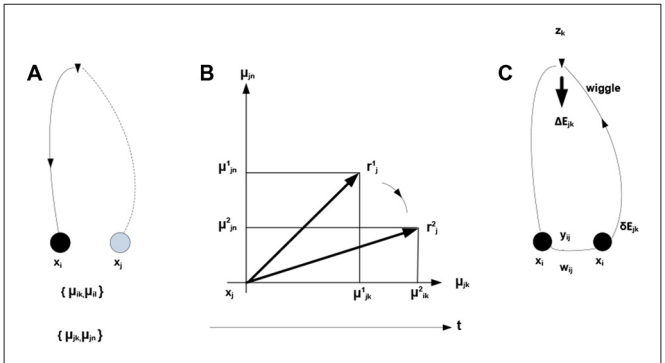


FIGURE 2 | Development of associative links. μ_{ik} denotes probability that neuron x_i fires in the presence of condition z_k. **(A)** Firing of neuron x_i initiates mobilization of neuron x_j having response field overlapping with that of x_i (both neurons respond to condition z_k). **(B)** Mobilization involves adjustments in x_j—the overlap component (x_{jk}) is amplified and the non-overlap one (x_{jn}) inhibited. The adjustment amounts to rotating a neuron's response vector r_j (changing from r_j^1 to r_j^2). **(C)** Firing of x_j (sending a "z_k wiggle") triggers release of energy ΔE_{jk} invested, in part, in producing synaptic modifications establishing an associative link y_{ij} of strength w_{ij}. Mobilization and firing of x_j consume energy δE_{jk} (other energy expenditures are not represented).

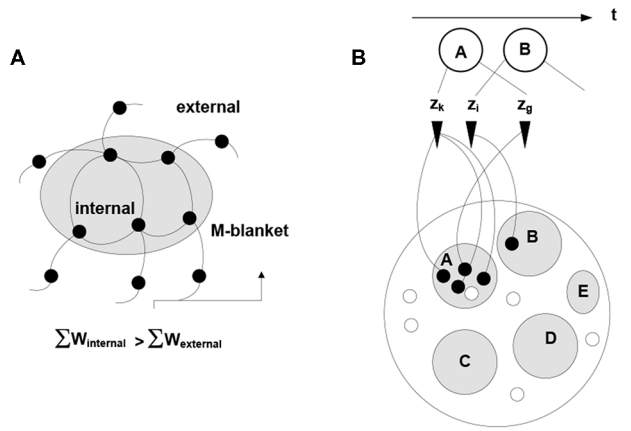


FIGURE 3 | Induction of Markov blankets and formation of neuronal packets. (A) Packets are internally cohesive and weakly externally coupled neuronal groups forming in associative networks as a result of a phase transition. Surface tension in the phase-separating surface causes groups to fold into bounded cohesive units capable of withstanding entropic erosion (Yufik, 1998a, 2013). Folding induces Markov blankets at the packet boundaries, that is, makes packets statistically independent (to a degree) from their surrounds. **(B)** The mechanism recovers persistent stimuli groupings A, B,...interspersed within the stream; thus giving rise to perception of bounded entities, or objects persisting though (re-emerging in) multiple episodes.

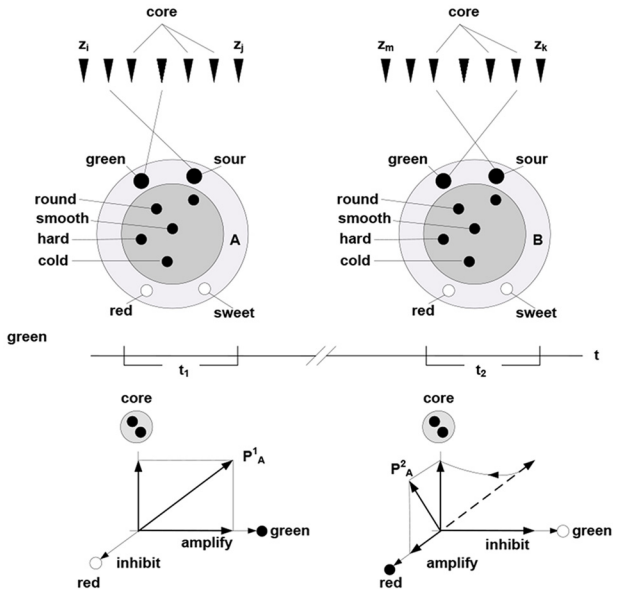


FIGURE 4 | Packet representation of a round, smooth, hard and cold object changing from green and sour in the time period t_1 to sweet and green in the period t_2 (persisting no longer because the object was eaten). The subset of neurons in the fringe admits different transition trajectories between the initial and final condition obtained by selective inhibition-amplification of the constituent neurons. The {green, sour} → {red, sweet} transition (behavior) amounts to rotation of packet vector $\mathbf{P_A}$ in the hyper plane determined by the fringe subset. Behavior of the object over time is represented by the P^1_A to P^2_A rotation.

circumscribe a reference set for the hypothesis, i.e., confine validation probes to the packet internals. Boundary energy barriers discourage but do not prohibit switching reference sets, because unsuccessful probing causes the process to transit to another packet. The packet mechanism is thermodynamically-motivated: energy intakes over time are increased while losses are reduced. If the environment changes, causing diminishing intakes and mounting losses, packets dissolve and are re-constituted.

Representing Behavior

In this formulation, cohesive and bounded neuronal packets act as functional units in the inference process. Stated formally, packet vectors (population vectors) are established on the collectives comprising response vectors of the constituent neurons $\mathbf{P_A} = (r_k, r_h, \ldots, r_g)$, here $\mathbf{P_A}$ is population vector established on packet A. Allocating packets entails their adaptive adjustments, via selective inhibition and amplification of the constituent responses. The persistence of packets establishes an invariant (slowly varying) core in the setting of a variable periphery, which amounts to formation of a hyperplane in the packet's response space; thereby confining rotation of the packet vector. **Figure 4** illustrates representation of behavior via packet vector rotation (ripening apple changes from green and sour to red and sweet).

The rotation of a packet vector does not violate the object's self-identity established by the packet or the ability to induce rotation at will, including reversal (e.g., the green and sour object I experienced earlier and the red and sweet object I experience now are one and the same object, which is established, in part, by my ability to revert to the earlier experience and follow its transformation into the present). Reversibility is a determining characteristic of cognitive mechanisms that enables reasoning (no reasoning is possible if, having initiated a thought, one can't return to the starting point) and apprehending causality (Piaget, 1978).

Representing Coordination

In the present setting, the term "relationship" is taken to denote a form of coordination in the behavior of related objects. Imputing a particular form of coordination to changing (behaving) objects affords a model of the causal dependencies generating sensory data. Establishing coordination in the behavior of objects A and B involves the creation of a bi-directional mapping between the varying subsets (fringe subsets) in the corresponding packets—entailing a coordination of the rotation of packet vectors. **Figure 5** illustrates this notion using a task employed in Piaget, to examine development of understanding in young children: discovering how to use a toy catapult (a plank balancing on support) to hit target objects with a plastic ball. Performing the task requires one to understand that pushing down one side causes the other side to go up. That is, "co-instantaneous coordination" needs to be established (Piaget, 1978).

Three important observations are in order here. First, coordinating objects essentially constrains their behavior; i.e., reduces their degrees of freedom or complexity. Establishing coordination between objects in the course of some inference

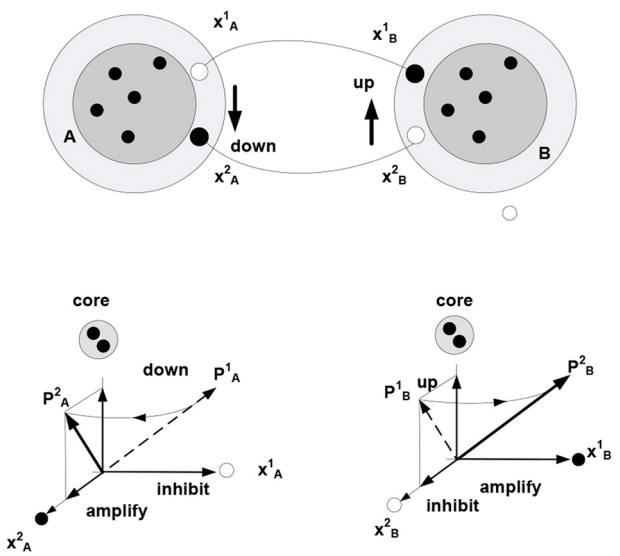


FIGURE 5 | The behavior of objects A and B (e.g., toy balance) is observed to be coordinated (moving A down is accompanied by B going up, etc.) Different movements correspond to (are represented by) different successions of firing in the fringes of packets A and B (e.g., neurons x^1_A and x^2_A represent different positions of A respective the vertical axis, successions $x^1_A \rightarrow x^2_A$ and $x^1_A \rightarrow x^2_A$ represent downward and upward movement, respectively). Bi-directional mapping between the fringe subsets in two packets establishes "co-instantaneous coordination" in the movement of packet vectors, $\mathbf{P_A} \between \mathbf{P_B}$ (symbol $\between$ denotes "co-instantaneous coordination" in the movement of packet vectors).

requires representations of the objects and their behavior (situated cognition) but does not reduce to simple recognition. That is, unlike objects and behaviors, coordination cannot be observed but has to be imputed, resulting in a compositional representation (iterative model), such that operations on one part of the composition produce particular changes in the other. For example, when thinking of pushing down one side of a catapult, one cannot help thinking that the other side will go up. The underlying mechanism is neither an image (although some visual predictions might be generated by the model) nor a linguistic expression, such as a rule (although some linguistic predictions might come to mind) but a forceful (energy consuming) mental activity directed at performing a particular work on a representation (vector rotation). **Figure 6** illustrates this notion.

In the absence of coordination, packets A and B are experienced as unrelated objects displaying mutually independent behavior patterns. Establishing coordination in the movement of packet vectors produces a generative model; that is, a coherent representative structure (model) and constrained operations on that structure (mental modeling), giving rise to the experience of a unified construct that combines objects in a meaningful relationship.

Figuratively, population vectors can be taken to represent the "consensus view" of the population, while vector rotation expresses changes in neuronal responses in the course of "settling on" a "consensus". According to the current proposal, understanding involves coordinated neuronal activities (Bressler and Kelso, 2001, 2016), in particular, coordinated rotation of population vectors comprising in a mental model, with the form of such coordination reflecting the form of mutual constraints (dependencies, relations) in the behavior of the entities represented by the populations. Consistent with that proposal, the experience of "grasp" accompanies the concluding stage in the modeling process that "settles" onto a consensus regarding relations among the participating entities. In short, settling onto the "consensus view" in a model corresponds to obtaining mutually coordinated vector rotations across the model representing a coherent account of the situation as it unfolds.

Second, exerting cognitive effort is hypothesized to be a correlate of consciousness (Yufik, 2013). Associative links and their spontaneous groupings (packets) are the product of learning; i.e., they condition the organism to emit recurring responses under recurring circumstances. Effortful composition of packets into mental models and model manipulations (e.g., coordinated rotation of packet vectors) serve to overcome the inertia of prior learning, when encountering and/or

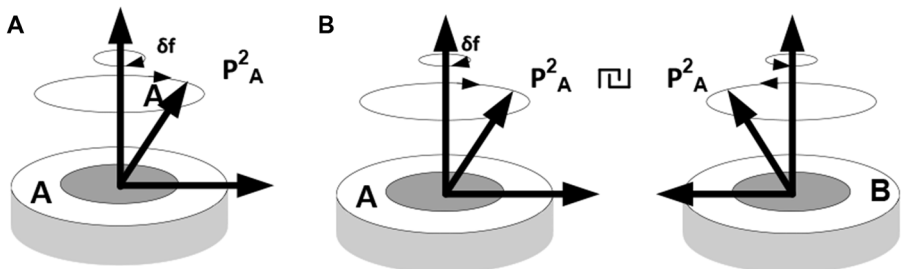


FIGURE 6 | Establishing behavioral coordination. (A) Rotation of packet vectors (selective inhibition-amplification of neuronal responses in the fringe subset) is analogous to rotating a shaft in a mechanism, e.g., a dc motor: it requires effort to produce movement (effort δf per unit angular displacement). **(B)** Establishing behavioral coordination creates a unified compositional structure such that vector rotations in one part of the composition produce (enforce) rotations in the other parts (e.g., visualizing one side of a catapult going down brings about, irresistibly, the image of the other side going up, or thinking of holding an object in the hand and releasing the hold brings about the image of a falling object, etc.).

anticipating unfamiliar conditions. Learning capabilities are common, to a varying degree, to all animal species, a superior adaptive efficiency in humans may be due to mechanisms allowing effortful suppression of the automatisms acquired in learning and/or adjusting their execution—depending on the circumstances at hand.

Third, coherent neuronal structures are thermodynamically beneficial; i.e., resisting decomposition and/or reorganization. For example, young children fail to understand that, when the target is moved away from the catapult, the ball's position on the plank needs to be shifted in the opposite direction. Failure is caused by the previously established basic coordination (reaching an object requires movement towards it, not away from it) precluding the requisite adjustments (children are incapable of a focused cognitive effort demanded by the adjustment).

Formally, coordination of packets defines an objective function over a vector space. In the nervous system, the function is implemented in a structure that is analogous (within limits) to Shannon's Differential Analyzer (DA; Shannon, 1941). The DA machine is composed of shafts connected by movement conveying devices such as gear boxes. When a shaft representing an independent variable is turned, all other shafts are constrained to turn accordingly. The implications of this analogy will be examined elsewhere, excepting the following observations.

(a) The objective function seeks maximization of energy efficiency, that is, vector (shaft) rotations are sought that maximize energy inflows at the expense of minimal rotation effort.

(b) A coherent model (tightly coordinated packets) collapses combinatorial complexity of the task and thus allows "intuitive" navigation of large combinatorial spaces, as in chess:

> *"Intuition is the ability to assess a situation, and without reasoning or logical analysis, immediately take the correct action. An intuitive decision can arise either as the result of long thought about the answer to the question, or without it"*
>
> *—(Beim, 2012, p. 10).*

The experience of "intuition" is produced by the ability to relate, via sufficiently tight coordinations, particular moves to the global objective (winning the game)—a move is "sensed" to improve or degrade the overall position (in the chess literature, this ability has been compared to a GPS in the player's mind showing whether moves take one towards or away from the goal (Palatnik and Khodarkovsky, 2014)). Such guiding intuition is not confined to chess but is a universal attribute of complex analysis and problem solving that is informed by coherent models.

> *"The mass of insufficiently connected experimental data was overwhelming. . .however, I soon learned to scent out that which was able to lead to fundamentals and to turn aside from everything else, from the multitude of things which clutter the mind and divert it from the essential"*
>
> *—(Einstein, 1949, p. 17).*

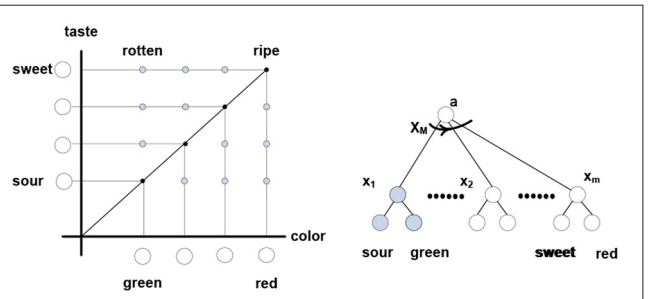


FIGURE 7 | Thinking "apple is ripening" involves rotating "apple" packet vector from the (sour-green) to (sweet-red) terminal positions via some intermediate angular positions. Neuron x_1 responds to co-firing of "sour" and "green" neurons, x_M responds to co-firing of "red" and "sweet," etc. Neuron x_M responds to the firing succession $x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_m$ formed of diagonal elements in the color-taste matrix. Thinking "apple is rotting" engages different elements residing in different rotation trajectories. In a simulation, firings can be associated with different values, contracting the matrix attributes and assigning value to the "ripening" trajectory; namely, the sum of values in the $x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_m$ firing succession (spur of the matrix).

Navigating and connecting massive sensory data requires a model that guides subsequent probes and enables determination (however approximate) of whether the data lies within the range of variation afforded by the model, or falls outside the range and invalidates the model. As per **Figure 1**, probabilistic prediction and inference are at the foundation of the modeling process.

(c) Coordinations in systems of nested packets can be expressed as optimization operations in vector spaces (Dorny, 1975) and as functions over tensors or multi-vectors (Clifford vectors) of geometric algebra (Hestenes and Sobczyk, 1999; Doran and Lasenby, 2003). Complexity reduction in such systems can involve rank reduction and tensor contractions.

(d) In the nervous system, complexity reduction can involve neurons responding to trajectories of packet vectors; that is, particular successions of their angular positions. In other words, such neurons respond to particular thinking patterns, as illustrated in **Figure 7**.

Summary

This section outlined a parsimonious theory of understanding where foundational ideas in systems neuroscience (Hebbian assembly) and probabilistic learning theory (variational free energy minimization) converge on the notion of a neuronal packet—a neuronal assembly "wrapped" in Markov blanket. Cognitive processes are defined as operations on neuronal packets providing a unifying formalism to express the function of understanding as well as phylogenetic and ontogenetic development of intelligence culminating in that function: allocating neurons—allocating cohesive neuronal groupings—adjusting groupings—apprehending coordinated adjustments—combining and coordinating groups (*mental modeling*). Psychologically, the process encompasses the progression from sensing, to perceiving, to understanding.

Mathematically, this formalism suggests operations on vector spaces (via a geometric calculus).

The ensuing theory grounds cognitive development in thermodynamics, suggesting a straightforward relationship between self-organization and evolution (packets are thermodynamically sculpted and operations evolve). Evolution engages an interplay between the internal (packet manipulations demand energy) and external processes (where the environment supplies energy), propelled by the need to improve energy efficiency. Organism-environment coupling is probabilistic, allowing a dual account: doing work to extract energy manifests as sampling and information gathering. The energy-saving tendency to maintain cohesive and stable packets is motivated by the minimization of surface tension in the packet boundary surface. Surface tension is a fundamental parameter expressing the thermodynamically favored direction of internal processes in any system. In a neuronal system, favored processes include increasing cohesion (reducing interface area in individual packets) and merging (reducing the total interface across the packet set). Minimization of surface tension entails minimization of a thermodynamic free energy in packet surfaces and equates to avoiding surprise (minimizing variational free energy in probabilistic inference). On that theory, packets are the substrate of inference.

One might ask whether the solutions that minimize variational free energy are stable and—from a technical perspective—are these functionals convex. By virtue of the dynamic and itinerant nature of biological systems (especially in the context of a circular causality implicit in self organization), it is highly unlikely that the energy functionals describing behavior are convex. Heuristically, this means that there will be many minima—or solutions. The implicit multi-stability provides a nice mathematical image of speciation—and indeed variants within any phenotype. In other words, there is no unique free energy minimum, in the same sense that there is no unique phenotype; each system adapts to its own econiche—finding its own solution.

The notion that quasi-stable neuronal packets—and their manipulation—underlie perception resonates with theories that associate perceptual units with quasi-stable solutions in mean field models; for example, neural field models that account for the neurogeometry of the cortex and the impact of visual input (e.g., Sarti and Citti, 2015). According to Sarti and Citti (2015), in the absence of visual input, quasi-stable solutions correspond to hallucinatory patterns. Notwithstanding the possibility of quasi-stable neuronal clusters engendering hallucinatory experiences, our theory predicates mental modeling on the formation of quasi-stable packets that maintain their integrity throughout episodes of absent and/or varying input. Such quasi-stable units allow the experience of continuing, self-identical objects that arise from (i.e., are superposed upon) discontinuous and varying sensory streams. More generally, the neuronal packet model is compatible with the mean field models that furnish a dynamics of neuronal systems from metastability and symmetry breaking—and associating system behavior under stimulation with quasi-stable states and active transient responses (Wilson and Cowan, 1972, 1973; Bressloff et al., 2002). Examining conceptual commonalities and reconciling differences between these models may help overcome their inherent limitations (e.g., Destexhe and Sejnowski, 2009) and offer synthetic perspectives.

ANALYSIS

This section compares the proposal in the preceding section to other theories described in "Theories of Understanding". Since our proposal rests on the notion of neuronal packets, we discuss how the idea conforms to the principles of neuroscience and present some recent data concerning the properties of neuronal structures consistent with those attributed to neuronal packets. Finally, we consider an approaches to understanding motivated by complementary ideas based on "intuitive physics engines".

Comparing Theories

The theories in "Theories of Understanding" complement our formulation. Moreover, they appear to reflect different facets of understanding, as conceptualized above. The "foundational theory of understanding" (Newton, 1996), which grounds understanding in self-directed (volitional, attentive) activities reconciling sensory inflows with memory structures and current goals, is consistent with our theory that associates understanding with goal satisfaction via self-directed allocation of neuronal resources. The idea that evolution has gradually shifted response targets away from the sensory periphery, producing internal efferent-afferent loops that can be decoupled from the motor output (Humphrey, 2000, 2006) is formally expressed in the model of self-adaptive resource allocation.

The key insights in the theory of understanding by Chart (2000) appear to be formally expressed and substantiated by our treatment. Chart (2000) derives understanding from simulations involving effortful (work- consuming) operations on mental models built of "mutors":

> *"Mutors are both the building blocks and the motors of mental models. . .mutors are active: they actually do the work on the input, and produce the output. They are not rules by which the input can be transformed into the output; rather, they are machines which effect the transformation"*
>
> *—(Chart, 2000, p. 47).*

These intuitive notions correlate closely with the idea of effortful vector rotation and other ideas (see **Figure 6**; note similarities between Chart's theory and Shannon's DA. The theories also differ in that one is centered on the work requirement and the other is oblivious to it).

The *doing work* requirement in Kauffman (2000), predicating intelligence on the ability to invest energy in performing thermodynamic work cycles directed, in part, on erecting constraints for the subsequent energy releases, appears to be fully upheld in our theory (e.g., boundary energy barriers constrain composition and movement of packet vectors thus constraining energy release in vector rotation which,

in turn, constrains condition at the boundary). The idea of associating intelligence with "approximate constrained optimization" in the service of need satisfaction (Glasser, 1984; Werbos, 1996, 1998) is inherent in the notion of probabilistic resource optimization. Our proposal ascertains a reciprocal and complementary relationship between probabilistic resource optimization via resource grouping, statistical explanation (Salmon, 1970) and probabilistic inference, as discussed above.

Simplification (Kitcher, 1981) and compression—postulated to be the definitive characteristic of explanation ("comprehension is compression", Chaitin, 2006)—are the product of enfolding, collapsing multiple resources into a single unit. In essence, alternating enfolding–unfolding serve to break large combinatorial problems into sets of much smaller ones, yielding profound complexity reduction. Furthermore, simplification is isomorphic with complexity minimization inherent in minimizing variational free energy and, by implication, thermodynamic complexity costs.

Finally, our theory gives operational expression to some of the central claims in the psychological theory of understanding. Developmental psychology predicates development of a capacity to understand, from infancy to maturity, on the growing ability to conduct "co-instantaneous mental coordinations" and thus apprehend relations abstracted from the current sensory input:

> *"...to coordinate data yielded by his own actions the child must appeal to unobservable, deductive relations which transcend his actions"*
>
> *—(Piaget, 1978, p. 12).*

Our proposal defines processes underlying "mental coordinations" and makes them responsible for all levels of understanding, from handling toys to formulating abstract theories. From the resource optimization standpoint, coordinating packets in nested packet groupings provides a scalable mechanism for compression and complexity reduction. From the psychological standpoint, coordination combines disparate and unrelated entities into "situations" imbued with meaning. That is, meaning is imputed by relations.

Neuronal Packets

A "neuronal packet" is a system-theoretic idea derived from conceptualizing the nervous system as a probabilistic resource optimization system with self-adaptive capabilities (Yufik, 1998b). The starting point was attempting to formulate Hebbian assemblies (Hebb, 1949, 1980) as material entities: what makes assemblies distinct, how does the system "know" where one assembly ends and another begins? Once formed, why wouldn't assemblies succumb to entropic erosion and dissolve momentarily? Drawing on Haken (1983, 1993), packets were hypothesized to be formed by phase transitions in associative networks and sculpted by an interplay between thermodynamic forces (reduction of thermal free energy in the inter-phase surface) favoring coalescence and forces of lateral inhibition resisting coalescence. This interplay dynamically optimizes responses: through lateral inhibition, packets capture regularities in the sensory stream.

Arguably, the existence of boundary mechanisms was implicit in the notion of assembly, the consequences (structure variation, induction of meaning, etc.) were fully anticipated by Hebb:

> *"...we have come to a classical problem...the meaning of "meaning"... a concept is not unitary. Its contents may vary from one time to another, except for a central core whose activity may dominate in arousing the system as a whole. To this dominant core, in man, a verbal tag can be attached; but the tag is not essential. The concept can function without it, and when there is a tag it may be only a part of the "fringe". The conceptual activity that can be aroused with a limited stimulation must have its organizing core, but it may also have a fringe content, or meaning, that varies with the circumstances of arousal"*
>
> *—(Hebb, 1949, p. 133; see **Figure 5**).*

The notion of *intrinsic organization* of cortical activity "that is so called because it is opposed to the organization imposed by sensory events" (p. 121), the necessity for assemblies to be sustained over time (p. 121), the possibility of forming "latent" associations between stimuli that have never co-occurred in the past (p. 132), the "coalescence" of assemblies (p. 132), and numerous other ideas in Hebb (1949) place the packet concept within Hebb's framework.

The concept of a "neuronal packet" is consistent with other system-level theories of cognition. The theory of neuronal group selection (TNGS; Edelman, 1992, 1993; Edelman and Tononi, 2000) associates cognitive functions with the formation of "neuronal groups" and establishment of "re-entrant mappings" between groups (Edelman and Gally, 2013; see **Figure 6**). In Gestalt psychology, packets manifest in the notion of "gestalt bubbles" (Lehar, 2003a,b), or "segregated wholes" that enable meaning ("... meaning follows the lines drawn by natural organization; it enters into segregated wholes" (Köhler, 1947, p. 82)). Significantly, "segregated wholes" were subject to forceful manipulation (the idea organizing "force fields" in the brain that "extend from the processes corresponding to the self to those corresponding to the object" (Köhler, 1947, p. 177; 1948)). The idea of "forceful" interactions was later associated with the activity of consciousness: in the brain, consciousness is "put to work" exerting a controlling influence on the stimuli-triggered and volitional (self-generated) motor responses (Sperry, 1969). Interestingly, the notion of force fields as underlying perception has been revisited in the context of gauge theories for the brain using variational free energy as the underlying Lagrangian (Sengupta et al., 2016). Formally, this is closely related to the autopoietic destruction of (free energy) gradients in synergetic formulations of brain function (Tschacher and Haken, 2007).

A "neuronal packet" is a speculative concept—the implicit packets (or assemblies) are not amenable to direct observation but have to be inferred in terms of their functional connectivity and underlying conditional independence. However, recent empirical data appears to uphold the concept. Packets are thermodynamically plausible because their [re]use minimizes energy expenditure. That is, the possibility of re-use is inherent

in the packet idea. Reusable neuronal groups ("bubbles") were discovered in the hippocampus of awake, free-moving animals (mice; Lin et al., 2005, 2006; Tsien, 2007). Empirical verification was enabled by recent technical advances allowing simultaneous recording of activity of 260 neurons: recordings were made in the CA1 region in animals subjected to different perturbations (shaking, elevator drops, air puffs) and in the resting state. Multiple discriminant analysis (MDA) was carried out over half-second sliding windows in recordings accumulated over several hours, revealing the formation of distinct "bubbles", Or groupings of neuronal activity that were well separated in the functional 3-D space (contracted by MDA from the 520-D space). The ensuing bubbles represented "integrated information about perceptual, emotional and factual aspects of the events" (Tsien, 2007, p. 55). After the "bubbles" were formed, subsequent responses could be characterized in different compositions, e.g., an "earthquake" type situation begins in the "resting bubble", transits to the "earthquake bubble" and returns to the "resting bubble"—thus following a distinct trajectory in the functional space.

The possibility of resource tuning (changing resource characteristics depending on those of the task) is inherent in the concept of resource allocation (see **Figure 2**). Task-dependent changes in the receptive fields of individual neurons (see rotation of neuronal response vectors) have been demonstrated in a broad range of tasks and conditions including different stimulation modalities (auditory, visual) and durations of exposure (Fritz et al., 2003, 2007; Kohn and Movshon, 2004; Elhilali et al., 2007). For example, recordings of individual neurons in A1 in ferrets performing tone-discrimination tasks revealed distinct and predictable changes in spectro-temporal receptive fields ("task-specific signatures"; Fritz et al., 2007). In the earlier experiments, neurons in the prestriate area V4—in monkeys attending to visual stimuli—demonstrated robust attentional gating of their receptive fields: a neuron having two stimuli within its receptive field selectively suppressed its responses to one or the other stimulus depending on the task (Moran and Desimone, 1985).

Task-dependent changes in the responses of neuronal populations (rotation of population, or packet vectors) were demonstrated by Georgopoulos and his group in studies of neuronal correlates of target reaching in monkeys. Neurons in M1 are broadly tuned to the direction of movement, with each neuron exhibiting a preferred direction—defining the orientation and the magnitude of the neuronal response vector. It was shown that population response vectors—obtained as the vector sum of weighted neuronal response vectors over the population of responding motor neurons—track the direction of the hand movement (Georgopoulos et al., 1988, 1993). In a similar fashion, weighted sums of neuronal responses over populations of sensory neurons were shown to align closely with the overt characteristics of sensory processing (Jazayeri and Movshon, 2006). Furthermore, it was shown recently that population responses adapt to task variations, involving subsets of neurons particularly relevant to the current task ("high-precision neurons"; Purushotaman and Bradley, 2005).

The overall approach of conceptualizing cognitive processes as optimization of neuronal resources has received experimental support and theoretical emphasis in the recent studies of visual perception (Gepshtein et al., 2013) and the analysis of candidate mechanisms in the brain capable of anticipation and long-term planning ("prospective optimization"; Sejnowski et al., 2014). Perhaps, the most compelling argument in favor of the present theory can be garnered from the work reported by Ito (1993, 2008), Salman (2002), Baillieux et al. (2008); Ellis and Newton (2010), Murdoch (2010), and Rosenbloom et al. (2012) suggesting a possibility that mental activities are controlled by internal models in the cerebellum (Ito, 2008), with movement and thought engaging identical control mechanisms (Ito, 1993). On the theory that understanding boils down to packet coordination, pieces of the understanding puzzle seem to be falling in place. That is, the critical function of packet coordination hypothesized in **Figure 6** may be evident in the cerebellum.

Key components of "understanding" include value-assignment (reward likelihood attribution), packet mobilization and effortful, context-sensitive variation, packet coordination, output suppression and response selection. These components map, under a gross simplification, onto a functional neuroanatomy comprising prefrontal cortex (PC), subcortical structures; including the basal ganglia, thalamus, and cerebellum, and the limbic system (Rosenbloom et al., 2012). The orbitofrontal, anterior cingulate and dorsolateral regions in PC interact with each other and the limbic system and subcortical structures. In particular, the orbitofrontal cortex and limbic system participate in reward-attribution, while the dorsolateral and anterior cingulate regions "facilitate intellectually effortful decisions" (Rosenbloom et al., 2012, p. 256). Frontal areas are involved in response suppression, while the cerebellum mediates a key mechanism of understanding: packet coordination. Via the cerebellum, precise timing—necessary for sensorimotor coordination (Salman, 2002)—becomes an integral part of situational understanding that is manifest in the ability to not only compose, in the mind, coordinated activities fine-tuned to the current situation but also to identify proper moments for releasing and terminating them.

Energy barriers play a crucial role in coordinated timing. On the present theory, folding into packets creates a continuous energy landscape in associative networks (peaks and valleys form energy barriers that separate pools of neurons endowing them with a conditional independence that create Markov blankets). The implicit barriers may be regulated by the limbic system (regulation of the "cortical tone" (Luria, 1973)), via the classical ascending neuromodulatory systems. For example, down regulation (stress, fear, low motivation) raises energy barriers, while up regulation (joy, arousal, high motivation) lowers them. This sort of regulation or (neuromodulatory) arousal, directly affects cognitive performance as follows. Optimal performance requires optimal "cortical tone" (underlying the

Yerkes—Dodson law of optimal performance (Eysenck and Keane, 1995)). Excessive down regulation blocks attentive access to packet internals (as in suddenly forgetting a familiar name) or arrests attention within a packet (vacillation, inability to escape from recurring thoughts). By contrast, excessive up regulation precludes sustained focus and predisposes to spurious associations. In pathological extremes, the landscape is either flattened, turning sensory inflow into undifferentiated flux (e.g., Alzheimer's disease), or loses integrity and decomposes into pockets of narrowly constrained skills (e.g., autism). When a packet dissolves, the contents are not forgotten but irrevocably lost. We shall re-visit this point briefly in the discussion.

The mechanism of mental modeling is ubiquitous across species. For example, sensing a prey initiates hunting behavior in a snake. If the prey suddenly disappears, the snake starts searching for it but only in the vicinity of the location where the prey was last sensed. By contrast, a dog chasing a prey that goes out of sight (e.g., a rabbit disappearing behind bushes) can initiate an interception maneuver; i.e., running towards a location where the prey is likely to re-emerge (Sjölander, 1995). Figuratively, the snake's hunting model contains one packet whose boundaries are statistically determined and genetically fixed (the radius and duration of search are consistent with the behavior of animals typically consumed by snakes—thus yielding adaptive fitness). Dogs and other higher animals possess repertoires of specialized packets amenable to situation-sensitive variations (a prey's velocity, distances, etc.). Chimpanzees can combine some genetically available activities (reaching with a stick, piling up objects and climbing to obtain a reward reflect their genetic repertoire) but coordinating such activities appears to be approaching the limits afforded by their nervous system. Human modeling capabilities in infancy are rudimentary (e.g., at 6 months, infants search for a toy after it was covered but, if the toy is removed and placed (in full view) under a different cover, they keep searching for it where it was first perceived (Bower, 1974)). Human capabilities develop rapidly, from coordinating a few variables in handling toys (e.g., ball placement in a toy catapult, given the distance to the target) to coordinating deeply nested variable structures in the creation of abstract theories. We propose that the formalism of neuronal packets and packet coordination characterizes essential features of the underlying mechanism at all stages of cognitive development.

So far, understanding and mental modeling have been discussed in the context of problem solving and prediction (Toulmin, 1961), without addressing the impact of emotion on these cognitive activities. The thermodynamic framework suggests a natural expression of that impact (Yufik, 1998a), by identifying emotional control with thermoregulation and temperature with the level of arousal (it is interesting to note that Aristotle attributed to the brain the function of thermoregulation, Gross (1995)). In particular, the neuronal packet model represents boundary free energy (the height of packet energy barrier) U as a function of temperature approximated as $U(T) = \sigma - T d\sigma/dT$ where σ is a stability coefficient computed as the ratio of the summary strength of the internal vs. external associative links in the packet ($\sigma > 1$: such that the packet disintegrates when σ approaches unity, bringing U(T) in to the vicinity of kT, where k is the Boltzmann constant). Increasing T lowers the barriers while decreasing T (stress, fear, anxiety) results in their elevation. Low barriers enable easy (low energy cost) transitions between packets (expansive, compositional thinking) while elevated barriers hamper the transitions.

Temperature variations can be local (focused thinking) or global (diffuse). Diffuse temperature increases lower energy barriers and "shake up" the system, entailing re-distribution of neurons among packets, followed by focused (selective) manipulations in the resulting structures (the term "cognition" derives from the Latin "cogito" meaning "to shake together", "intelligence" derives from the Latin "intelligo" meaning "to select among", Koestler, 1964, p. 120). As noted earlier, the overall temperature dependency of the packet system approximates the Yerkes-Dodson law of performance (optimal levels of arousal yield optimal cognitive performance). More generally, temperature regulation engages global self-regulatory loops allowing the organism to reconcile conditions in the outside with those inside and thus maintain a form of homeostasis. Arguably, thermal regulation transcends the hierarchy of functional levels in the organism—from changes in the cell membrane permeability and neurotransmitter flow (e.g., changes in the release, reuptake and repriming of synaptic vesicles; the micro level) to changes in packet composition (the mesa level), and further to emotional shifts entailing changes in overt macro responses (advance or retreat; the macro level). These views are generally consistent with those formulated in Damasio and Carvalho (2013) and Damasio and Damasio (2016).

Alternative Theories

A recent theory of cognitive mechanisms involved in the understanding of physical scenes (e.g., a determining whether a stack of blocks is going to hold or to topple) derives understanding from the operation of an "intuitive physics engine" (IPE) combining simulation of interaction between objects with probabilistic inference, by treating simulation runs as statistical samples (Battaglia et al., 2013). Simulating interactions is the crux of the matter, how is this accomplished in IPE? To demonstrate human-like performance, IPE employs open dynamics engine (ODE[1]) offering a library of routines (equations, methods and algorithms) to simulate rigid body dynamics. If IPE succeeds in emulating humans, what would this tell about the mechanisms of scene understanding in the brain? Stated differently, what makes IPE brain-like?

Three constraints in employing the ODE library are claimed to qualify IPE as a theory of scene understanding: only elementary rules of physics are selected in ODE, Monte Carlo procedures inject probabilities into simulation runs, and inference calculations are carried out to a crude approximation. Consider applying these constraints in a toy catapult problem (e.g., balancing two objects on a plank): $w_1L_1 = w_2L_2$ is the most elementary rule, simulation varies the values of L_1 and L_2, probability distributions are associated with variation ranges

[1]http://www.ode.org

L_1 and L_2, and all calculations discard small terms and round the results. If that is what underlies understanding, the question remains: how is the rule $w_1L_1 = w_2L_2$ obtained, represented and exercised? The probabilistic inference and approximation components in IPE only postpone the inescapable conclusion that understanding boils down, literally, to mental arithmetic. With that, any human-like behavior can be readily imitated and explained (e.g., a child failing to understand that ball needs to be moved away from the center of the catapult when the distance to the target increases, has her Monte Carlo flip the sign, i.e., computes $L_2 - \Delta L_2$, instead of $L_2 + \Delta L_2$,).

In short, results in Battaglia et al. (2013) appear to demonstrate that combining methods of analytical mechanics with probabilistic inference allows rough and quick assessment of interaction dynamics in simple mechanical systems. Whether these results have anything to do with human understanding or intuition is open for debate. In lieu of entering the debate, this article has outlined a complementary approach to the issue.

DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH

Brain is complex, dynamic self-organizing system (Bressler, 1994; Singer, 2009). Self-organization requires a flow of thermodynamic energy through a system acting as a conduit between an energy source and energy sink. At equilibrium, energy transfer by thermodynamic forces is accompanied by generation of entropy. Deviations from equilibrium is accompanied by a decrease in the rate of entropy production, eventually producing conditions where stable structures emerge in the form of spatial (e.g., Bérnard cells), temporal (e.g., Belousov-Zhabotinsky reaction) or spatiotemporal structures (Glansdorff and Prigogine, 1971; Prigogine and Stengers, 1984, 1997; Prigogine, 1994; Bak, 1996; Jensen, 1998). The brain belongs in the continuum of self-organizing systems (Bressler, 1994; Kelso, 1995; Camazine et al., 2001). Sustained self-organization in far-from-equilibrium systems is contingent on the existence of internal mechanisms capable of removing entropy from the volume occupied by the system and depositing it outside the volume (Morowitz, 1978, 1979; England, 2013; Prokopenko et al., 2014). The development of intelligence implies a reduction of entropy within the brain's volume—to levels allowing emergence of stable structures that can both amplify energy inflows and direct the investment of a growing portion of that inflow towards creating more entropy reducing structure. In a sense, a self-organizing (self-adaptive) system keeps folding upon itself, producing increasing degrees of internal order. Human intelligence requires a degree of order, engendering stable but flexible structures (neuronal packets) and reproducible internal processes (thinking). This combination gives rise to the experience of interacting with an orderly environment amenable to understanding, as follows.

The requirements of facilitating energy import from the outside—and structure generation of the inside—converge when structures are flexible (but stable) and reflect regularities in the external conditions. With that, reciprocity is established between internal "objects" and environment. A self-organizing system becomes aware of the "objects"— including itself as an object; i.e., when objects become amenable to internal manipulation, establishing relations between objects expressing higher-order regularities in the environment. The availability of such manipulations rests on having reduced the rate of entropy production, down to levels that allowing reversibility of thinking. That is, no thinking is possible if one cannot: (1) dwell on object A; (2) switch from object A to object B and return to B; and (3) keep all the objects intact in the course of 1 and 2. Reversibility endows quasi-stable objects with self-identity, thus rendering thought possible and making the environment (the universe of persevering, self-identical objects) understandable. The relationship between reversibility and understanding is manifest in the foundational principles of psychology, logic and mathematics.

In psychology, this relationship was first articulated in the last century by Piaget, in the form of a reversibility principle and the notion that cognitive structures—and operations on those structures—in mature adults acquire the property of algebraic groups. In logic, the relation underlies The Law of Identity formulated by Aristotle as the key axiom from which reasoning derives. The Law of Identity ($\mathbf{A} \equiv \mathbf{A}$) (and the corollary of non-contradiction and excluded middle) asserts preservation of self-identity in things despite changes. Things neither appear nor disappear spuriously, they remain self-identical over time and do not change without a cause. Finally, in mathematics, the relation is expressed in the foundational principle of set induction and cardinality attribution formulated by Cantor (1915/1955):

> *"We will call by the name "power" or "cardinal number" of M the general concept which, by means of our active faculty of thought, arises from the aggregate M when we make abstraction from m and the order in which they are given"*
>
> *—(Cantor, 1915/1955, see Tiles, 1989, p. 99).*

In short, set is induced on a group by the "active faculty of thought" that treats the group, reversibly and alternatively, either as a manifold or as a unit abstracted from the manifold.

The criteria of causality are hard to explicate (e.g., leading to the recent notion of "graded causation" (Fitelson and Hitchcock, 2011; Halpern and Hitchcock, 2015)) but, nuances aside, causality concerns a relation between some A and B: changes in A are (or are not) the cause of changes in (B). By contrast, the set operation dwells on A. The operation underlies mathematics (and abstractive thinking in general) and enables compositionality; i.e., combining A and B into a new unit A, B $\rightarrow$ (AB) amenable to reversible decomposition (AB) $\rightarrow$ A, B, and so on, indefinitely.

According to the theory of neuronal packets, the above principles are rooted in (and express) packet unfolding/enfolding and inter-packet coordination (causality). Unfolding gives access to the packet's sensory contents, while enfolding abstracts from them. Alternating between enfolding and unfolding can be visualized as moving up and down a

cone; with the sensory data at the base. On the way up, the sensory component is reduced—and is completely removed (abstracted away) at the apex. Symbolic labels that could be attached at the apex (e.g., labels "apple" and "Apple computer") have no sensory overlaps with the corresponding objects. The essence of thinking is effortful packet manipulation, with the process alternating sporadically between imagining and reasoning (syntactic manipulation of labels). Crucially, the process is different from—and does not reduce to—*pattern recognition*. This contention will be discussed elsewhere.

The development of order in self-organizing systems implies the emergence of Markov blankets; i.e., encountering a confluence of conditions that allows the system to self-segregate, or fold into components that remain coupled to the system but acquire conditional independence. In living organisms, mechanisms start to form that regulate the "permeability" of the blankets, i.e., facilitating inflow of energy and matter necessary for sustaining independence and integrity at the level consistent with survival. One might imagine that further development creates higher-order regulatory mechanisms comprised of nested components "wrapped" in Markov blankets.

When analyzing the thermodynamic underpinnings of life, Schrodinger introduced the notion of negentropy extraction: "the device by which an organism maintains itself at a fairly high level of orderliness (low level of entropy) really consists in continually sucking orderliness from its environment" (Schrodinger, 2006, p. 73). Negentropy extraction involves active sampling and harvesting of information from the environment. The induction of Markov blankets and increase of order via partitioning of associative networks into nearly homogeneous subsets (neuronal packets) equates to internal generation of information (Salmon, 1970). Thermodynamic free energy is therefore diverted from dissipating organismal structure and is stored in ATP molecules at the packet surface, to be released in the work of composing and re-shaping packets for further free energy minimizing inference. Our theory defines the increase of order via constructing models as negentropy generation (orderliness is manufactured inside the system).

Minimization of boundary free energy can drive self-organization and self-assembly in microstructures (Syms et al., 2003) and influence first-order phase transitions, inducing critical phenomena (surface-induced order and disorder (Lipowsky, 1984)). The coexistences of phases in a first-order transition is described by Landau-Lifsitz potential with several minima, with spontaneous symmetry breaking (e.g., packet formation) on obtaining one of the minima (producing order and the disordered phase characterized by a vanishing order parameter (Lipowsky, 1984)). In general, identifying the thermodynamic variable with the surface area of a packet offers a hypothetical Lagrangian or Lyapunov function that poses some interesting analytic and practical questions. From a technical point of view, it motivates a formal analysis of the relationship between the surface area (thermodynamic free energy) and variational free energy. From a practical point of view, the surface area can be treated as an order parameter, which is either minimized or conserved—in accord with Hamilton's principle of stationary or least action.

Transition from negentropy extraction to negentropy generation encompasses a continuum of intelligent processes, from rudimentary (plant intelligence, e.g., Trevawas, 2002; Marder, 2013) to the most elaborate (human intelligence). In the latter, a spectrum of mechanisms can be involved operating in conjunction with neuronal mechanisms; e.g., from limbic neuromodulation to glial cell function (Chung et al., 2015); from synaptic processes to microtubules (Penrose, 1997). All such mechanisms exploit thermodynamic forces to optimize energy extraction and utilization in the interest of survival (e.g., sunflowers tracking the sun). Accordingly, the formalism of self-adaptive resource optimization applies across the continuum of biological intelligence. Emulating biological intelligence in artifacts would require a range of designs, including analog (super-Turing network (Siegelmann, 1999; Cabessa and Siegelmann, 2011)), digital and digital-analog hybrids.

Our proposal associates self-organization in the physical substrate with minimization of free energy, and asserts isomorphism between variational and thermodynamic expressions of free energy. Under both expressions, the process involves self-partitioning in the substrate yielding internally cohesive and externally weakly coupled (statistically quasi-independent) components. As astutely noted by a reviewer, the concept of energy minimization resonates with some classical techniques in pattern analysis (e.g., energy minimization in Hopfield networks) and image processing. In general, minimization of an "energy functional" is used to obtain image segmentation into "meaningful" regions ("objects") having uniform feature intensity and separated by non-uniform, low-intensity patches. Minimization can be sought of some local energy-like expression (Lucas and Kanade, 1981) or a global energy functional (Horn and Schunck, 1981; Bruhn et al., 2005). In the former case, the "energy functional" takes the form $E(u, B) \rightarrow \min$ where u is the smoothed image and B is a curve segmenting the image (i.e., the union of "object" boundaries; Mumford and Shah, 1989; Shah, 1992).

Mathematical ideas motivating boundary detection by minimizing energy functionals (Mumford and Shah, 1985) appear to be converging on our proposal postulating free energy minimization in the interface or boundary separating neuronal packets from the surrounding structure, thus providing further support to the hypothesis that packets underlie perception of "objects." Note that our overall proposal deals with models of input (rather than percepts) and thus calls for expanding the conceptual basis and the corresponding mathematical apparatus, as compared to those employed in image processing. In particular, the free energy minimization requirement is associated not only with segmenting images into packets ("objects") but, crucially, with the subsequent operations on packets, such as coordinated rotation of packet vectors. In other words, the energy functional needs to be extended to include minimization over two variables: the boundary energy and the action. We believe that examining relations between energy-like function minimization in image processing and

variational and thermodynamic energy minimization in mental modeling is likely to yield informative and practically useful results, presenting a challenge for further research.

It is interesting to note that the vector manipulation formalism adopted in the present theory overlaps, to a degree, with the theory of morphogenesis in Thom (1975). In particular, the theory expresses morphogenesis (change of form) in a system M in terms of a vector field X on M determining the system's macroscopic dynamics. However, the overlap is limited since the intent was to "construct an abstract, purely geometrical theory, independent of the substrate of forms and the nature of the forces that create them" (Thom, 1975, p. 8). Similar attempts can be found in other system-theoretic studies of complex structures (e.g., Casti, 1979). Most system theories, including Thom (1975), focus on the general conditions of stability and resilience; i.e., the system's ability to absorb external disturbances without dramatic consequences for its steady-state and transient behavior. By contrast with system-theoretic proposals, the present proposal resonates with the objective reinstating the primacy of action and bodily grounded experiences in the theory of intelligence (Nunez and Freeman, 2014) and is interested in the physical properties of the substrate and the forces, seeking to relate them to resilience and adaptive changes. Nonetheless, system theories offer a rich mathematical apparatus and key insights (e.g., concerning the role of topological factors in biological morphogenesis (Thom, 1975)), that may contribute to a comprehensive theory of cognition.

Summary

Life emerges in networks of interacting material entities under a confluence of conditions that allow regions in the network to fold into bounded units statistically independent from the environment. Sustaining life requires regulating the flow of energy and matter through the boundary. The dual requirement of maintaining independence from the environment, while extracting sustenance from it, is resolved in progressively improving regulatory mechanisms ascending from the boundary to the internals. The progress is enabled by folding in neuronal networks and culminates in mental modeling involving manipulation of folded units (packets).

A detailed examination of the above hypothesis suggests a metaphor of brain function that comprises Bayesian and Aristotelian components, as follows. The interaction between an organism and its environment is probabilistic (no action is guaranteed to yield the expected outcome), necessitating Bayesian inference to predict and prepare for counterfactual outcomes before their onset; i.e., the cybernetic or Bayesian brain (Conant and Ashby, 1970; Knill and Pouget, 2004; Seth, 2014). Self-organization creates structures and operations in the system allowing logical inference; i.e., the Aristotelian brain. The Aristotelian brain builds on the foundation of the Bayesian brain in the course of self-adaptive resource optimization. The need to invest work in operating on structures equilibrates the Aristotelian-Bayesian system in the brain: self-partitioning into packets establishes both reference sets for Bayesian inference and a trade-off between the amount of cognitive work the system can invest and the amount of surprise it can tolerate.

The self-adaptive resource optimization framework (Yufik, 1998b, 2002; Yufik and Malhotra, 1999; Yufik and Sheridan, 2002) offers a simple account of cognitive processes, highlighting the crucial role of Markov blanket induction in neuronal systems, as a pivotal optimization mechanism.

From the perspective of Bayesian inference, induction equates to dynamic partitioning of large inference problems into a hierarchical succession of simpler problems, minimizing complexity (through dimension reduction) with the least loss of accuracy. Anticipatory inference (e.g., counterfactual prediction) is integral to optimization. This formalism is consistent with the functional organization of memory, distinguishing long-term (model parameters) and short-term (postdictive) components: in this (Bayesian) setting structure learning and inference can be expressed as optimization on vector constructs, such as Clifford vectors or tensors (e.g., Dorny, 1975; Smolensky, 1990; Doran and Lasenby, 2003).

From the perspective of physics, abductive reasoning equates to placing associative networks into regulated variational free energy landscapes where cohesive subnetworks ("bubbles") reside in valleys separated by energy barriers. This (variational and thermodynamic free) energy landscape defines expenditures (energy consumption and dissipation) in terms of the computational complexity—accuracy trade-offs and motivates optimization (Sengupta et al., 2013). From the perspective of psychology, induction underlies the unparalleled efficacy of human reasoning, by enabling transition from sensation to perception and from perception to understanding.

From the perspective of systems neuroscience, the function of understanding appears to be mediated by the Aristotelian-Bayesian brain via collaborative engagement of the thalamo-cortical system (associative network), the limbic systems (emotive thermoregulation) and the cerebellum (coordination). The theoretical perspective offered in this article is based on a fundamental, cornerstone of systems neuroscience (Hebbian assembly), by attributing biophysical properties to the assemblies that, arguably, are implicit in—and have been anticipated by—the original concept.

Finally, from the perspective of technology, implementation of the optimization and induction mechanisms speaks to a transition from machine learning to machine understanding. Advances in machine intelligence over the last half century have been associated primarily with perfecting techniques for computing weight distributions in fixed topology (perceptron-type) networks yielding a mapping between the input and output vectors (learning, pattern recognition). The store of algebraic ideas that have been employed in the task is rich, going back to Tichonov's regularization and iterative error reduction methods by Gauss, but finite and appearing, despite the recent strides (e.g., deep learning), to be nearing exhaustion. Simulation of understanding involves networks of varying topology and operations on dynamic vector structures, with the weights intact. Implementing such simulations could exploit algebraic ideas that have been largely untapped, promising advances in autonomous systems and other critical applications that, arguably, are not accessible via the methods of machine learning. These distinct but

complementary perspectives indicate possible avenues for further investigation.

AUTHOR CONTRIBUTIONS

Both authors collaborated in writing the article. All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

YMY would like to express his gratitude to Thomas B. Sheridan, formerly of the MIT, for the help, encouragement and inspiration he provided and the ideas and insights he generously shared. KF was funded by the Wellcome Trust (Ref: 088130/Z/09/Z). We would also like to thank our reviewers for invaluable guidance in presenting these ideas.

REFERENCES

Baillieux, H., De Smet, H. J., Paquier, P. E., De Deyn, P. P., and Mainën, P. (2008). Cerebellar neurocognition: insights into the bottom of the brain. *Clin. Neurol. Neurosurg.* 110, 763–773. doi: 10.1016/j.clineuro.2008.05.013

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*. New York, NY: Copernicus Press.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U S A.* 110, 18327–18332. doi: 10.1073/pnas.1306572110

Beim, V. (2012). *The Enigma of Chess Intuition: Can You Mobilize the Hidden Forces in Your Chess?* Alkmaar, Netherlands: The New in Chess Publisher.

Bower, T. G. R. (1974). *Development in Infancy*. San Francisco, CA: W.H. Freeman and Co.

Bressler, S. L. (1994). "Dynamic self-organization in the brain as observed by transient cortical coherence," in *Origins: Brain and Self-Organization*, ed. K. H. Pribram (New Jersey, NJ: Lawrence Erlbaum Associates Publishers), 536–545.

Bressler, S. L., and Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition. *Trends Cogn. Sci.* 5, 26–36. doi: 10.1016/s1364-6613(00)01564-3

Bressler, S. L., and Kelso, J. A. (2016). Coordinations dynamics in cognitive neuroscience. *Front. Neurosci.* 10:397. doi: 10.3389/fnins.2016.00397

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., and Wiener, M. C. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural Comput.* 14, 473–491. doi: 10.1162/089976602317250861

Bruhn, A., Weickert, J., and Shnörr, C. (2005). Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. *Int. J. Comput. Vis.* 61, 211–231. doi: 10.1023/B:VISI.0000045324.43199.43

Bunge, M. (1979). *Causality and Modern Science*. New York, NY: Dover.

Cabessa, J., and Siegelmann, H. T. (2011). "Evolving recurrent neural networks are super-Turing," in *Int. Joint Conf. Neural Networks* (San Jose, CA), 3200–3206.

Camazine, S., Deneubourg, J.-L., Franks, N. R., Sneyd, J., Theraulaz, G., and Bonabeau, E. (2001). *Self-Organization in Biological Systems*. Princeton, NJ: Princeton University Press.

Cantor, G. (1915/1955). *Contributions to the Founding of the Theory of Transfinite Numbers*. Trans., P. E. B. Jourdain (Dover).

Carnap, R. (1962). *Logical Foundations of Probability*. Chicago, IL: The University of Chicago Press.

Casti, J. L. (1979). *Connectivity, Complexity, and Catastrophe in Large-Scale Systems*. New York, NY: John Wiley.

Chaitin, G. (2006). The limits of reason. *Sci. Am.* 294, 74–81. doi: 10.1038/scientificamerican0306-74

Chart, D. (2000). *A Theory of Understanding. Philosophical and Psychological Perspective*. Burlington, VT: Ashgate Publishing Co.

Chung, W.-S., Welsh, C. A., Barres, B. A., and Stevens, B. (2015). Do glia drive synaptic and cognitive impairment in disease? *Nat. Neurosci.* 18, 1539–1545. doi: 10.1038/nn.4142

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Systems Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Craik, K. (1943). *The Nature of Explanation*. Cambridge, MA: Cambridge University Press.

Cushing, J. T. (1994). *Quantum Mechanics. Historical Contingency and the Copenhagen Hegemony*. Chicago, IL: University of Chicago Press.

Damasio, A., and Carvalho, G. B. (2013). The nature of feelings: evolutionary and biological origins. *Nat. Rev. Neurosci.* 14, 143–152. doi: 10.1038/nrn3403

Damasio, A., and Damasio, H. (2016). Exploring the concept of homeostasis and considering its implications for economics. *J. Econ. Behav. Organ.* 126B, 125–129. doi: 10.1016/j.jebo.2015.12.003

Davidson, D. (1970). *Essays on Actions and Events*. New York, NY: Clarendon Press.

Davidson, D. (1993). "Thinking causes," in *Mental Causation*, eds J. Heil and A. Mele (Oxford, NY: Clarendon Press), 3–17.

Destexhe, A., and Sejnowski, T. J. (2009). The Wilson-Cowan model, 36 years later. *Biol. Cybern.* 101, 1–2. doi: 10.1007/s00422-009-0328-3

Dieks, D., and de Regt, H. W. (1998). Reduction and understanding. *Found. Sci.* 3, 45–59. doi: 10.1023/A:1009630119534

Di Ventra, M. D., and Pershin, Y. V. (2013). On the physical properties of memristive, memcapacitive and meminductive systems. *Nanotechnology* 24:255201. doi: 10.1088/0957-4484/24/25/255201

Doran, C., and Lasenby, A. (2003). *Geometric Algebra for Physicists*. Cambridge, MA: Cambridge University Press.

Dorny, C. N. (1975). *A Vector Space Approach to Models and Optimization*. New York, NY: John Wiley & Sons.

Edelman, G. M. (1992). *Bright Air, Brilliant Fire. On the Matter of the Mind*. New York, NY: Basic Books.

Edelman, G. M. (1993). Neural Darwinism: selection and reentrant signaling in higher brain function. *Neuron* 10, 115–125. doi: 10.1016/0896-6273(93)90304-a

Edelman, G. M., and Gally, J. A. (2013). Reentry: a key mechanism for integration of brain function. *Front. Integr. Neurosci.* 7:63. doi: 10.3389/fnint.2013.00063

Edelman, G. M., and Tononi, G. (2000). *A Universe of Consciousness: How Matter Becomes Imagination*. New York, NY: Basic Books.

Einstein, A. (1949). "Autobiographical Notes," in *Albert Einstein: Philosopher-scientist*, ed. P. A. Schlipp (La Salle, IL: Open Court Publishing), 13–69.

Elhilali, M., Fritz, J. B., Chi, T.-S. Shamma, S. A. (2007). Auditory cortical receptive fields: stable entities with plastic abilities. *J. Neurosci.* 27, 10372–10382. doi: 10.1523/jneurosci.1462-07.2007

Ellis, R. D., and Newton, N. (2010). *How the Mind Uses the Brain (To Move the Body and Image the Universe)*. Chicago, IL: Open Court.

England, J. L. (2013). Statistical physics of self-replication. *J. Chem. Phys.* 139:121923. doi: 10.1063/1.4818538

Eysenck, M. W., and Keane, M. T. (1995). *Cognitive Psychology. A Student's Handbook*, 3rd Edn. (Hove, UK: Psychology Press).

Fitelson, B., and Hitchcock, C. (2011). "Probabilistic measures of causal strength," in *Causality in the Sciences*, eds P. M. Illari F. Russo and J. Williamson (Oxford: Oxford University Press), 600–627.

Feynman, R. P., Leighton, R. B., Sands, M. (1964). *The Feynman Lectures on Physics Volume II.*, Reading, MA: Addison-Wesley.

Freeman, W. J., and Holmes, M. D. (2005). Metastability, instability, and state transition in neocortex. *Neural Netw.* 18, 497–504. doi: 10.1016/j.neunet.2005.06.014

Freeman, W. J., Kozma, R., Vitiello, G. (2012). Adaptation of the generalized Carnot cycle to describe thermodynamics of cerebral cortex. *Proc. IEEE WCAI, IJCNN*, Available Online at: http://escholarship.org/uc/item/4087h3bs#page-1

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2013). Life as we know it. *J. Royal Society, INTERFACE*, Available Online at: http://rsif.royalsocietypublishing.org/

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K., Sengupta, B., and Auletta, G. (2014). Cognitive dynamics: from attractors to active inference. *Proc. IEEE.* 102, 427–445. doi: 10.1109/jproc.2014.2306251

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear. Res.* 229, 186–203. doi: 10.1016/j.heares.2007.01.009

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Fuchs, A., Kelso, J. A. S., and Haken, H. (1992). Phase transitions in the human brain: Spatial mode dynamics. *Int. J. Bifurcation Chaos.* 2, 917–939. doi: 10.1142/s0218127492000537

Gentner, D., and Stevens, A. L. (Eds). (1983). *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Georgopoulos, A. P., Kettner, R. E., and Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* 8, 2928–2937.

Georgopoulos, A. P., Taira, M., and Lukashin, A. (1993). Cognitive neurophysiology of the motor cortex. *Science.* 260, 47–52. doi: 10.1126/science.8465199

Gepshtein, S., Lesmes, L. A., and Albright, T. D. (2013). Sensory adaptation as optimal resource allocation. *Proc. Natl. Acad. Sci. U S A.* 110, 4368–4373. doi: 10.1073/pnas.1204109110

Glansdorff, P., and Prigogine, I. (1971). *Thermodynamic Theory of Structure, Stability and Fluctuations.* New York, NY: John Wiley & Sons, Inc.

Glasser, W. (1984). *Control Theory: A New Explanation of How We Control Our Lives.* New York, NY: Harper and Row.

Gross, C. G. (1995). Aristotle on the brain. *Neuroscientist.* 1, 245–250.

Hadamard, J. (1954). *An Essay on the Psychology of Invention in the Mathematical Field.* New York, NY: Dover.

Haken, H. (1983). *Synergetics, An Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology.* New York, NY: Springer-Verlag.

Haken, H. (1993). *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices.* New York, NY: Springer-Verlag.

Halpern, J. Y., and Hitchcock, C. (2015). Graded causality and defaults. *Br. J. Philos. Sci.* 66, 413–457. doi: 10.1093/bjps/jaxt050

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory.* New York, NY: John Wiley & Sons.

Hebb, D. O. (1980). *Essay on Mind.* Hillsdale, NJ: LEA Publisher.

Hempel, C. G. (1962). "Deductive-Nomological vs. Statistical Explanation," in *Minnesota Studies in the Philosophy of Science*, (Vol. III), eds H. Feigel and G. Maxwell (Minneapolis, MN: University of Minnesota Press), 98–131.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York, NY: Free Press.

Hestenes, D., and Sobczyk, G. (1999). *Clifford Algebra to Geometric Calculus. A Unified Language for Mathematics and Physics.* Dodrecht: Kluwer.

Horn, B. K. P., and Schunck, B. G. (1981). Determining optical flow. *Artif. Intell.* 17, 185–203. doi: 10.1016/0004-3702(81)90024-2

Humphrey, N. (2000). *How to Solve the Mind-Body Problem.* Thoverton, UK: Imprint Academic.

Humphrey, N. (2006). *Seeing Red: A Study in Consciousness.* Cambridge, MA: The Belknap Press of Harvard University Press.

Ito, M. (1993). Movement and thought: Identical control mechanisms by the cerebellum. *Trends Neurosci.* 16, 448–450. doi: 10.1016/0166-2236(93)90073-u

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nat. Rev. Neurosci.* 9, 304–313. doi: 10.1038/nrn2332

Jazayeri, M., and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nat. Neurosci.* 9, 690–696. doi: 10.1038/nn1691

Jensen, H. J. (1998). *Self-Organizing Criticality. Emergent Complex Behavior in Physical and Biological Systems.* Cambridge, MA: Cambridge University Press.

Johnson-Laird, P. N. (1983). *Mental Models. Towards a Cognitive Science of Language, Inference and Consciousness.* Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1989). "Mental models," in *Foundations of Cognitive Science*, ed. M. I. Posner (Cambridge, MA: MIT Press), 469–499.

Johnson-Laird, P. N. (2003). "The psychology of understanding," in *The Nature and Limits of Human Understanding*, eds P. N. Johnson-Laird and A. J. Sanford (London: T & T Clark), 3–46.

Kasparov, G. (2007). *How Life Imitates Chess.* New York, NY: Bloomsbury.

Kauffman, S. (2000). *Investigations.* New York, NY: Oxford University Press.

Kelso, J. A. S. (1995). *The Dynamic Patterns. The Self-Organization of Brain and Behavior.* Cambridge, MA: The MIT Press.

Kitcher, P. (1981). Explanatory unification. *Philos. Sci.* 33, 337–359.

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Koestler, A. (1964). *The Act of Creation.* London: Penguin Group.

Köhler, W. (1947). *Gestalt Psychology.* New York, NY: A Mentor Book.

Köhler, W. (1948). *The Mentality of Apes.* London, UK: Routledge and Kegan.

Kohn, A., and Movshon, A. (2004). Adaptation changes the direction tuning of macaque MT neurons. *Nat. Neurosci.* 7, 764–772. doi: 10.1038/nn1267

Kozma, R., Puljic, M., Balister, P., Bollobás, B., and Freeman, W. (2005). Phase transitions in the neuropercolation model of neural populations with mixed local and non-local interactions. *Biol. Cybern.* 92, 367–379. doi: 10.1007/s00422-005-0565-z

Lakoff, G. (2003). "How the body shapes thought: Thinking with all- too-human brain," in *The Nature and Limits of Human Understanding*, ed. A. J. Sanford (London: T & T Clark), 49–74

Lear, J. (1988). *Aristotle: the Desire to Understand.* New York, NY: Cambridge University Press.

Lehar, S. (2003a). Gestalt isomorphism and the primacy of subjective conscious experience: A Gestalt bubble model. *Behav. Brain Sci.* 26, 375–408; discussion 408–443. doi: 10.1017/s0140525x03000098

Lehar, S. (2003b). *The World in Your Head: A Gestalt View of the Mechanism of Conscious Experience.* Hillsdale, NJ: LEA Publishing.

Libet, B., Freeman, A., and Sutherland, J. K. B. (Eds). (2000). *The Volitional Brain: Towards a Neuroscience of Free Will.* Thorverton: Imprint Academic.

Lin, L., Osan, R., Shoham, S., Jin, W., Zuo, W., and Tsien, J. Z. (2005). Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus. *Proc. Natl. Acad. Sci. U S A.* 102, 6125–6130. doi: 10.1073/pnas.0408233102

Lin, L., Osan, R., and Tsien, J. Z. (2006). Organizing principle of real-time memory encoding: neural clique assemblies and universal neural codes. *Trends Neurosci.* 29, 48–57. doi: 10.1016/j.tins.2005.11.004

Lipowsky, R. (1984). Surface-induced order and disorder: critical phenomena at first-order phase transitions. *J. Appl. Phys.* 55, 2485–2490. doi: 10.1063/1.333703

Lucas, B., and Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (Vancouver, BC), 674–679

Luria, A. R. (1973). *The Working Brain.* New York, NY: Basic Books.

MacLennan, B. L. (1998). "Mixing memory and desire: Want and will in neural modelling," in *Brain and values. Is a Biological Science of Value Possible?*, ed. K. H. Pribram (New Jersey, NJ: Lawrence Erlbaum Associates), 31–42.

Marder, M. (2013). Plant intelligence and attention. *Plant Signal. Behav.* 8:e23902. doi: 10.4161/psb.23902

Margenau, H. (1959). *The Nature of Physical Reality.* New York, NY: McGraw-Hill Education.

Moran, J., and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science.* 229, 782–784. doi: 10.1126/science.4023713

Morowitz, H. J. (1979). *Energy Flow in Biology: Biological Organization As a Problem in Thermal Physics.* Woodbridge, CO: Ox Bow Press.

Morowitz, H. J. (1978). *Foundations of Bioenergetics.* Woodbridge, CO: Ox Bow Press.

Mumford, D., and Shah, J. (1985). "Boundary detection by minimizing functionals," in *Proc. IEEE CS Conference Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA), 22–26.

Mumford, D., and Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* 42, 577–685. doi: 10.1002/cpa.3160420503

Murdoch, B. E. (2010). The cerebellum and language: historical perspectives and review. *Cortex.* 46, 858–868. doi: 10.1016/j.cortex.2009.07.018

Newton, N. (1996). *Foundations of Understanding. Advances in Conscious Research.* Amsterdam: John Benjamins Publishing Co.

Noe, A. (2004). *Action in Perception.* Cambridge, MA: The MIT Press.

Nunez, R. E., and Freeman, W. (Eds). (2014). *Reclaiming Cognition: The Primacy of Action, Intention and Emotion.* Thorverton, UK: Imprint Academic.

Palatnik, S., and Khodarkovsky, M. (2014). *The Chess GPS: Improvement of Your Position.* Washington DC: Wildside Press.

Penrose, R. (1997). *The Large, the Small and the Human Mind.* Boston, MA: Cambridge University Press.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness.* New York, NY: Oxford University Press.

Piaget, J. (1950). *The Psychology of Intelligence.* New York, NY: Harcourt Brace.

Piaget, J. (1954). *The Construction of Reality in the Child.* New York, NY: Basic Books.

Piaget, J. (1976). *The Grasp of Consciousness: Action and Concept in the Young Child.* Cambridge, MA: Harvard Univ. Press.

Piaget, J. (1977). *The Development of Thought: Equilibration of Cognitive Structures.* New York, NY: The Viking Press.

Piaget, J. (1978). *Success and Understanding.* Cambridge, MA: Harvard Univ. Press.

Piaget, J., and Inhelder, B. (1969). *The Psychology of the Child.* New York, NY: Basic Books.

Poincare, H. (1952). "Mathematical discovery," in *Science and Method*, ed. H. Poincare (New York, NY: Dover), 46–63.

Pribram, K. H. (1998). "On brain and value: Utility, preference, play and creativity," in *Brain and Values: Is a Biological Science of Values Possible*, ed. K. H. Pribram (New Jersey, NJ: Lawrence Erlbaum Associates Publishers), 43–54.

Prigogine, I. (1994). "Mind and matter: beyond the Cartesian dualism," in *Origins: Brain and Self-Organization*, ed. K. H. Pribram (New Jersey, NJ: Lawrence Erlbaum Associates Publishers), 3–15.

Prigogine, I., and Stengers, I. (1984). *Order Out of Chaos.* New York, NY: Bantam.

Prigogine, I., and Stengers, I. (1997). *The End of Certainty.* New York, NY: Simon and Schuster.

Prokopenko, M., Polani, D., and Ay, N. (2014). "On the cross-disciplinary nature of guided self-organization," in *Guided Self-Organization: Inception*, ed. M. Prokopenko (Berlin: Springer), 3–18.

Purushotaman, G., and Bradley, D. C. (2005). Neural population code for fine perceptual decisions in area MT. *Nat. Neurosci.* 8, 99–106. doi: 10.1038/nn1373

Razi, A., and Friston, K. J. (2016). The connected brain: causality, models, and intrinsic dynamics. *IEEE Signal Proc. Mag.* 33, 14–35. doi: 10.1109/msp.2015.2482121

Rosenbloom, M. H., Schmahmann, J. D., and Price, B. H. (2012). The functional neuroanatomy of decision-making. *J. Neuropsychiatry Clin. Neurosci.* 24, 266–277. doi: 10.1176/appi.neuropsych.11060139

Salman, M. S. (2002). The cerebellum: new insights into the role of the cerebellum in timing motor and cognitive tasks. *J. Child Neurol.* 17, 1–9. doi: 10.1177/088307380201700101

Salmon, W. C. (1970). *Statistical Explanation and Statistical Relevance.* Pittsburgh, PA: University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World.* Chicago, IL: Princeton University Press.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation.* Pittsburgh, PA: University of Pittsburgh Press.

Sanford, A. J. (1987). *The Mind of Man: Models of Human Understanding.* New Haven, CT: Yale University Press.

Sarti, A., Citti, G. (2015). The constitution of visual perceptual units in the functional architecture of V1. *J. Comp. Neurosci.* 38, 285–300. doi: 10.1007/s10827-014-0540-6

Schooler, J. W., Ohlsson, S., and Brook, K. (1993). Thoughts beyond words: when language overshadows insight. *J. Exp. Psychol. Gen.* 2, 166–1843. doi: 10.1037/0096-3445.122.2.166

Schrodinger, W. (2006). *What is Life?* New York, NY: Cambridge University Press.

Sejnowski, T., Poizner, H., Lynch, G., Gepshtein, S., and Greenspan, R. J. (2014). Prospective optimization. *Proc. IEEE Inst. Electr. Electron. Eng.* 102, 799–811. doi: 10.1109/JPROC.2014.2314297

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and efficiency in the nervous system-a synthesis. *PLoS Comput. Biol.* 9:e1003157. doi: 10.1371/journal.pcbi.1003157

Sengupta, B., Tozzi, A., Cooray, G. K., Douglas, P. K., and Friston, K. J. (2016). Towards a neuronal gauge theory. *PLoS Biol.* 14:e1002400. doi: 10.1371/journal.pbio.1002400

Seth, A. (2014). "The cybernetic brain: from interoceptive inference to sensorimotor contingencies," in *Open MIND*, eds T. Metzinger and J. Windt (Frankfurt AM: MIND Group), 1–24.

Shah, J. (1992). Properties of energy-minimizing segmentations. *SIAM J. Control Optim.* 30, 99–111. doi: 10.1137/0330007

Shannon, C. E. (1941). Mathematical theory of the differential analyzer. *J. Math. Phys.* 20, 337–354. doi: 10.1002/sapm1941201337

Shastri, L. (2001). "Biological grounding of recruitment learning and vicinal algorithms in long-term potentiation," in *Emergent Neural Computational Architectures Based on Neuroscience–Towards Neuroscience-Inspired Computing*, eds J. Austin, S. Wermter and D. Wilshaw (Berlin: Lecture Notes in Computer Science, Springer-Verlag), 348–367.

Siebeck, U. E., Litherland, L., and Wallis, G. M. (2009). Shape learning and discrimination in reef fish. *J. Exp. Biol.* 212, 2113–2119. doi: 10.1242/jeb.028936

Siegelmann, H. T. (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit.* Cambridge, MA: Birkhauser Boston Inc.

Sierpinska, A. (1994). *Understanding in Mathematics.* London: The Falmer Press.

Singer, W. (2009). The brain, a complex self-organizing system. *Eur. Rev.* 17, 321–329. doi: 10.1017/s1062798709000751

Sjölander, S. (1995). Some cognitive break-through in the evolution of cognition and consciousness and their impact on the biology of language. *Evol. Cogn.* 1, 3–11.

Sloman, S. (2005). *Causal Models. How People Think About the World and the Alternatives.* New York, NY: Oxford University Press.

Smolensky, P. (1990). Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems. *Artif. Intell.* 46, 159–216. doi: 10.1016/0004-3702(90)90007-m

Sperry, R. W. (1969). A modified concept of consciousness. *Psychol. Rev.* 76, 532–536. doi: 10.1037/h0028156

Syms, R. R. A., Yeatman, E. M., Bright, V. M., and Whitesides, G. M. (2003). Surface-tension powered self-assembly of microstructures–state-of-the-art. *J. Microelectromech. Syst.* 12, 387–417. doi: 10.1109/jmems.2003.811724

Thom, R. (1975). *Structural Stability and Morphogenesis. An Outline of a General Theory of Models.* Reading, MA: W.A. Benjamin, Inc..

Thompson, E., and Varela, F. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends Cogn. Sci.* 5, 418–425. doi: 10.1016/s1364-6613(00)01750-2

Tiles, M. (1989). *The Philosophy of Set Theory.* New York, NY: Dover Publications.

Toulmin, S. (1961). *Foresight and Understanding.* London: Hutchison.

Trevawas, A. (2002). Mindless mastery. *Nature.* 415:841. doi: 10.1038/415841a

Tsien, J. Z. (2007). The memory code. *Sci. Am.* 297, 52–57. doi: 10.1038/scientificamerican0707-52

Tschacher, W., and Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organised pattern formation. *New Ideas Psychol.* 25, 1–15. doi: 10.1016/j.newideapsych.2006.09.002

van Fraasen, B. (1980). *The Scientific Image*. Oxford: Clarendon Press.

von Wright, G. H. (1971). *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

Werbos, J. P. (1994). "Self-organization: Reexamining the basics and an alternative to the Big Bang," in *Origins: Brain and Self-Organization*, ed. K. H. Pribram (Hillsdale, NJ: Lawrence Erlbaum Associates Publishers), 16–52.

Werbos, J. P. (1996). "Optimization: A Foundation for Understanding Consciousness," in *Optimality in Biological and Artificial Networks?* eds S. Levine and W. S. Elsberry (Hillsdale, NJ: Lawrence Erlbaum Associates Publishers), 19–42.

Werbos, J. P. (1998). "Values, goals and utility in engineering-based theory of mammalian intelligence," in *Brain and Values: Is a Biological Science of Values Possible*, ed. K. H. Pribram (Hillsdale, NJ: Lawrence Erlbaum Associates Publishers), 55–76.

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/s0006-3495(72)86068-5

Wilson, H. R., and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik.* 13, 55–80. doi: 10.1007/bf00288786

Yantis, S. (1992). Multi-element visual tracking: attention and perceptual organization. *Cogn. Psychol.* 24, 295–340. doi: 10.1016/0010-0285(92)90010-y

Yufik, Y. M. (1998a). "Virtual associative networks: a framework for cognitive modelling," in *Brain and Values*, ed. K. H. Pribram (New York, NY: Lawence Erlbaum Associates), 109–177.

Yufik, Y. M. (1998b). Probabilistic resource-allocation system with self-adaptive capabilities. US Patent 5,794,224.

Yufik, Y. M. (2002). "How the mind works: An exercise in pragmatism", in *Proceedings of the 2002 International Joint Conference Neural Networks, 2002. IJCNN 02* (Honolulu: HI), 2265–2269.

Yufik, Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. doi: 10.1016/j.chaos.2013.04.010

Yufik, Y. M., and Malhotra, R. (1999). Information blending in virtual associative networks: a new paradigm for sensor integration. *Int. J. Artif. Intell. Tools* 8, 275–290. doi: 10.1142/s0218213099000191

Yufik, Y. M., and Sheridan, T. B. (1997). Virtual networks: new framework for operator modelling in complex systems. *Annu. Rev. Control.* 20, 179–195. doi: 10.1016/s1367-5788(97)00016-3

Yufik, Y. M., and Sheridan, T. B. (2002). Swiss army knife and Ockham's razor: modelling operator's comprehension in complex dynamic tasks. *IEEE Trans. Syst. Man Cyber. A Syst. Hum.* 32, 185–199. doi: 10.1109/tsmca.2002.1021107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on