The long and winding road to a science of consciousness

E.A. Stoll
stoll@westerninstitute.org
Western Institute for Advanced Study
Denver, Colorado, USA

Revised: 2 April 2023

Abstract

Historically, it has been difficult to articulate exactly what consciousness is, much less uncover the mechanisms underlying this phenomenon. Reframing the hard problem – from the impossible task of explaining 'the subjective experience of being' toward explaining 'perceptual information content, predictive models of the world and the self that grow over time, and non-deterministic behavioral outcomes' – has made the problem more neuroscientifically tractable. The new framework of Conifold Theory provides a mechanistic process by which these features emerge naturally from neural network computations, if certain anatomical and physiological criteria are met. This report details over two dozen approaches to studying the problem, describing the path taken over many years toward developing a science of consciousness. Each of these approaches has made enormously valuable contributions to the research question at hand, by articulating some key feature of consciousness and offering a method for tackling the gaps in our understanding. Building on this hard-won progress, the new framework of Conifold Theory vindicates and elaborates critical components of previously-proposed theories of consciousness, while offering a more complete and detailed explanation of the physical mechanisms underlying this phenomenon. In retrospect, it seems the field, which has often appeared to be groping in the dark, was making steady progress toward answering its biggest question all along.

Introduction

Consciousness is characterized by a cohesive stream of multi-modal perceptual experience, which is continuously updated with incoming sensory input; the formation of cognitive constructs, which predict cause-effect relationships between actions and outcomes; and the ability to select context-appropriate behavior, based on incoming sensory data and relevant memories. All three features of consciousness are expected to have some neural basis, with perception (Elgueda et al., 2019; Griffiths et al., 1998; Kanwisher et al., 1997; Olthof et al., 2019; Pins & Ffytche, 2003; Solomon & Lennie, 2007), predictive modeling (Auksztulewicz et al., 2018; Forseth et al., 2020; Kok et al., 2016; Schwiedrzik et al., 2014; Snyder et al., 2015), and decision-making (Beck et al., 2008; Hanks, 2011; Mostert et al., 2015) all dependent upon the electrochemical activity of cortical neurons.

And yet, the exact relationship between neural activity and the intangible features of consciousness has been difficult to understand. We collect data about our local environment through individual sensory apparatus and process these data in separate regions of the brain, but we nonetheless experience a bound, cohesive *perception* of the world around us (Engel & Singer, 2001; von der Malsburg, 1995). Similarly, the emergence of non-material *cognitive* constructs from neural activity has long been considered problematic, with the very nature of these constructs difficult to explain in terms of physical laws (Collell & Fauquet, 2015; Rees et al., 2002). And thirdly, identifying some mechanism for directing *willed behavior* has proven difficult, particularly within the classical assumptions of causal determinism (O'Connor & Franklin, 2020).

A new approach, Conifold Theory, resolves all three of these phenomena – perception, cognition, and non-deterministic behavior – into a single process of neural computation. By modeling how cortical neurons retain sensitivity to probabilistic particle behavior during the encoding process, it becomes apparent how these computational units cyclically generate and compress information, building a cohesive, accessible representation of incoming sensory data and gleaning predictive value from this information to direct behavioral outcomes. By retaining a focus on the mechanics of probabilistic coding, Conifold Theory provides an explanation of cortical neural network function which accounts for key features of consciousness, including streaming perceptual experience, predictive processing, spontaneous structural remodeling to store memory, and goal-directed behavior. This new theory builds on the progress achieved by neuroscientists, philosophers, mathematicians, and information theorists by reconciling a number of previously-articulated concepts within a larger framework.

The aim of this article is to detail the critical theoretical advances made by those who have deeply considered the question of consciousness, exploring the contributions of each approach in order to recognize how we arrived at our current state of knowledge. The key frameworks addressed here include cognitive models such as Global Workspace Theory, the Somatic Marker Hypothesis, the Phenomenal Self Model and the Self-Organizing Meta-Representational Account; eliminative theories such as the Multiple Drafts Model and Attention Schema Theory; alternative physical theories of consciousness like Orchestrated Objective Reduction Theory and Conscious Electromagnetic Field Theory; neuroscientific approaches such as Predictive Processing and Semantic Pointer Architecture; and finally, mathematically grounded approaches like Integrated Information Theory, which focus on the key role of information in conscious processes. During this exercise, it shall become apparent that each theoretical framework has made enormously useful contributions to the research question at hand, by articulating some key feature of consciousness and positing a method for tackling the gaps in our understanding. By building on this hard-won progress, the new framework of Conifold Theory offers a great deal of explanatory power. However, this comprehensive theory could not have been created without the fundamental concepts underpinning cognition having been articulated by philosophers and without the notion of consciousness as a physical process of neural computation having been established through decades of neuroscientific investigation.

Moving forward, it will be useful to celebrate how far we have traveled with diligent effort and decide how we will push our understanding even further. Indeed, describing the relationship between the mind and the brain in clear, neuroscientifically-grounded terms should benefit the cognitive sciences enormously, with valuable implications for basic research and clinical practice. The progress made in merging the fields of neuroscience, non-equilibrium thermodynamics, and information theory should also yield significant advances in computing. In summary, the field of cognitive science has arrived at an enormously promising moment, and it is worth acknowledging the hard work that went into laying the foundation for a mature science of consciousness.

An overview of key theoretical approaches

Metaphysics, epistemology, and the role of sensation in perceiving the world

At the dawn of the modern age of reason, philosophers began to attack the problem of consciousness alongside the problem of existence itself, finding the metaphysical question of *what we are* to be tied to the epistemological question of *how we come to know anything at all*.

The enlightenment philosopher John Locke argued that human beings are born without innate knowledge or character, instead building an understanding of the world through lived experience (Locke, 1689). In a view that is enormously compatible with modern neuroscience, he rejected innate knowledge and placed the responsibility for gaining understanding solely on human endeavor, via the dual processes of perception and reflection. He wrote: "Let us then suppose the mind to be, as we say, white paper, void of all characters, without any ideas; how comes it to be furnished? Whence comes it by that vast store, which the busy and boundless fancy of man has painted on it, with an almost endless variety? Whence has it all the materials of reason and knowledge? To this I answer, in one word, from *experience*…. First *our senses*, conversant about particular sensible objects, do *convey into the mind*, several distinct *perceptions* of things…. Secondly, the other fountain, from which experience furnisheth the understanding with ideas, is the *perception of the operations of our own minds* within us."

Yet this view – that we furnish our minds with knowledge solely through effort and work – was not universally shared. Locke's contemporary René Descartes despaired that epistemology was intrinsically flawed, and doubted the bodily senses could ever reveal the true nature of things (Descartes, 1641). He argued that ideas and the objects they represent are separate, and a person could only be sure of the ideas, not the things themselves: "Even as I speak, I put the wax by the fire and look: the shape is lost, the size increases; it becomes liquid and hot; you can hardly touch it, and if you strike it, it no longer makes a sound. But does the same wax remain? It must be admitted that it does; no one denies it, no one thinks otherwise. So what was it in the wax that I understood with such distinctness? Evidently none of the features which I arrive at by means of the senses; for whatever came under taste, smell, sight, touch or hearing has now altered – yet the wax remains… I must therefore conclude that the nature of this piece of wax is in no way revealed by observation, but is perceived by the mind alone."

Both Locke and Descartes wisely separated perception, which focuses on external stimuli, from reflection, which focuses on the internal state, and both philosophers noted the inherent subjectivity of all thought processes. There is always an 'I' – with memories, beliefs, biases, and intentions – doing the thinking. Yet while Locke believed in the existence of an objective reality that was perceived by the sensory apparatus, Descartes took the unreliability and incompleteness of sensory information to mean the world was not truly accessible to the mind at all. He contended that his own thoughts certainly existed, but nothing else could be ascertained.

Locke acknowledged that our cognitive processes might be imperfect, but they provided the best instruments we have at our disposal to attack the problem. Indeed, in the following years, a decision to embrace the mind as an epistemological tool allowed humanity to proceed with investigations into physics. Understanding the laws of nature, including temperature-dependent state changes in materials like the wax in the fire, vindicated the idea that our senses *can* collect practical information about the world and our reason *can* explicate these observations. Our faculties of sensation and reason in turn allow us to accumulate knowledge about the world and build technology that is enormously useful in navigating the world around us. The scientific method formalizes this process, with descriptive observation permitting the construction of cognitive models, or theories, regarding the structure and operation of the world. These theories can then be articulated into testable predictions, to set expectations for the specific events that should be observed in a specific context. The results of these scientific investigations provide valuable feedback, with empirical results allowing us to hone our theories and construct a better knowledge of reality. A human being does this informally every day, but the scientific method ensures that we consciously eliminate anomalous data and discard earlier models as new information comes along.

The goal of modern cognitive science is to subject the mind itself to scientific enquiry. This effort may yield not only progress in metaphysics – understanding *what we are* – but also help us to address the epistemological question of *how we come to know things* and the moral question of *what we should do with this incredible power*. Yet not everyone wants the mind to be subjected to scientific enquiry. Descartes' dualistic view of reality and the self remains influential, particularly among those who prefer to believe that consciousness is a spiritual substance rather than a physical process. These two different perspectives, introduced by Locke and Descartes, not only indicate a continuing discrepancy in how we perceive ourselves as human beings, but also how much we trust in reality, and whether we believe ourselves to have moral responsibility for our actions.

Phenomenal and Access Consciousness

The twentieth century saw great progress in considering the problem of consciousness, building on a strong foundation of scientific advances in physics and biology. In 1945, the philosopher Maurice Merleau-Ponty took an important step toward articulating the nature of consciousness with his publication of *Phenomenology of Perception* (Merleau-Ponty, 1945). In this work, Merleau-Ponty argued against Cartesian dualism and in favor of a more dialectical concept of consciousness. This approach usefully built on Locke's philosophical groundwork by differentiating between reality, the bodily process of acquiring data about reality, and the subjective experience of perceptual awareness. In this view, the irreducible conscious experience is not equivalent to the biophysical processes which permit the collection of data through the sensory apparatus. These biophysical processes do however contribute to the stream of perceptual experience, which in turn forms the foundation of the gestalt self that reflects on reality, the bodily self, and the nature of this contemplative, observational existence. Merleau-Ponty asserted "the primacy of perception" in our efforts to understand the world, placing the physical body at the center of the debate by suggesting it deserves attention as both subject (the one understanding) and object (the one being understood). This work built on the work of earlier philosophers, particularly Locke, by placing both the body and the mind squarely within reality.

In the 1990s, the philosopher Ned Block articulated the concept of consciousness more clearly by differentiating between phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness), arguing that these are two separate aspects of conscious existence which are often erroneously conflated (Block, 1995). P-consciousness has 'experiential properties' known as 'qualia', including sights, sounds, smells, emotions, and desires. Meanwhile, A-consciousness is defined as a state whose representational content is "poised for use as a premise in reasoning, poised for rational control of action, and poised for rational control of speech". P-consciousness is the pure experience of being in the present moment – for example, a child riding a bike might feel the air on her skin and the giddiness of speeding downhill; she might smell hot asphalt and freshly-mown grass, hear dogs barking, see the trees and houses blurring past. A-consciousness is the awareness of having an experience and being able to use this information – for example, to press the brakes at the right moment, or holler a greeting to a neighbor. The very nature of P-consciousness – the cohesive, multi-modal stream of perceptual experience – has been dubbed 'the hard problem' by the philosopher David Chalmers, because this phenomenon seems unamenable to standard methods of cognitive science and unexplainable in terms of computational or neural mechanisms (Chalmers, 1995). Understanding what perceptual experience is, and what causes it, may aid in understanding whether it contributes at all to the cognitive processing that occurs between sensory input and motor output.

The question arises as to how these two key aspects of consciousness are related. Block argues that a lack of A-consciousness would not guarantee a lack of P-consciousness; being unable to report an experience does not necessarily imply that experience is not happening (Block, 1995). In other words, P-consciousness could potentially exist without A-consciousness. This might be the case, for example, in non-human animals and infants. Yet Chalmers has pointed out that access to perceptual experience is essentially an availability of perceptual content, and this global access always seems to accompany perceptual experience (Chalmers, 1997). A converse problem is whether a lack of P-consciousness would guarantee a lack of A-consciousness. Block argues A-consciousness cannot exist without P-consciousness, since ‘P greases the wheels of A’, providing perceptual material to be aware of and report on (Block, 1997). Meanwhile, Chalmers argues that P-consciousness is merely an epiphenomenon, completely superfluous to the mechanisms of information processing, since one can imagine a ‘zombie’ that has the same causal machinery without having any perceptual experience at all (Chalmers, 1997). If zombies are possible, and functionally isomorphic to human beings, then perceptual experience itself is not necessary to form access states. If zombies are *not* possible, and perceptual experience is indeed an integral part of the computational processes underlying generalized intelligence, then P-consciousness may contribute to A-consciousness, and both may contribute to causal processes in the brain.

Developmental studies suggest that P-consciousness does permit A-consciousness, with incoming sensory data cohesively merging together and accumulating over time to form a cognitive model of how the world operates (Csibra et al., 2000; Grossmann et al., 2007). The reliance on P-consciousness for building accurate cognitive models is also evidenced by a reliance on sensory feedback during motor learning (Herrmann et al., 2004; Olthof et al., 2019; Pistohl et al., 2015). The mental experience also seems to drive the selection of behavior. Patients with severe damage to the visual cortex, who do not admit to perceiving a presented stimulus, can name the object with surprising accuracy when prompted with a choice task – yet one interesting feature of blindsight is that “subjects never initiate on their own any actions informed by perceptions from the blindfield… [it seems] that information must normally be represented in phenomenal consciousness if it is to play any role in guiding voluntary action” (van Gulick, 1989). These empirical studies suggest that P-consciousness, or at least the neural correlates underpinning it, is the starting point for all other conscious activities: the reporting of experiences, the calculation of optimal movements to achieve a given outcome, and the initiation of flexible, goal-directed behavior. This evidence is far more in line with Block's view of consciousness as an active loop of collecting information and using this information to act effectively in the world than Chalmers' view of P-consciousness as a mere epiphenomenon. However, the latter view may be usefully posed as a null hypothesis, with phenomenal consciousness assumed to do nothing unless it is proven to have some causal role in directing cortical neuron activity and subsequent behavioral outcomes.

First-Order and Higher-Order Theories

First-order theorists, like Block, maintain that the only thing required for phenomenal conscious awareness of a given stimulus is the cortical processing that is directly prompted by that stimulus (Block, 2011a, 2011b). Phenomenal conscious states, in this view, are unrefined states of awareness of the local environment, permitted by the acquisition of sensory data. Additional processes, such as attention, working memory, and metacognition, merely allow cognitive access to, and introspection about, the first-order state. By contrast, David Rosenthal and other higher-order theorists argue that a first-order state resulting from neural activity in visual cortex or auditory cortex is not sufficient to bring a stimulus into conscious awareness (Rosenthal, 2005). Here, an access state is *required* for perception. Evidence for this claim includes the "compelling subjective impression that peripheral vision is [not] impoverished" when there is in fact quantitatively less sensory data collected in the periphery of the retina than the foveal region (Lau & Rosenthal, 2011). This powerful and consistent illusion is not caused by a detailed first-order representation, but rather a top-down process of 'filling-in' missing or impoverished data. Higher-Order Theory further predicts that changes in phenomenal conscious awareness *can* occur in the presence of an unchanging stimulus and *only* occur upon changes in the activity of higher cortical regions (Rosenthal, 2008).

According to Higher-Order Theory, a person is not conscious of a first-order state, except by virtue of the higher-order representation (HOR) of it (Brown, 2015). The HOR allows *access* to the first-order state. Higher-order theorists use the term "introspection" for this additional level of representation, with "introspection" referring to situations when an individual is attentively and deliberately focused on the first-order state, accessing the content of this perceptual experience (Rosenthal, 2005). And so, to be consciously aware of a *higher-order state* requires yet a further higher-order representation (Block, 2011a). A state of higher-order awareness in the absence of any particular first-order representation is useful, permitting an individual to think about thinking in a general cognitive sense. This HOR of a HOR is called a HOROR (Brown, 2015). While this framework is quite useful for considering access consciousness as a lens for observing reality, the predictions of Higher-Order Theory have not entirely held up to scrutiny (Malach, 2011).

HORS can represent internal emotional states as well as multi-modal sensory data collected from the external environment (LeDoux & Brown, 2017). Higher-order theorists posit that brain mechanisms which give rise to conscious emotional experiences "are not fundamentally different from those that give rise to perceptual conscious experiences." In this view, both internally-derived and externally-derived data must involve higher-order representations of lower-order information, encoded within the cortical neural network. "Thus, subcortical circuits are not responsible for feelings, but instead provide lower-order, non-conscious inputs that coalesce with other kinds of neural signals in the cognitive assembly." As a result, the conscious awareness of fear, and behaviors linked to defensive survival responses, are not driven by subcortical circuits. Rather, subcortical circuits provide raw input data for higher-order access, with proprioceptive information entering the prefrontal circuitry alongside sensory data in each modality. The active modules in higher cortical regions may then represent a schema of the fear-inducing stimulus, relevant visual and auditory cues, *and* a schema of the self which is potentially endangered by the stimulus. In this view, there must be a self in order to feel fear: "If you are not aware of being afraid, you are not afraid; if you are not afraid, you are not feeling fear." This view is consistent with classical neurology practice, in which ablation of the prefrontal cortex, rather than the amygdalar circuitry, was thought to be the most effective method of reducing emotionality and emotionally-driven behavior in human patients (Gallea, 2017). Indeed, the shameful history of psychosurgery has taught us that, without an intact prefrontal cortex, a human being has a diminished understanding of the self as a persistent entity, along with a reduced emotional affect and a reduced behavioral drive (Stuss, 1991). Higher-Order Theory asserts that emotions and percepts do not have any representational existence *unless* prefrontal cortical circuits permit these data to coalesce into an interpretable representation. In this way, Higher-Order Theory defines a meaningful connection between streaming perceptual experience and self-awareness.

Global Workspace Theory

An alternative framework for explaining the cognitive architecture underlying conscious experience is called Global Workspace Theory (Baars, 1997). This theory, initially proposed by Bernard Baars then elaborated by Stanislas Dehaene, compares the experience of consciousness to a theater. There, focal consciousness acts "as a bright spot on the stage, directed there by the spotlight of attention" while "the entire stage of the theater corresponds to working memory". The whole event is produced continuously for the benefit of an audience of one, with unconscious processes working behind the scenes, out of view, off-stage. The immersive experience of the theater provides a singular stream of information – vast, yet also intrinsically limited in its representation of reality.

The Global Workspace does not arise solely with prefrontal cortex activation, but rather through a "neuronal avalanche" (Dehaene, 2014). Information about the world, entering through each sensory apparatus, is integrated by sequential processing steps in thalamic and cortical regions of the brain, then 'selected' by attentional processes. The resulting spotlight of attention, molded by coincident events across sensory modalities, is oriented toward the collection of additional sensory data, which are then broadcast throughout the cortex. In this way, a jumble of inputs across sensory modalities, occurring close together in space and time, can converge onto a single coherent interpretation – for example, a playground filled with screaming children. This singular *interpretation* of objects and events within the global workspace, arising from the functions of perception and memory, represents the conscious state of the organism. Here, intrinsic mental constructs, not extrinsic reality or sensory input representing extrinsic reality, form the basis of conscious experience (Dehaene et al., 1998). In this view, cortical neural networks prefer to rehearse existing information sets rather than reacting to externally-derived information, engaging in the latter process only as necessary. Higher regions of cortex gate their input sources, permitting sensory data to be released into the global workspace only as needed. Rather than objectively instructing the system on the external state of the environment, these signals merely result in the selection of one pre-existing internal representation among many.

This framework leads to some fairly specific predictions about the brain regions involved in completing tasks which require attention and effort. The theory distinguishes two computational spaces: a global workspace composed of sparse but heavily-interconnected neurons with long-range axons, and individual modules dedicated to sensory perception, movement initiation, memory, reflective evaluation, and attention (Dehaene et al., 1998). These modules contribute to the global workspace, with the workspace itself selectively mobilizing or suppressing the contribution of specific neurons within each module. During task performance, the contributing neurons become spontaneously coordinated, forming the global workspace. This theoretical framework leaves aside the mechanisms by which top-down processes ‘decide’ the contributions of individual modules to system-wide computational processes, and why these computational processes give rise to a stream of perceptual experience, but any theory which leads to the formulation of testable predictions about the relationship between perceptual experience and neural activity, with measurable endpoints, certainly provides a useful advance. Indeed, some predictions have borne out, with experiments demonstrating a role for frontal-parietal regions of cortex in conscious information processing. One key experiment explored the correlation between cortical activity and motor output, showing a delayed activation of higher cortical regions during quick task-switching, when behavioral response times are impaired due to a computational bottleneck (Hesselmann et al., 2011). Another key experiment demonstrated that reported stimulus visibility and objective task performance were correlated with each other, and both were impaired in patients with white matter tract lesions due to multiple sclerosis; variation in the severity of both subjective impairment and task impairment corresponded to the extent of damage (Reuter et al., 2009). These studies tie together the structural and functional integrity of cortical neural networks with the perceptual experience of sensory stimuli and the organization of response behavior. Indeed, there is little doubt that these cortical modules are involved in completing cognitive tasks, and the strength of populational activity in these brain regions is correlated with the amount of attention required to complete the task.

Recurrent Processing Theory

Another cognitive architecture that has been set forth to explain the conscious experience is Recurrent Processing Theory, proposed by the neuroscientist Victor Lamme (Lamme, 2006). This framework agrees with Global Workspace Theory in that separate modules, such as cortical regions dedicated to individual sensory modalities, contribute to phenomenal consciousness. But Recurrent Processing Theory diverges in describing access consciousness. Here, access consciousness is not simply the selection of information arising from multiple source modules and impinging on the prefrontal cortex, but rather recurrent processing through extensive cortico-cortical connections, which occurs *after* that feedforward sweep of information processing.

In this view, modular activity in primary visual cortex or primary auditory cortex are not sufficient to perceive the stimulus in either modality; in this general sense, Recurrent Processing Theory is more in line with Higher-Order Theory than First-Order Theory. Yet directing modular inputs into prefrontal cortex is not sufficient for conscious awareness of the stimulus either. While this activity may lead to the unconscious priming of attention, it does not in itself prompt conscious awareness.

Local recurrent processing *within* modules, followed by widespread recurrent processing *between* modules, binds the content together, so the properties of a stimulus can be consciously and cohesively perceived (Vandenbroucke et al., 2016). Some fragile memory formation during the initial unconscious processing, within the relevant sensory module, permits the subsequent identification of a minimally-attended object, once attention, memory, and cognitive processes are integrated with that sensory information (Pinto, Vandenbroucke, et al., 2017). In short, once long-range recurrent processes are in effect, the data that is cached in local modules becomes *accessible* – these data are called upon and brought into the conscious awareness. For this reason, the identification of isolated objects or the interpretation of simple scenes requires little recurrent processing, but the parsing of complex scenes relies heavily on feedback loops between visual cortex and frontoparietal cortex (Groen et al., 2018).

This theoretical framework is a useful approach for three key reasons. Firstly, this model relies on some of the unique anatomical and physiological properties of cortical neural networks, which are not present in simpler spinal reflex circuits, to explain conscious processes. Cortico-cortical feedback loops go some way toward explaining why conscious awareness of environmental stimuli co-occurs with cortical neuron activity, but not with activity in spinal reflex circuits. Secondly, this model explains why there are time delays in event-related potentials relating to object recognition in V1, particularly in the context of complex scenes. Certainly, these data imply that conscious processing of sensory information requires re-orientation toward the attended scene to extract further data, before the interpretation of the scene can be properly determined. Thirdly, this model is useful because it postulates that a specific, observable aspect of a biological neural network provides a reliable correlate of consciousness (Tsuchiya et al., 2015). This kind of objective gauge, paired with measurable endpoints, is required for any theory of consciousness to be testable and falsifiable (Kleiner & Hoel, 2021).

The only problem with this theory is that it implies any biological or synthetic neural network with recurrent processes should be conscious – not only of stimuli in the environment, but also of itself as an observer. This logical conclusion presents an ontological problem, as many of the deep neural networks around today (such as AlphaGo) would then already fit the definition of consciousness, despite a notable lack of generalized problem-solving ability and a corresponding lack of unprompted, unprogrammed behavioral output (Sabokrou et al., 2017; Silver et al., 2016; Wang et al., 2017). Perhaps recurrent processing is *required* for consciousness, but it is not *sufficient*. Other factors may be necessary to achieve conscious awareness of attended stimuli in the environment, as well as conscious awareness of the self as a capable observer.

Somatic Marker Hypothesis

The unique experience of *being* a self remains difficult; it is not just a matter of being in a position in space and time to observe external events. There is something about the act of being, and having a self who *experiences* that act of being, that requires explanation. The Somatic Marker Hypothesis, proposed by the neuroscientist Antonio Damásio, aims to address this problem. The Somatic Marker Hypothesis asserts that emotions and their biological underpinnings contribute heavily to cognitive awareness and subsequent decisions (Damásio, 2000). This theoretical framework significantly reduces the focus on externally-derived sensory data in constructing the subjective mental state. Instead, the Somatic Marker Hypothesis contends that conscious mental states are largely internally-derived, arising in response to the autonomic nervous actions of the body, such as changes in heartbeat and blood pressure. This approach was posited in opposition to classical neuroscience, which holds that behaviors are selected largely as a response to external cues in the environment. Over the past two decades, the Somatic Marker Hypothesis has contributed greatly to the formulation of Higher-Order Theory, Global Workspace Theory, and Recurrent Processing Theory, which all posit a key role for emotional inputs in governing the mental state and driving the construction of cognitive models. In this view, the subjective perceptual experience which most informs goals and actions is that representing the bodily state – and indeed, there is *no concept of the self* without a mental representation of the bodily state.

In this view, the self is a central entity, combining sensory inputs, proprioceptive feedback, genetic predispositions, and stored memories from previous experiences. Damásio explains: "The non-conscious neural signaling of an individual organism begets the proto-self, which permits… core consciousness, which allows for an autobiographical self, which permits extended consciousness. [And] at the end of the chain, extended consciousness permits conscience." The emergence of a self-construct depends upon physical processes, from induced changes in gene expression to autonomic feedback loops. As such, explanations of the self tend to focus on the physical "I" – the one with *that particular set of genes*, the one experiencing *that racing heartbeat*.

The Somatic Marker Hypothesis specifically calls for a role of emotional information processing in cognitive goal formation and behavioral choice. For example, in a gambling task, patients with damage to the prefrontal cortex exhibited impaired decision-making, which was paired with an inability to drive an anticipatory rise in skin conductance when engaging in risky decisions (Bechara et al., 1996). The results of this experiment suggest that disabled emotional feedback signals impair the decision-making process. However, these findings are confounded by potential common cause, and may be better explained by the faulty assessment of risk magnitude through impaired executive function *and* the faulty top-down control of autonomic processes by the prefrontal cortex, rather than a bottom-up failure to perceive the emotional valence of a situation (Tomb et al., 2002).

Yet regardless of how autonomic processes contribute to the conscious state, we can appreciate that conscious states do depend on bodily cues. Certainly, our conscious awareness encompasses a mix of internal and external stimuli, with data arising from proprioceptive inputs as well as external sensory inputs. And certainly, information about the present bodily state influences the self-concept, the formation and prioritization of goals, and the cognitive estimates of feasible action in the world.

Overall, the concept of a cognitive architecture which takes into account multiple types of sensory data to inform decision-making is incredibly useful, regardless of whether externally-derived cues from the local environment or proprioceptive cues from the body turn out to dominate.

Phenomenal Self Model

To explain the very nature of the self and the usefulness of this cognitive construct, the philosopher Thomas Metzinger has proposed the Phenomenal Self Model (Metzinger, 2003). This framework provides an analytic approach to generalizing the notion of the subjective self, so that it can be applied to any potentially conscious entity – biological or engineered. The Phenomenal Self Model has several key properties. First, there is a sense of ownership over the very concept of the self. The cognitive construct identifies with *being* the cognitive construct. Second, the self-model is dependent on the unique perspective of the body in space. The self always perceives the world from some perspective, in a manner limited by the range of the sensory apparatus and the relative position of the body within space. Third, the self-model evolves with time. The construct narrates itself, fixing each episode of experience into some *representation* of the relationship between itself and every object and event perceived. The phenomenal self-model therefore emerges as an instrument for intelligent information processing, certain of its own global unified properties and available as a resource for attentive perception and the initiation of behavior.

The phenomenal self is posited to be real. It "actually exists, not only as a distinct theoretical entity but something that will be empirically discovered in the future" (Metzinger, 2003). Yet the phenomenal self cannot exist on its own, as a distinct ontological entity; it is instead a "dynamic, ongoing process creating very specific representational and functional properties" (Metzinger, 2007). In this view, the self is a constructed representation of the relationship between subject and object; it exists to provide a useful interface for interacting with reality – particularly to select flexible, adaptable behaviors that promote survival of the body and cohesiveness of the self-concept.

Metzinger emphasizes that having a cognitive model of the self is a pre-condition for suffering, since suffering is a characteristic of the self. There has to be some idea that "it is *myself* who is suffering right now, it is my *own* suffering I am undergoing" (Metzinger, 2017). Suffering occurs when the conscious system exists, knows it exists, and identifies with some negative valence state. For example, if some perceived stimulus is interpreted to put the body in danger of annihilation, the consciousness arising from the neural network contained within that body is also interpreted to be in danger of annihilation. The converse is also true. If the self-model unexpectedly disintegrates, the body is in great danger of disintegrating as well, if it has not done so already – certainly, if the self does not perceive itself to be a coherent entity, it will not act to protect its coherent state. In this view, suffering is not only an indicator of self-awareness; it is also a discrepancy measure between the amount of control the system has and the amount of control the system wishes to have over its circumstances. For there to be a discrepancy, there must be a 'self' with goals and beliefs, with a knowledge of existing within the world, with some implicit or explicit belief that *one is a coherent entity capable of achieving goals through directed action*, and with the belief that existence is only worthwhile *if effortful action does achieve the desired goals*. Following this logic, Metzinger asserts that suffering can be defined by two markers: *the loss of control* and *the disintegration of the self*.

This approach is not only useful from a metaphysical standpoint, for understanding the nature of conscious awareness, but also from a moral standpoint, for examining the ethical imperatives which accompany having a self. Others have elaborated on this idea. The neuroscientist Michael Gazzaniga has posited the idea that conscious events may involve, as a necessary condition, interaction with a "self" module, a sort of "executive interpreter" in the brain (Gazzaniga, 2011). This view asserts that there can be no conscious experience of anything happening, unless it is happening *to* someone. Here, the self is the one who perceives, the one who acts. The self is therefore a morally responsible entity, even if it emerges naturally from unconscious processes. Here, the self is the sum of all inputs – from the genes inherited from the parents to the environmental influences occurring in the present moment, from the stored memory of past events to hopeful expectations for the future. The self is the one who is doing the perceiving, who is initiating the action. Gerard Edelman and others have built a framework that is compatible with this view, called Neural Darwinism, which suggests that cognitive concepts, paired with neural network activity, are simply the optimized end state of all those factors converging to most effectively guide advantageous behavior (Edelman, 2004). But these theories are challenged to explain what the self *is*, or what it is perceiving when it perceives the world.

Self-Organizing Meta-Representational Account

By purposefully placing *the self* at the center of a unitary consciousness which emerges from neural information processing, the cognitive scientist Axel Cleeremans has framed the Self-Organizing Meta-Representational Account (Cleeremans, 2019; Cleeremans et al., 2020). Here, conscious awareness is not just about being aware of some object or event in the environment, but rather being aware of being aware of that object or event. There is some value assigned to the object or event, in relation to the observer, and the observer comes to know himself by understanding his relationship to objects and events in his environment. For example, anyone can drink a glass of wine, but an expert oenologist will consciously appreciate that glass of wine much more than a naïve drinker, at the same time identifying as someone who appreciates wine and knows a lot about it. These properties are inseparable from each other. And importantly, expertise takes time; learning takes time. In this view, the brain learns to be an expert on *itself*. In taking the time to learn what it has observed about the world and what it is capable of doing in the world, the brain comes to know itself. It might learn which phenomenal states it prefers to be in, and it might learn strategies to get into these states. From unconscious processes emerge goals, and *that* is what constitutes a conscious entity. Cleeremans argues: “Having reasons for doing things is precisely what differentiates conscious agents from agents such as AlphaGo which, despite exhibiting superhuman skill when doing things, remains unable to do so for reasons of its own.”

The Self-Organizing Meta-Representational Account is built on Cleereman's radical plasticity hypothesis (Cleeremans, 2008, 2011). In this view of consciousness as a recurrent neural process, the brain not only strives to predict the effects of its own motor output on the world, but also the consequences of activity in one region of cerebral cortex on other regions. As the brain continuously and unconsciously redescribes its own activity to itself, it develops a system of meta-representation that elaborates on first-order representations. Emotional states and ingrained cognitive concepts interact with incoming sensory information, forming value propositions about the incoming sensory data. In this way, the brain learns about itself, and develops a theory about itself. This theory is constructed in relation to the environment, and objects in the environment are in turn interpreted through this cognitive model.

The process of building goals and preferences from self-knowledge is, in this theory, a process of learning to evaluate the *quality* of first-order representations (Cleeremans, 2008, 2011). Weak signals are insufficient to create a conscious percept, and may simply prompt orientation toward the stimulus to gather more data, while particularly strong signals can be handled automatically, with unconscious processes. Intermediate representations are therefore the primary target of cognitive processing. Conscious attention is required to perceive and organize a response to intermediate representations, and this is accomplished by subjecting the attended object or event to several processing loops. First, a relationship must be assessed between expectation and observation (*e.g.* noticing the previous state of an object, versus its current state). Second, a relationship must be established between self and other, to assess the relative value of what is happening to the bodily self (*e.g.* evaluating how the existence and trajectory of this object might affect me). Third, a relationship must be established between perception and action (*e.g.* modeling the expected or predicted observation if one does nothing, versus the expected or predicted observation if one takes some specific action). Putting these constructs together, the individual creates a cognitive model of the object or event in the environment, in relation to the bodily self. Any object or event that has little meaning in relation to the self-concept may therefore have little representation in consciousness.

The Self-Organizing Meta-Representational Account is built on three assumptions (Cleeremans et al., 2020). The first assumption is that neural information processing itself is unconscious. That is, consciousness depends on higher-level mechanisms rather than on the intrinsic properties of neural signaling. The second assumption is that information processing is graded, cascading in a continuous flow from posterior cortex to anterior cortex. This process allows information to rise into conscious awareness as evidence accumulates over time. The third assumption is that 'plasticity is mandatory' – with every experience leaving a trace in the brain. In this view, learning is unintentional; the construction of cognitive models is unintentional; the brain simply encodes what it learns through experience. This view somewhat leaves open the neural mechanisms by which an intentional self emerges, and the mechanisms by which the intentional self begins to mold its own cognitive models. Understanding these mechanisms may eventually help us to explain what exactly the self *is*, what it is perceiving when it perceives the world, and how exactly these intangible cognitive phenomena emerge from neural processes.

Conscious Electromagnetic Field Theory

It is now abundantly clear that cortical neural networks encode sensory information and store memories, with unique mental states being associated with unique patterns of neural activity. Yet there must be some place that houses the global workspace or the meta-representational account, a physical process or setting that allows data collected by the senses to be played out in the mind in a coordinated, cohesive fashion. Because consciousness requires the coordinated function of many different brain regions, there must be something about the very process of *integrating information together* which is key to understanding consciousness. Somehow a cohesive experience must emerge from cortical neural activity.

Aiming to address the hard problem, a theory has been put forth to explain consciousness in physical terms, as a natural emergent property of electrical events in neural networks. This approach is called Conscious Electromagnetic Information (CEMI) theory. The neurophysiologist Susan Pockett, a proponent of CEMI theory, argues that consciousness "is not material in the usually accepted sense, but neither is it some kind of non-physical spook" (Pockett, 2000). Rather, she states, it is simply "a local, brain-generated configuration of, or pattern in, the electromagnetic field." The geneticist Johnjoe McFadden, another proponent of CEMI theory, has proposed that every time a neuron fires an electrical signal, it generates a disturbance in the surrounding electromagnetic field, which creates a representation of the information contained in the neural network (McFadden, 2002).

This theoretical framework asserts that the electromagnetic field generated by the brain *unites* the information encoded in the simultaneous activity of millions of neurons, binding this information together in the time domain (McFadden, 2013). The approach asserts that extrinsic information entering through the senses must retain some truth about the nature of the represented object or scene, and this 'gestalt' must be extracted from the neural coding.

Critics of CEMI theory argue that electromagnetic fields are a common occurrence in nature and have no specific mechanism for manifesting consciousness, and furthermore, there is no good way to formally test such a hypothesis. Meanwhile, proponents of the theory argue it is a legitimate line of inquiry because it is testable and falsifiable. One key prediction of the theory, that exogenous EM fields should be able to influence neural activity, appears to be supported by a number of studies (Carrubba, 2008; Heusser, 1997). Moreover, endogenous electrical fields, produced by populations of neurons in one region of the cortex, have been shown to affect the electrochemical activity of neighboring but unconnected neurons under physiological conditions (Anastassiou, 2011; Fröhlich, 2010). Yet the main problem with this theory is that it posits electromagnetic fields produce consciousness or *are* consciousness. In this view, it is unclear what is special about a biological nervous system; given the current constraints of the theory, a standard kitchen appliance could easily be as conscious as a human being with an equivalently sized electrical field. A more nuanced explanatory framework is needed here – one that connects the distinctive features of cortical neural networks with the key features of conscious experience, one that articulates the physical mechanisms by which electrical activity generates qualia, and one that excludes the possibility of consciousness in systems without the distinctive hallmark of goal-directed behavior.

Orchestrated Objective Reduction Theory

A different approach to considering the problem of consciousness in discrete physical terms involves considering whether quantum phenomena formally contribute to neural information processing. While quantum mechanics are generally appreciated to apply only at the atomic scale or smaller, the physicist Roger Penrose has suggested the phenomena might apply to biological cells as well, since they are sensitive to events occurring at a molecular and atomic level (Penrose, 1989). Essentially, Penrose proposes the idea that electrons in biological cells are entangled in a quantum superposition state, which is so unstable that it necessarily collapses into a single reality. In this theory, the coherent particles resolve into a decoherent state spontaneously – for example, when the energy difference between neighboring matter reaches a critical threshold. This theory relies on physical processes to cause a sudden and coordinated collapse of the coherent state, and therefore is called Orchestrated Objective Reduction (Orch-OR) Theory.

This view emerged from solid work in physics. For decades, Penrose has explored how matter might cause space-time curvature and dilation starting at quantum scales, with this gravitational curvature increasing as the particles move probabilistically through time, generating entropy (Penrose, 1979, 2008). Quantum superpositions are expected to cause opposing curvatures. As this separation continues, the number of superposition states expands over some time evolution, until space itself splits and forms multiple possible worlds. But the formation of these vector spaces is expected not to be irreversible or permanent. Because the superpositions are unstable, they will self-collapse at some point under objective reduction. This collapse is proposed to be associated with consciousness; in Penrose's view, this natural process *is* consciousness. As a result, many thermodynamic systems, including the entire universe, might be conscious and capable of effecting causation.

Stuart Hameroff, an anesthesiologist, has added to the original Orch-OR hypothesis, focusing on hypothetical quantum-scale changes to the cytoskeletal structure of the neuron. Hameroff proposed that a momentary electron coherence, guiding a synchronized oscillation of water molecules within the molecular structure of the microtubule, causes an electromagnetic event that allows the quantum tunneling of electrons through the cytoskeleton upon the collapse of the coherent state (Hameroff & Penrose, 1996, 2014). Although all cells in the body have microtubules, this theory proposes that quantum decoherence processes within the cytoskeletal structure somehow manifest perceptual experience in the nervous system, but not in other organs.

The focus on cytoskeletal dynamics is quite distant from the neuroscientific understanding of the brain as a synaptic computer, with shifting electrochemical potentials based on the Hodgkin-Huxley model of integrate-and-fire neurons. But Hameroff cites an interesting feature of action potential initiation in cortical neurons, which is incompatible with the original Hodgkin-Huxley model, namely that sodium ion channels open simultaneously rather than sequentially (Naundorf et al., 2006). He argues this finding points to some hidden variable affecting neuronal firing – and further asserts that identifying that hidden variable is key to understanding how consciousness enters the equation and influences subsequent behavior. Despite this valid argument, there must be a viable mechanism for quantum oscillations to materially contribute to neuronal signaling. On this point, Orch-OR fails, because the specific predictions made by the theory have been disproven. Independent estimates suggest that timescales for quantum decoherence are whole orders of magnitude shorter than the timescales of cytoskeletal dynamics (Tegmark, 2000). Additionally, microtubule assembly processes have been demonstrated to be inconsistent with the concept of these molecules acting as quantum computational units (McKemmish et al., 2009). Yet the biggest problem with the theory in its current form is simple: there is no reason that microtubules would cause consciousness to manifest in neurons, but not in other cells of the body. After all, every cell in the body contains a similar cytoskeleton, comprised of the same types of molecular lattices, obeying the same physical laws. This theory does not explain why the brain generates consciousness but other organs do not. Furthermore, microtubules are not involved in neural information processing; signal propagation is mediated by changes in the electrochemical potential of the neural membrane. Assuming that perception, cognition, and behavioral choice are related to microtubule dynamics requires discarding absolutely everything we know about neural structure and function. This is an untenable approach, considering the proponents of the theory do not even attempt to explain how all of neuroscience is wrong, and memories are stored in microtubules rather than synaptic connections. For these reasons, Orch-OR is not widely accepted. However, the idea behind it is usefully creative and has led to much-needed experimental and theoretical work on the possibility of quantum mechanical phenomena occurring in the 'warm, wet, and noisy' environment of the brain.

Holonomic Brain Theory

Mysticism has been a persistent issue in the study of consciousness, and has long permeated efforts to study how exactly psychological phenomena emerge from physical processes in the brain. In the early twentieth century, the gestalt theorist Wolfgang Köhler and others articulated the singular and irreducible nature of perceptual awareness. Köhler was a contemporary of Maurice Merleau-Ponty; both of these thinkers strove to differentiate between external reality, the act of sensation, and the unitary, experiential nature of perception (Köhler, 1947; Merleau-Ponty, 1945). More recently, the concept of binding information together into a cohesive whole has become critical to studying the neural correlates of perception, memory formation, predictive error propagation, decision-making, and the development of context-dependent expectations (Engel & Singer, 2001; Harris & Gordon, 2015). Yet the formalization of this field was gradual. In the mid-to-late twentieth century era of scientific and cultural exploration that followed the advent of the information revolution, some ideas were posed which partially merged physics with mysticism. Although neurons were coming to be appreciated as individual binary computing units – like transistors – the concept of information as a *global binding property* in biological neural networks grew popular, particularly upon the discovery of holography (Gabor, 1948).

Holography provides a well-understood method by which discrete packets of information are bound into a cohesive whole, and this relevant prospect has been appreciated for some time. A number of researchers from various fields – including the inventor of holography Dennis Gabor, the quantum physicist David Bohm, the neuroscientist Karl Pribram, and the cognitive scientist Christopher Longuet-Higgins – considered holography a potentially useful framework for describing the unitary nature of consciousness. These researchers drew two key parallels between consciousness and holography. The first is that, like holograms, memory is a non-localizable phenomenon, although it contains information that is encoded by a physical structure (Pribram & Meade, 1999). In other words, holograms provide a good *metaphorical* analogy to the intangible nature of memory. The second interesting parallel is that performing Fourier transforms on electrical oscillations in cortical neural networks naturally reproduce the inter-spike intervals observed in these systems (Gabor, 1968; Longuet-Higgins, 1968). In other words, there is a temporal component to electrical noise that seems to convey information in the system, even contributing to the timing of signaling outcomes.

However, proponents of the idea did not specify any mechanism or process by which neural networks could physically produce holograms, and so this view remained a vague concept rather than a scientific theory. These gestalt approaches, with no clear mechanistic basis, were quickly replaced by computationally-inspired theories of parallel information processing. The latter approaches aspire to explain how complex pattern detection and decisional outputs naturally emerge from neuronal ensembles, with each neuron acting as a binary computational unit (Ackley et al., 1985; McClelland et al., 1987). This minimalist view, favored in both neuroscience and computer engineering, remains the gold standard for a parsimonious account of neural computation. And yet, this approach does not explain the qualitative, subjective, cohesive, and experiential nature of cognition; it also requires that neural networks are not structured in a completely random manner, but rather start out with some initial connectivity suited to the task at hand. The emergent properties of cortical neural networks – from phenomenal consciousness to synchronized firing patterns and spontaneous self-remodeling – are simply not well-accommodated in these reductionist classical computing theories.

The metaphorical connection posed by holonomic brain theory is intriguing, and the key concept of binding information across the time domain, as proposed by Gabor and Pribram, has proven highly useful in modern neuroscience (Buzsaki & Draguhn, 2004; Csibra et al., 2000; Engel & Singer, 2001; Harris & Gordon, 2015; Tseng et al., 2016; Whittington et al., 2010). However, the concept of 'electrical oscillations' in the 'synaptodendritic web' proposed by Karl Pribram attempted only a vague allusion to the physical properties of quantum systems; the vocabulary used for discussing the anatomical features of cortical neurons was non-standard to say the least; and none of the researchers made much effort toward reconciling the 'holonomic brain' idea with the mechanisms underpinning cortical neuron signaling. Without making any specific predictions about the relationship between neural activity and holographic information, this approach remained primarily conceptual. These many disqualifying factors may explain why this highly intuitive idea – that 'holographic information' might underlie the unitary nature of conscious experience – had trouble gaining traction in the mainstream neuroscience community. In the absence of any clear mechanism or testable hypothesis, this conceptual framework has remained mired in the realm of mysticism.

Perceptual Interface Theory

On the opposite side of the theoretical spectrum from mysticism is the prospect of eliminating the very idea of consciousness, or even all of reality itself. The latter approach has been posited by the cognitive scientist Donald Hoffman (Hoffman, 2019). This surprising conclusion arose from exploring the thorny issue of veridical versus non-veridical perception. To address the issue, Hoffman and his colleagues devised a mapping system that takes an evolutionary-fitness approach to modeling the ability of an entity to interact with reality and perceive it accurately (Hoffman & Prakash, 2014; Hoffman & Singh, 2012). This model demonstrates that organisms only perceive what it is useful to perceive; in other words, they do not perceive reality itself, but rather operate a perceptual interface that is useful for ensuring individual survival. The resulting theoretical framework is known as the Perceptual Interface Theory. It is worth considering this model and its assumptions in some detail.

The mathematical model underpinning this theory identifies the minimum number of components needed for a system to perceive the world and act upon that information, using incoming data to drive behavior which either benefits survival or risks it. The bare minimum number of components are: a world or external reality to observe, W; a perceptual map of that world, P; a space of experiences, X; an algorithm that allows the entity to choose a new action given its experiences, D; a mapping of possible actions in relation to the world, A; and the space of actions themselves, G. These six components map nicely onto a hierarchical cognitive architecture – there must be some external reality to be observed; there must be incoming sensory information which, regardless of accuracy, provides some perceptual map of the world; there must be a space of experience; there must be a decision-making algorithm; there must be a way for the decision to be implemented within the neural network; and finally, there must be some action or behavior resulting from that implementation.

The model is reasonable enough, but its implications are elaborately warped. Proponents of Interface Theory assert there is no reality, even though W is a key factor in this model (Hoffman, 2019). Furthermore, proponents of the theory assert that reality could never be perceived, even if it did exist, because there is too great a cost for computing accuracy or 'truth'. The computational model that forms the basis of this argument is very specific: it presents various resource quantities, with a fitness payoff maximized for medium amounts of the resource (e.g. oxygen), then evaluates how evolution would shape the sensory states of organisms in this condition. The sensory mapping that results from this model does represent the amount of oxygen present – with headache signifying low quantities of oxygen, unawareness of any ‘resource deficit’ at medium quantities, and dizziness emerging at high oxygen levels. Yet despite this elegant simulation, Hoffman argues that sensory systems will usually reach a binary representation, not a full distribution reflecting the true complexity of the environment. This conclusion is bizarrely at odds with his own model, and also empirically wrong. Here, Hoffman does not take into account combinatorial coding, which is common in sensory systems across the animal kingdom, and which allows biological systems to accurately encode a range of stimuli within a single modality with no priors required (Malnic et al., 1999; Solomon & Lennie, 2007; Taberner & Liberman, 2005). This semi-redundant coding strategy carries lower costs and conveys far more information than non-redundant coding, so this observed characteristic of neural systems increases both efficiency and accuracy (Chambers et al., 2019; Meister, 1996). The lack of this well-established critical feature in the computational models underpinning Perceptual Interface Theory seems to have led to the erroneous conclusion that neural networks encode fitness, not truth. Including combinatorial coding in each model of evolutionary fitness reveals that neurons encode both.

Taking this key functional characteristic of neural processing into account, it is abundantly clear that biological neural networks *do* encode data about reality and that perception *does* reflect the status of an actual external environment. Even if the organism only perceives a sliver of reality – by perceiving photons as colors, mechanical waves as sounds, average energies as temperatures – these data are useful, and in fact they are useful *because* they reflect reality. And even if the organism occasionally experiences non-veridical perception, that error can be overcome through further orientation toward a stimulus, using the sensory apparatus and the capacity for reason. Our knowledge of reality may be incomplete, but that does not mean reality does not exist at all. It is far more parsimonious to posit that there is an external world out there and consciousness is a useful computing process which evolved to help organisms navigate this reality. However, some researchers just prefer to reject reality. Others have taken the complementary approach of positing that, regardless of the existence of reality, phenomenal consciousness itself may not exist.

Multiple Drafts Model

In 1991, the philosopher Daniel Dennett published a book promisingly titled *Consciousness Explained*. Yet instead of explaining the biophysical mechanisms underlying the phenomenon, as might be expected from the title, Dennett eloquently argued that consciousness does not exist at all – that our very experience of being is an illusion maintained by an automaton (Dennett, 1991). Our selves, he claimed, are flimsy constructs which last only as long as working memory – we can hold a few salient details about the world, but there is no such thing as qualia or any lasting sense of self. This account is known as the Multiple Drafts Model, because it purposefully employs an authorship metaphor. Here, "information entering the nervous system is under continuous editorial revision" and over time, what emerges is "something rather like a narrative stream or sequence". The continuous editorial revision is achieved by "parallel, multi-track processes of interpretation" which are distributed around the brain. So, although we are certainly the authors of our actions and moral decisions, there is no experience and no actual narrator having that experience – what exists is only a chaotic flow of information.

This view has led some disappointed readers to dub the book *Consciousness Explained Away*. Even Dennett himself argues that only a theory of consciousness which explains the phenomenon in terms of unconscious processes can be valid, arguing that "to explain is to explain away" (Searle et al., 1997). One is left to wonder if Dennett has ever been in the presence of live music, or a golden retriever puppy, or if he is perhaps lacking a representational capacity the rest of us have. Yet while insisting that qualia simply do not exist, he notes that people do keep bringing it up: "The persuasive imagery of the Cartesian theater keeps coming back to haunt us – laypeople and scientists alike – even after its ghostly dualism has been denounced and exorcised." Dennett insists it is naïve to believe that experiential perception actually occurs, arguing that phenomenal consciousness is better explained in terms of how abstract cognitive processes affect behavioral output.

The philosopher John Searle has tenaciously argued against this view, pointing out that even having the conversation disproved Dennett's logic (Searle et al., 1997). What are people talking about, when they claim to have personal experiences of a qualitative nature, if these did not exist? How and why would people fake such a thing? Dennett's response is that there can be competence without comprehension (Searle et al., 1997). Our behaviors are simply the results of habits built up over time, combining to form the appearance of a cohesive self. The very act of reportability is flawed; there is no way to be certain that phenomenal experiences reported by people have actually occurred.

"I regard his view as self-refuting," contends Searle, "Because it denies the existence of the data which a theory of consciousness is supposed to explain." But that is the very point. Qualia and the reporting of phenomenal conscious experience are not considered valid data in the materialist view, *because* they are flawed, subjective, and only accessible to the person experiencing them. Researchers who subscribe to this view argue that qualia – the feeling of pain, the redness of blood, the sound of a voice, the welling of emotion, the warmth of a blanket – are just a delusion, or some trickery of neural function. Some researchers have run with this concept, arguing that consciousness is a mere illusion, sustained by attentional processes.

Attention Schema Theory

The neuroscientist Michael Graziano has dug into this idea, developing Attention Schema Theory as a way of erasing the hard problem (Webb & Graziano, 2015). Attention Schema Theory proposes that the brain constructs a model of the very process of attention, forming an 'attention schema' just like it constructs a 'body schema' to represent the bodily structure and function. Here, the mental self is a cognitive model of its own internal processes. The framework is broadly compatible with research on attentional processes, which reveals that attended stimuli have a much greater effect on memory and behavior than unattended stimuli (Wilterson et al., 2020). From an empirical standpoint, this framework is useful; separating the psychological process of *awareness* from the psychological process of *attention,* particularly with stimulus masking experiments, has allowed researchers to parse these distinguishable processes, in order to determine how each contributes to reportability and subsequent behavioral choice (Webb et al., 2016). From a theoretical standpoint, this framework is also useful, productively allowing researchers to understand how social cognition might form, with individuals modeling the attention and awareness of others (Graziano & Kastner, 2011).

Yet Attention Schema Theory does not aim to explain the phenomenal aspects of consciousness. Instead, it waves the problem away, asserting that consciousness is a process of directing attention, while streaming perceptual awareness is simply a 'mirage' sustained by the computations of the brain (Graziano, 2021). In a similar vein, the behaviorist Nick Chater argues the mind is 'flat' – that consciousness is an illusory phenomenon with no basis in reality, equivalent to the process of selective attention at best (Chater, 2018). Likewise, the philosopher Keith Frankish argues the qualitative properties of stimuli are illusions, although he allows that subjective judgements about these qualitative properties correspond to something real. In this latter view, introspection tracks *the patterns of reaction* that perceptual experiences produce. These patterns reflect the *significance* of stimuli, or the 'affordances' they offer to the organism (Frankish, 2016). This approach to cognitive neuroscience as a 'black box' underlying behavioral output does nothing to explain the undeniable phenomenon of perception that we all (presumably) experience. Instead it merely re-categorizes qualia and streaming perceptual experience as something *not worth explaining*.

Indeed, many neuroscientists today agree with a strictly reductionist view – that perceptual awareness is simply not real. It is instead a mere by-product of neural activity which requires no further elucidation. This view is in dramatic contrast to that of most laypeople, who insist they do experience the wonder of a qualitative, cohesive stream of perceptual experience. Most individuals who are naïve to the subject area would say they do have thoughts, and these thoughts are something categorically distinct from neural activity, even if thoughts do arise from neural activity. And yet, a growing number of neuroscientists and philosophers have come to believe the entire phenomenon of consciousness is an illusion.

The major roadblock that stands in the way of discovering a physical mechanism for consciousness is simple: *the very definition of thought is that it is immaterial*. If one believes this world is entirely natural, made of matter and energy, then thought *cannot* exist, being neither of these things. Because the existence of thought is incompatible with the prior belief that our world is comprised of already-defined substances, the very concept of thought must be discarded, because this is the most parsimonious approach. This widespread view asserts that phenomenal consciousness is essentially non-existent, strictly equivalent to neural processes and requiring no further explanation than these. Because this interpretation is so common, it is worth considering how far neuroscience has gotten in explaining perception, cognition, and decision-making in purely reductionist terms.

Neuroanatomical Correlates of Consciousness

The relentlessly reductionist approach to identifying the contents of the mind has deep roots in the history of neuroscience, going back nearly 130 years. From the time of Santiago Ramon y Cajal, who established the existence of neurons and worked tirelessly to catalogue their diverse anatomical features; to David Hubel and Torsten Weisel, who recorded the characteristic electrical signals of individual neurons as they responded to external stimuli in real time; to the researchers currently engaged in the modern science of connectomics, who strive to monitor mental processes through functional imaging of the brain; neuroscience has continually endeavored to explain the very stuff of thought in purely material terms. This work has helped to define the neural correlates of consciousness – the minimal structural and functional requirements for attaining a conscious percept (Crick & Koch, 1990).

Neurology has offered valuable insight, as the natural history of individuals acquiring injury inevitably provides an opportunity to study the deep correlations between anatomy, physiology, reported perceptual experience, and behavior in human beings (Kean, 2014; Ramachandran, 2011; Sacks, 1985). Yet while incidental events provide anecdotal data, systematic treatments provide systematic data. This is notable in the case of severe epilepsy, where palliative treatment may include the severing of cortical connections between hemispheres to prevent the spread of ictal activity between the two halves of the brain. This surgical intervention may sound drastic, but this treatment has a long demonstrated history of safety and efficacy, with both significant seizure reduction and an improvement in quality-of-life measures (Asadi-Pooya et al., 2008; Sperling et al., 1999). Yet some cognitive deficits occur after corpus callosotomy, generally related to either language tasks or spatial reasoning tasks that require cooperation between the two sides of the body (Franz et al., 2000). These effects can be revealed by perceptual tasks conducted after the surgery. Split-brain patients with standard left-hemisphere dominance in language ability are expected to maintain effective verbal and gestural responses to stimuli presented in the right visual half-field, but to have difficulty in responding to stimuli presented in the left visual half-field, due to the crossover of projections from the retinal field (Kita & Lausberg, 2008). However, studies which prompted these patients to verbally identify stimuli in the left visual half-field, and state their confidence level in their report, were able to respond with surprising accuracy (Pinto, Neville, et al., 2017). These results suggest that split-brain patients processing visual information in the right hemisphere are capable of accessing these data in the left (language-generating) hemisphere; as such, severing the corpus collosum may split the availability of visual percepts, disrupting the holistic nature of the visual field, but this procedure does not create two independent conscious perceivers within one brain.

Psychophysics experiments have also proven useful in probing the neural correlates of consciousness, particularly paired with functional imaging or electrophysiological recordings of key brain regions involved in perceptual tasks. Early studies identified the neurons in visual cortex which are dedicated to processing features, edges, and motion; in both humans and non-human mammals, these cells are arranged in cortical columns that exhibit a spatial mapping of the visual field (Hubel & Wiesel, 1959). Binocular rivalry tasks, in which the eyes are shown two different patterns, cause visual perceptions to become unstable, oscillating between the image presented to one eye and the image presented to the other eye (Logothetis & Schall, 1989). Usefully, this well-characterized perceptual effect is associated with an objective readout of eye movement, so experimental results are not dependent on subjective reporting, typically plagued by timing and confidence errors (Logothetis & Schall, 1990). Recording from individual cortical neurons during this task reveals these bi-stable perceptual states are associated with signaling in orientation-selective cell populations in V4, although many neurons in the visual cortex are not selective (Leopold & Logothetis, 1996). Neuronal activity in the inferior temporal cortex is less ambiguous, with many cells exhibiting firing patterns which correlate well with the perceptual dominance of one stimulus or the other (Sheinberg & Logothetis, 1997). The distinction between neurons that reflect conscious perception and those which merely encode retinal mappings is apparent in axonal projections as well as electrophysiological activity in the neuronal populations themselves. A *dorsal stream* of projections from primary visual cortex, running through occipitoparietal cortex to the *posterior* region of inferior temporal cortex, is thought to contribute to the spatial mapping of visual stimuli, while a *ventral stream* projecting from visual cortex to extrastriate cortex and the *anterior* portion of inferior temporal cortex provides information on the qualitative features of a stimulus, which are needed for the conscious processing of these stimuli (Goodale & Milner, 1992; Kravitz et al., 2013). Perceptual tasks which require attending to either the location or the color of objects in complex visual scenes allows researchers to parse the brain regions associated with encoding 'where' information in the dorsal stream and 'what' information in the ventral stream; intriguingly, the hippocampus receives multiple input streams and has the unique characteristic of encoding both 'where' and 'what' categories of information (Harel et al., 2013).

These experiments demonstrate that perceptual experience can be broken down into component processes, which are encoded in different brain networks.

The reductionist approach, taken to its logical extreme, has led some neuroscientists to posit the seat of consciousness in a specific area of the brain – perhaps one highly connected to other areas, whose function is necessary for sustaining alertness and perception (Crick & Koch, 1990). In particular, the claustrum has been suggested by Francis Crick and Christof Koch to be the defined region which organizes the operation of consciousness (Koch, 2014). Koch has pointed out structural and functional evidence for considering this region as a key neural correlate for wakeful awareness, stating: "No abrupt and specific cessation and resumption of consciousness have previously been reported, despite decades of electrically stimulating the forebrain of awake patients in the operating room. But [this region is] different. Here, consciousness as a whole appeared to be turned off and then on again." Other neuroscientists have theorized the seat of consciousness to be elsewhere in the brain. Based on studies of differential brain region activation during perceptual tasks, like the ones cited above, the prefrontal cortex and the hippocampus have been proposed to direct attention and "stream" conscious episodes across time (Herweg et al., 2016; Newman & Grace, 1999). In the words of neuroscientists James Newman and Anthony Grace: "The hippocampus is the primary recipient of inferotemporal outputs and is known to be the substrate for the consolidation of working memories to long term, episodic memories." Since memories are key to experience and the sense of self, this region might tie perceptions together and house the very phenomenon of self-awareness. And yet, identifying a region of the brain that is *required* for consciousness is not the same as identifying what consciousness *is*. There is something about cortical neural activity that is driving the manifestation of phenomenal conscious experience. Understanding this problem not only requires considering the structures of the cerebral cortex, and the connections between these structures, but also the physiological properties of cortical neural networks, and the way in which neural activity in sparsely-distributed cell populations is bound together in time to create a cohesive perceptual experience.

Neurophysiological Signatures of Consciousness

A striking feature of the mammalian central nervous system is the coordinated synchronous firing of sparsely-distributed neurons across the neural network. That is, neurons in far-flung cortical regions fire together, in phase with each other (Buzsaki & Draguhn, 2004; Engel & Singer, 2001; Harris & Gordon, 2015; Whittington et al., 2010). These regular bursts of synchronized neural activity, which spread across the cortex several times per second, are thought to be critical for the binding of perceptual experience, memory retrieval, and optimization of behavioral output (Engel & Singer, 2001; Harris & Gordon, 2015). Synchronous activity is observed at multiple frequency bands, including theta (4-8 Hz), alpha (8-12 Hz), beta (13-30 Hz), and gamma (a wide range of high-frequency oscillations, from 30 to 150 Hz). Higher-frequency oscillations can be embedded within the lower-frequency oscillations, with several of the former taking place during the longer period of the latter type (Buzsaki & Draguhn, 2004). Neurophysiological studies have demonstrated the presence of strong, synchronous global activity in conscious animals and changes in the frequency spectrum during sleep, coma, and anesthesia (Schiff et al., 2014).

Slow oscillations, particularly theta rhythm, originate from hippocampus and increase in power during the sensorimotor integration phase of task learning (Grion et al., 2016). Meanwhile, gamma frequency oscillations originate from the midbrain but spread across the thalamus and cortex, augmenting the processing of sensory information by routing gaze or auditory attention toward the stimulus (Goddard et al., 2012). As a result, oscillations are thought to underlie the unitary nature of consciousness, with synchronous activity across different brain regions binding information across the time dimension and directing output behavior (Singer & Gray, 1995; von der Malsburg, 1995). Interestingly, in support of this hypothesis, introducing additional gamma oscillations in an experimental setting does enhance the perceptual binding effect in a working memory task, in a manner that is dependent on the phase of the induced stimulus and the phase of the endogenous gamma oscillation (Tseng et al., 2016). And phase-locking between individual neurons and the network-wide oscillation is observed during free behavior (Chrobak & Buzsaki, 1998).

Some computational models demonstrate that oscillations can arise spontaneously through stochastic processes, with noise in the system contributing to the formation of resonant frequencies in coupled subnetworks and, with further recruitment, coherent oscillations across the entire system. However, this effect is primarily observed with epileptic-like spontaneous activity; normal high-frequency oscillations show a limited ability to recruit broader regions of the network, suggesting that spontaneous firing patterns may not produce the distinctive synchronous activity observed in a normal healthy brain (Stacey et al., 2011). Yet some researchers maintain that neural networks form oscillations naturally, with order arising from stochastic activity through multiple independent molecular mechanisms, and the network spontaneously ceasing further neuron recruitment before the synchronous activity generates an epileptic state (Whittington et al., 2010).

Of course, an alternative interpretation is that neural oscillations are manifestations of top-down processing, rather than bottom-up phenomena. In this vein, a theoretical model posited by Christoph Herrmann suggests that gamma activity reflects the cognitive process of matching sensory input with memory (Herrmann et al., 2004). If so, stimuli which are harmonious, or which match existing memory, should evoke a greater gamma burst compared to non-harmonious or unfamiliar stimuli. In support of this hypothesis, studies have shown stronger gamma frequency power in response to hearing words versus pseudowords (Pulvermuller et al., 1996) and in response to seeing to faces versus abstract unidentifiable shapes (Rodriguez et al., 1999). Also, unlikely events which impart novel information are associated with greater neuronal synchrony in areas of cortex associated with sensorimotor integration. For example, a shape with non-coherent visual features (such as a square plaid pattern) which nevertheless moves as a single object, will evoke greater neuronal synchrony than similar-size shapes that contain coherent visual features (Thiele & Stoner, 2003). Therefore, surprising-yet-consistent events provoke high neuronal synchrony; the resulting network-wide oscillations bind perceptual stimuli and promote the formation of memory (Zarnadze et al., 2016). In summary, gamma frequency oscillations seem to be neurophysiological correlates of perceptual awareness during a wakeful state (John, 2002). However, the mechanistic relationship between neural activity and phenomenal consciousness remains uncertain.

Unlimited Associative Learning

At the turn of the new century, Eva Jablonka and Simona Ginsburg took a new approach to thinking about consciousness as a mode of being, which could be differentiated from non-consciousness by an evolutionary transition marker (Birch et al., 2020; Ginsburg & Jablonka, 2010). The goal of this theoretical framework is to identify the evolutionary transition marker denoting the categorical existence of consciousness, similar to a transition marker denoting 'life' that separates living beings from non-living beings. It is worth delving into the comparison. The definition of life notoriously includes a laundry list of characteristics, including 1) maintenance of a thermodynamic boundary, 2) regulation of the internal milieu, 3) stability, 4) metabolism, 5) growth, 6) reproduction, and 7) death. Yet there is a commonality in these traits: the evolutionary transition marker denoting 'life' is unlimited heritability, with DNA providing information storage, the opportunity for information sharing, and an endless possibility for change (Jablonka & Szathmary, 1995). Likewise, consciousness is characterized by a laundry list of characteristics, including 1) a singular, unified perception and the ability to differentiate between percepts, 2) the integration of information over time to form memories, 3) the global accessibility of integrated information, which allows incoming data to be evaluated in light of previous experience, 4) an awareness of the self as distinct from others, 5) selective attention toward events that are salient to the self, 6) embodiment, or the feeling of what it is like to be that entity, and 7) agency, or the ability to act with intention. [The researchers do not argue these characteristics are *necessary* for consciousness, but they do argue these characteristics, in tandem, may be *sufficient* to mark out consciousness. This model rejects the existence of zombies and the concept of pan-psychism, given the lack of evidence for either; it asserts instead that consciousness is exactly what it looks like: a property of living beings.] The goal is to find a singular transition marker for consciousness which would imply all the rest, just as unlimited heritability provides a singular transition marker for life that implies all other accompanying properties.

One key characteristic of consciousness that has been noted by Jablonka and Ginsburg is the capacity for unlimited associative learning. This property is interestingly similar to the transition marker for life – it is "a within-lifetime analogue of unlimited heredity" (Ginsburg & Jablonka, 2010). While living systems can *store and share information*, generating lineages with the possibility of open-ended variation, and thereby expanding the set of possible futures, conscious systems can *store and share information*, thereby expanding the range of possible futures within a single lifetime. The common factor here is *the storing and sharing of information*, which provides a distinct evolutionary advantage. Information storage allows progress to be 'saved'. Meanwhile, information sharing allows a group of individuals – either cells or systems – to learn about their environment for far less cost than exploratory behavior (Lachmann et al., 2000). Therefore, any categorically new kind of information storage or categorically new method of information sharing will indicate a major leap in evolution. As such, cognitive memory is posited to be an inflection point in evolutionary history, equivalent to the advent of genetic information storage. [Written language may constitute a similar inflection point, allowing information storage, information sharing, and the potential for cross-pollination of ideas.]

Ginsburg and Jablonka argue that consciousness is characterized by a categorically new kind of information storage (with learned events being stored in memory) and information sharing (with learned events being shared through language). For this reason, consciousness can be described in terms of Unlimited Associative Learning. Unlimited Associative Learning is an evolutionary transition marker, characterized by the ability to make five kinds of associative links: 1) linking discriminable features across and within sensory modalities, to create compound percepts, 2) linking novel stimuli to subsequent reward or punishment, to yield first-order conditioning, 3) linking chains of associations between stimuli, to yield second-order conditioning, 4) de-linking the temporal association between conditioned and unconditioned stimuli, to escape the constraints of immediacy in learning, 5) re-writing the association after the link is devalued or extinguished, to remain open to new possibilities in a dynamic environment. Together, the features of Unlimited Associative Learning are expected to form a natural cluster of capabilities which underlie consciousness. So, the theory predicts these features of learning should co-occur, that damage to one should affect the operation of the others, and that all features of consciousness should be present in animals with these learning abilities (Birch et al., 2020). As such, this theory provides a testable theory of consciousness which ties this capacity deeply into biology.

This evolutionary approach of considering the *utility* of consciousness allows us to consider how and why it may have arisen – and what other species may share this capacity. It is a top-down approach of thinking about the problem, and it is complemented by bottom-up approaches, which explore exactly how mammalian neural networks consciously navigate the world by learning about it. Yet it is worth noting that before learning can occur, features of the environment must be detected, encoded, and bound into a cohesive stream of information. The identification and categorization of stimuli within a single sensory modality provides *something to learn about*, and as such, the visual system is a good starting point for considering the problem of consciousness.

Artificial Neural Networks

Stripping down perception to its simplest computational processes has yielded fantastic insights for both neuroscience and computing. Feature detection and feature unification in visual scenes are readily achieved by layered biological neural networks. Light signals in the environment are captured by photoreceptors in the retina and converted to electrical signals, which are then sent to cells downstream. Signals from the retina project primarily to two regions in the brain – the superior colliculus, which directs eye muscles to orient toward moving stimuli, and the lateral geniculate nucleus of the thalamus, which sends the sensory data onward to the occipital cortex. Primary visual cortex (V1) receives inputs from the thalamus and exhibits a retinotopic map, which enhances contrast but still includes blind spots (Tootell et al., 1998). Secondary visual cortex (V2) receives input from V1. These neurons may encode features of stimuli in the receptive field, exhibiting selective responsivity to size, color, shape, or orientation (Hegde & Van Essen, 2000). V3, V4, and V5 receive inputs from primary and secondary visual cortex. Neurons in V5 are tuned to the speed and direction of moving visual stimuli, and lesions in this area cause people to observe the world in a series of frames rather than continuous motion (Beckers & Zeki, 1995; Mestre et al., 1992; Paradis et al., 2000; Plant et al., 1993). Meanwhile, the fusiform gyrus of inferior temporal cortex has a notable specificity for faces (Kanwisher et al., 1997), and other regions of inferior temporal cortex contain highly selective feature-detecting neurons (dubbed 'grandmother cells') which seem to code very specific stimuli through the activation of sparse ensembles (Kreiman et al., 2000).

Biological neural networks accomplish feature detection easily. And yet, computers still have trouble identifying and categorizing novel objects, particularly if they are presented in a different orientation or illumination from previously-presented data (Faraji & Qi, 2016; Kim et al., 2018). Deeply considering the computational strategies of biological neural networks (BNNs) informed the early design of artificial neural networks (ANNs), driving advances in computer engineering. Indeed, the inclusion of hidden processing layers is a fundamental trait shared in common between biological and artificial neural networks, since this structural design dramatically improves feature detection (Zhang et al., 1990). At every stage of development, artificial neural networks have improved by more closely hewing to solutions that were first discovered by biological evolution.

The field was essentially founded by John Hopfield, who established the essential structure and operation of an artificial neural network (Hopfield, 1982). The original model called for each computational unit to remain in a binary state until it was selected for updating; very quickly, the model came to include graded representations and noise (Hopfield, 1984). Newer, convolutional neural networks incorporated recursive loops to improve feature detection through both feedforward and feed-back processes (Hopfield, 1987). Perception then improves with learning. There are two routes toward learning in ANNs: the introduction of training data, allowing the system to create an appropriate input-output map before being introduced to novel data (supervised learning) or the presentation of untagged data, allowing the system to create a lean internal representation of categories by itself (unsupervised learning). Both strategies are commonly used in the field today.

Optimal coding in an artificial neural network is achieved through the employment of an energetic cost function; the system iteratively searches for a local minimum in the cost function, and in doing so, reliably undergoes gradient descent toward a good local solution (Ackley et al., 1985). Yet overfit is a common problem; once an artificial neural network finds a solution, it does not readily exit a local minimum to explore further, the way biological organisms do when resources are sufficient. Furthermore, the recursive loops most often used in ANNs are not biologically relevant; in cortical neural networks, there is no back-propagation of errors, only feed-forward propagation. Only in the past few years has a method of equilibrium propagation been developed, which moves error signals forward in a more biologically-inspired manner but still retains the same speed and accuracy as back-propagation (Bellec et al., 2020; Scellier & Bengio, 2019). One key lesson from the design and engineering of artificial neural networks is that feature detection and categorical discrimination is not just a product of bottom-up processing – these computations require top-down processing, which improves with learning. Yet it is worth noting that all features of Unlimited Associative Learning are not met in artificial neural networks. While ANNs can integrate distinct features into a compound representation, exhibit first-order conditioning by linking a novel stimulus with a cost or reward, and make second-order associations, they do not excel at managing spatially and temporally disparate associations, tending instead toward locality. They also have difficulty exploring alternatives once a 'good-enough' solution has been found. ANNs still have much to learn from BNNs. But it is worth considering quite how much of conscious information processing can be modelled in ANNs.

Adaptive Resonance Theory

To combine the processes of feature detection, pattern recognition, categorization, attentional shift, and learning into a single, streamlined algorithm, in a manner similar to biological neural networks, Stephen Grossberg and longtime collaborator Gail Carpenter created Adaptive Resonance Theory (ART). This artificial neural network model achieves successful object identification by comparing bottom-up sensory data with a top-down memory template (Carpenter & Grossberg, 1987; Carpenter et al., 1991). The memory template either is introduced via training set (supervised learning) or emerges over time as the system collects data (unsupervised learning), with input data honing the memory template (Carpenter et al., 1991). A vigilance factor allows modules with relevant memory templates to focus attention on salient inputs, yielding a selective activation of appropriate circuitry (Carpenter & Grossberg, 1987). Whenever a 'mis-match' occurs, and a parallel search among nodes does not yield any good match, a new node is added. Whenever a 'good-enough' match occurs, a synchronous resonant state naturally emerges in the system, reflecting both top-down attentional focus and bottom-up learning of categorical fit. With this matching strategy, ART permits adaptive learning, with a certain amount of plasticity accommodating new memory formation while existing memories remain stable. ART is an artificial neural network that is run on classical von Neumann computer architecture, not a stand-alone hardware-instantiated system, like the brain. However, this layered simulation of neural network architecture, following clear rules, accomplishes many of the computational tasks previously reserved for biological entities, and as such, has been offered as a general theory of cognition, with the system *learning* to better perceive input data (Carpenter, 2001).

Grossberg and Carpenter have developed newer iterations of the original ART framework, enhancing learning efficacy and biological relevance at every step. The first iteration, ART1, self-organizes recognition categories from arbitrary sequences of binary input patterns. The second iteration, ART2, expands that capability by also handling analog inputs. ART3 incorporates a simulation of a chemical synapse, which allows inputs in higher modules to be weighted and allows the activation status of a node to be reset in a time-dependent manner (Carpenter & Grossberg, 1990). All three models have a bi-directional hierarchical architecture, much like the recursive connections in the cerebral cortex, but later iterations of ART have improved contrast enhancement and better error correction, since the competitive dynamics between nodes are sharpened during template matching. Another advance in the theoretical framework is ARTMAP, which links the success of a category assignment to the overall category size, simply by modifying the vigilance parameter in response to errors (Carpenter & Grossberg, 1991). This local feedback mechanism triggers a more effective search-and-match process, thereby distinguishing notable rare events along a spectrum of related events, in which small factors may signal drastically altered consequences. This mechanism is similar to an animal learning to distinguish slight color changes in a banana which reliably indicate the level of reward.

The early ART models classify inputs according to the set of features they possess, with each feature represented by a binary value indicating presence (1) or absence (0). Later ART models incorporate fuzzy set theory, allowing the system to classify inputs with an ambiguous set of features, or a pattern of fuzzy membership values between 0 and 1 which indicates the *extent* to which each feature is present (Carpenter et al., 1992). While previous approaches used lateral inhibition in the lower layers to suppress noisy input data, this newer method embraces noise. Here, artificial neural networks act on probability distributions, rather than real-valued weights, which allows for uncertainty to factor into category assignment – a key factor in learning about new objects while retaining previous knowledge. This approach supports supervised learning, particularly in highly variable datasets like handwriting samples and medical images (Carpenter et al., 1995). By combining all this progress into a single model, Grossberg and Carpenter were able to create FUSION-ARTMAP. FUSION has the added advantage of making multiple input-parsing channels converge on a higher-order node, with each reporting some level of confidence in their assessment, so that higher-order nodes can integrate these disparate and uncertain data (Asfour et al., 1993). By gathering multiple 'opinions', allowing evidence to accumulate, and delaying the identification of an object if confidence is low, this approach supports unsupervised learning (Carpenter & Ross, 1995). Each of these steps is critical to creating a system capable of *learning* how to better perceive input data; the systematic construction of ART models throughout the 80s and 90s carefully set the stage for the explosion of advanced artificial intelligence research over the past two decades.

One common approach within the modern field of artificial neural network design is employing a Markov chain. This method is particularly useful for modeling stochastic processes as a function of time, and it can be applied to noisy signal detection in artificial neural networks with imperfect information, to help them model the 'true' state of an external dataset (Gamerman & Lopes, 2006). A Markov chain describes a sequence of events, in which the probability of each event occurring depends on the state attained during the previous event. That cyclical process of Bayesian inference allows a neural network to change its state in discrete time steps. This is the approach taken by a newer variation of Adaptive Resonance Theory, called Temporal-Difference Fusion Architecture for Learning and Cognition (TD-FALCON). The TD-FALCON method of unsupervised reinforcement learning employs a Markov decision process (Tan et al., 2008). In a Markov decision process, the network has some distribution of possible states $s_1, ... s_n \in S$ and some distribution of possible actions $a_1, ... a_n \in A$; each action yields some distribution of possible observations $o_1, ... o_n \in O$. The likelihood of a state transition is an action-dependent process, given by the probability that action a in state s at time t will lead to state s_{t+1}, given by $P(s_{t+1}|s_t,a_t)$. The expected reward (or cost) of undergoing a state transition is given by $s_t \times a_t \rightarrow o_t$, with the observation providing feedback on the cost or reward associated with that computational cycle. The 'policy' of the network will maximize the weighted sum of expected rewards. This approach aids in unsupervised learning, by allowing the system to update its 'knowledge'. And as we will discover, the use of this kind of Bayesian inference method and the incorporation of noise into the system state better approximates the biology of a cortical neural network. But first, it is worth considering exactly how much information is contained in the clean output signals themselves – how much *meaning* can be assigned to a spike train.

Semantic Pointer Architecture

One of the truly challenging questions of consciousness studies is how neural computation gives rise to intangible concepts – that is, how neural information processing generates *meaning*. Cognition involves the use of *concepts* – from ordinary, everyday physical items to abstract, intangible ideas like love and beauty. Concepts aid in object categorization and effective communication, and they can be combined to build more complex ideas or to describe the relationships between things. Identifying the link between neural information processing and concept formation is the driving force behind Semantic Pointer Architecture, a theoretical framework devised by Chris Eliasmith. This theory holds that conceptual content is represented by the neural output signal. Semantic Pointers are "neurally instantiated, symbol-like representations that can be transformed in numerous ways to yield further representations that function to support cognitive processes like categorization, inference, and language" (Blouw et al., 2016). Semantic Pointers are not equivalent to concepts; rather, they are vectors encoded by the spiking activity of a neuronal ensemble. This theory aims to tie concepts to patterns of activity across neuronal populations, by modeling how representation manifests during the encoding process. This is notoriously difficult, since biological neural networks seem to have multiple ways of creating concepts – by setting a prototype (the best example of the category), by calling upon exemplars (typical examples of the category), or by creating a theory (rules guiding category membership). Young children tend to use prototypes, while older children incorporate exemplars in decision-making on categorization tasks (Hayes & Taplin, 1993a, 1993b). Prior knowledge reduces the reliance on exemplars, and leads to decisions that are based on general rules for assigning category membership (Harris et al., 2008; Thibaut et al., 2018). Therefore, concept formation seems to change over the course of development and learning, with mature neural networks gaining the capacity for abstraction as the cognitive toolkit grows. This increasing level of abstraction is exactly the process that Semantic Pointer Architecture was designed to capture.

The Semantic Pointer Architecture is built on the close study of biological systems and the diligent modeling of artificial neural networks (Eliasmith, 2005). The computational unit is a leaky integrate-and-fire neuron, allowing synaptic inputs to accumulate over time before a threshold membrane potential is reached, so the spiking output of the cells is a non-linear function of the input current. The local population of cells converges on the next layer, and the entire network exhibits stable attractor dynamics. A neural engineering framework, instantiated within classical computing architecture, compiles the underlying differential equations and computes connection weights across the network during operation. Like ART, selective attention is achieved through modulatory gain control of task-relevant neuronal populations (Bobier et al., 2014). Neuronal heterogeneity and noise during signal integration – hallmarks of mammalian cerebral cortex – are largely eliminated from this model, because these factors affect accuracy, speed and cost. Early models excluded all forms of dendritic non-linearity and stochasticity, on the argument that these factors were not needed to effectively model the nervous system (Eliasmith, 2005); however, later iterations have now begun to incorporate these factors (Stockel & Eliasmith, 2021). In updating the neural engineering framework, Eliasmith has shown that neuronal heterogeneity and input noise accomplish the same thing – both alter signal timing and signal amplitude, by permitting a stimulus to differentially affect the voltage state of neurons within a single layer (Hunsberger et al., 2014). The result of increasing either neuronal heterogeneity or input noise is that within-layer neurons encode different aspects of the stimulus, which allows the neuronal population as a whole to encode more information. However, it is worth noting that, if neuronal heterogeneity and input noise accomplish the same thing, it is not entirely clear why biological systems retain both features, particularly since noise is so costly.

In the Semantic Pointer Architecture, each neuron represents a vector. With a single preferred input (*e.g.* red or 650nm wavelength), the neuron displays a clear binary output signal, either firing or not, in a manner that reflects the presence or absence of the color red in the receptive field; by encoding two stimulus dimensions (*e.g.* red triggers firing and blue restrains firing), the output of the neuron traces a tuning curve across color spectrum, much like neurons in primary visual cortex (Singh & Eliasmith, 2006). This method improves feature detection, since the resulting neuronal output – the firing rate – provides far more information about the quality or intensity of the stimulus feature to the next layer of neurons. With each neuron encoding, in its firing rate, the degree to which a stimulus matches preferred qualities, the neuronal ensemble as a whole can construct something similar to a *concept* of how well the item matches each dimension of a category. Together, these vectors form a topological space, which represents objects in a symbol-like manner (Voelker et al., 2021). Semantic pointers support the construction of more complex concepts, as basic features that are sparsely-encoded in lower layers (*e.g.* four legs, flat surface, no head) are bound together to mark higher-order categories (*e.g.* a table). This architecture nicely balances generalization and specificity, allowing extrapolation of prior knowledge to new situations without loosening the category completely (Gosmann & Eliasmith, 2016). The model performs categorization experiments in an interesting way, with significantly better performance than human subjects on recognizing prototype patterns and somewhat worse performance on categorizing high-distortion images (Blouw et al., 2016).

Semantic Pointer Architecture demonstrates how much improvement can be made to artificial neural network operations by hewing to biological realities – focusing not only on structural aspects like integrative feed-forward layers of parallel core processors, but also on functional aspects like how fuzzy category membership can be encoded by sparse ensembles of computational units. What naturally emerges from this framework are concepts with uncertain category memberships. Yet it is worth noting that semantic pointers do not exist in the same way that streaming perceptual experience exists for us. Indeed, the Semantic Pointer Architecture proves that artificial neural networks can do a lot of intelligent computation without experiential perception, self-awareness, or thought-based manipulation of concepts. The extensive explanatory power of this approach comes with a clear explanatory limitation: this theoretical framework simply cannot explain why cortical neuron activity is paired with a qualitative perception of concepts. Eliasmith and his colleagues argue the goal of the theoretical framework is to exploit the functionality of the model system, rather than use it to account for neural data directly (Blouw et al., 2016). And yet, the model does account extremely well for biological neuron activity during oculomotor tasks (MacNeil & Eliasmith, 2011) and cerebellar reaching tasks (DeWolf et al., 2016). But while this architecture models the functionality of subcortical circuitry extremely well, it cannot model voluntary motor output or the cohesive stream of perceptual content which emerges from cortical neural network activity. As a result, this excellent reductionist model represents the intrinsic limits of spiking neural networks in recapitulating the fascinating emergent properties of the mammalian cerebral cortex.

Decision To Engage

Spike trains certainly convey information, both at the level of the individual neuron and relative to the phase of the broader cell population. Yet despite the relative success of synaptically-weighted integrate-and-fire models for linking neural computation with percept recognition in artificial neural networks, one key factor is missing when rate coding is the only parcel of information delivered to the next processing module. This output measure is the end result of a computation; to understand how such a clean firing pattern emerges requires investigating how exactly information is compiled in each cortical neuron during sensorimotor integration, and that means addressing the noisiness of the input signal. This leads us to probabilistic coding, where *stochastic noise* – in both the stimulus presentation and the encoding of that stimulus – yields a far more biologically accurate model of neural computation (Beck et al., 2008; Maoz et al., 2020). In this approach, neural coding is viewed not so much as 'a pattern of clean output signals' which match some categorical truth about the external world, but rather as a process of integrating multiple ambiguous upstream inputs in order to extract a signal from the noise. The neuroscientists Mike Shadlen and Roozbeh Kiani have embraced this biological reality to develop a deeper understanding of the neural codes underlying decision-making – with single-unit activity and cellular population dynamics triggering a system-wide "decision to engage" with the local environment (Shadlen & Kiani, 2013).

This approach is built on hundreds of diligent studies on the correspondence between neural activity, sensory input and motor output during perceptual tasks. A subject will observe dots moving around randomly on a screen and must decide whether the net direction of motion is to the left or the right. Since each dot is on-screen for only a short time, and the coherent motion is only a small fraction of the total movement, this difficult task requires the subject to accumulate small amounts of data in support of one hypothesis or another. Direction-selective neurons in V5 respond to motion in the receptive field, increasing their firing rate in response to stimuli with high coherence in the preferred direction. Meanwhile, associative neurons in the lateral intraparietal cortex (LIP) represent accumulated evidence for each of the choices. Neurons in this region *integrate* information from visual cortex and project to areas that control eye movement and spatial attention, making them critical components of the sensorimotor loop. These LIP neurons start out in an uncertain state, with no bias toward a decision, then gradually accumulate evidence over the course of seconds to reach a decision (Churchland et al., 2008). If the stimulus is extinguished before a decision has been reached, the decision will favor whichever direction had accumulated more evidence. The decision to move the eye in one direction or the other, to track the general direction of movement, is informed by the weight of evidence from a bounded integration of upstream signals; if incoming data from the sensory apparatus indicates a clear, coherent pattern of motion, the decision will be both speedy and accurate, but if the data are uncertain, the decision may be slow or inaccurate (Kiani et al., 2008). Confidence in the decision is affected not only by the weight of evidence, but also by the amount of time taken to make the decision; if neurons in intraparietal cortex fail to send a clear, coherent signal indicating the consensus decision of the cell population by the time the stimulus has disappeared, subjects report uncertainty in identifying the percept and difficulty in completing the task accurately (Kiani et al., 2014). Changes of mind may occur, particularly when a decision must be made in the context of noisy sensory data. The evidence accumulates over time until it reaches a criterion level, triggering a decision, and while that information is being used to initiate behavior, the original decision is either reversed or reaffirmed by additional signaling from the intraparietal cortex (Resulaj et al., 2009).

This work establishes two critical ideas. The first is that stochastic noise is a fundamental component of any realistic sensory input, and neural computation involves extracting the signal from the noise, so that contextually-appropriate behavior can be initiated. The second idea is that decisions are central to the study of consciousness; decisions are essentially the fulcrum between sensory input and motor output. This approach has led Shadlen and Kiani to develop a theory of consciousness in terms of a "Decision To Engage" (Shadlen & Kiani, 2013). In this view, decisions are a common feature to all aspects of consciousness. Wakefulness is a decision to engage the environment; perception is a decision to engage with incoming sensory input; acting with intention is a decision to author change in the world. Consciousness is therefore an ongoing narrative, correlated with neural activity and evolving as evidence is accumulated and processed. The neural network selects how much attention to give the world, decides which objects and events are worth attending to, and organizes the action of the body, in order to receive rewards or secure survival. The Decision-to-Engage Model rejects the idea of a global workspace, arguing instead that neuronal ensembles are sufficient to calculate output behaviors based on sensory input. Some regions of intraparietal cortex project to other association areas rather than motor areas, leading to cognitive complexity, such as thinking about thinking or thinking about what another individual is thinking. As attention is continuously redistributed in service of evidence accumulation, the subject can actively interrogate their environment.

This model hews to neuroanatomical and neurophysiological realities, with the brain reaching a decision to act by integrating noisy signals from multiple input sources. A decision on how to act in the world, based on incoming sensory data, is achieved by integrating evidence until sufficient information is collected or external constraints force a choice to be made (Kiani et al., 2014; Kiani et al., 2008). This approach relies on *Bayesian statistics* to model evidence accumulation in the presence of noise (Beck et al., 2008; Maoz et al., 2020). It should be noted, however, that this method is not the same as *Bayesian inference*, where the brain calculates the discrepancy between predicted events and observed events – by applying a Kalman filter to quantify and minimize uncertainty in each computational cycle, in a manner that is now commonly used in ANNs (Aitchison & Lengyel, 2017). A BNN or an ANN that is engaged in Bayesian inference explicitly encodes incoming data in the context of ongoing neural population dynamics, which are shaped by development and learning, then updates its prior beliefs as necessary, when an abundance of data requires it (Echeveste et al., 2020). These two methods rely on probabilistic coding, yet they differ in one important way. The Bayesian statistical approach allows researchers to model the neuronal firing rate during a single perceptual decision task – focusing on the state of each computational unit preceding the decision, the accumulation of stochastic "evidence" from upstream sources, and the end result of a stereotyped response to the task once a decision is made (Churchland et al., 2011). Meanwhile, the Bayesian inference approach allows prior experience to be taken into account, with a cognitive understanding of the present scenario emerging as incoming sensory data are actively compared to historical data through sampling-based probabilistic inference (Echeveste et al., 2020). This method of calculating the discrepancy between expectation and observation, using Bayesian inference methods, is known as Predictive Processing.

Predictive Processing Model

Some truly insightful models of consciousness arise from the observation that sensory perception does not occur independently of prior experience. In a particularly significant advance which builds on Higher-Order Theory, Lucia Melloni and her team showed that perception is continuously shaped by prior expectations (Melloni et al., 2011). The group also identified differential neural pathways involved in encoding 'qualitative' expectations about a stimulus versus 'existence' expectations about the stimulus, corresponding to the ventral and dorsal streams (Auksztulewicz et al., 2018). Notably, the reliance on prior expectation in sensory perception – and the divergent neural pathways encoding particular categorical expectations – were discovered in both visual and auditory modalities. Yet familiarity is complex; perception is both enhanced by previous experience and reduced by adaptation to a stimulus, with these opposing processes encoded by different cortical networks (Schwiedrzik et al., 2014; Snyder et al., 2015). In short, 'perception' seems to be the process by which prior beliefs meet present inputs (regardless of the sensory modality under investigation) and 'attention' is the process of allocating resources toward data acquisition (to reduce the difference between expectation and observation). The specific goal of scanning a scene, with intent to gather a certain kind of information, is also a kind of expectation: it will determine how the observer scans the scene, how they distribute their attention, and what kind of information they gain. As a result, both prior experience and current objectives affect the perceptual experience, as well as subsequent decision-making based on this information. And so, a particularly useful approach to considering how neural coding contributes to conscious perception has been modeling how the brain encodes information as *a divergence from expectations*. This theoretical framework is known as the Predictive Processing Model. It differs from Decision-to-Engage in that decisions are not only made from an accumulation of information, but also an estimate of the divergence of that new information from prior experience.

The Predictive Processing Model asserts that expectations, which are based on prior experience, generate stimulus templates in the deep layers of visual cortex; these templates bias how the world is perceived. Meanwhile, superficial layers of cortex encode prediction errors, thereby prompting a re-evaluation of the cognitive model or the stimulus itself. This framework has built heavily on the work of cognitive neuroscientists Peter Kok and Floris de Lange, who have demonstrated that within every primary sensory processing region, there are two groups of neurons – prediction units (P) and prediction error units (PE) – which together act to encode the difference between predictive models and sensory inputs (Kok, Jehee, et al., 2012; Kok et al., 2016). In this model, predictive coding arises from deep cortical layers, in a feed-forward manner, and prediction errors arise from more superficial layers, providing feedback. In support of this hypothesis, large BOLD signals are observed upon stimulus presentation in deep cortical layers, while omission of an expected stimulus is correlated with a larger BOLD signal in superficial layers (Kok, Jehee, et al., 2012; Kok et al., 2016).

By comparing incoming sensory data with expectations, the brain makes an energy-saving shortcut in computation. It only has to cue up the expected sequence of events in a given context, then encode the error from that cognitive model. In light of these findings, many researchers consider the brain to be a prediction engine. If predictions are inaccurate, prediction errors are generated, and the cognitive model is updated. If predictions are accurate, no changes need to be made to the encoding structure or the cognitive model, and the task can be completed unconsciously. As a result, the brain has a singular goal: to minimize prediction error (Hohwy, 2017). The Predictive Processing Model poses consciousness as a process of regulation and control, rather than a process of discovering the world as it truly is. Notably, some proponents of this model resist calling it a theory of consciousness, rather arguing that it does 'useful work for consciousness science' by addressing the properties of cortical neural networks that co-occur with consciousness (Seth & Hohwy, 2021).

Regardless of this reasonable limitation, the model naturally yields a useful, novel interpretation of the *advantage* provided by conscious information processing. That is, predictive processing naturally generates cognitive models that focus on *the underlying causes* of sensory data. First, the brain creates a hypothesis regarding the structure and operation of the world; then, it tests the hypothesis by acquiring sensory data. The prediction error can be minimized in one of two ways: either by changing the internal cognitive model in light of the prediction error, or by keeping the model constant and changing the orientation toward the stimulus so as to collect data that is compatible with the existing model (Hohwy, 2017). As noted by the cognitive scientist Jakob Hohwy, a proponent of the theory, this framework allows researchers to differentiate between the often-conflated concepts of attention and perception, with increased attention toward a stimulus enhancing any prediction error and prompting existing predictions to be updated through further data collection (Hohwy, 2012). This model is in line with experimental results in perceptual tasks, with attention effectively reversing the sensory attenuation of predicted signals (Kok, Rahnev, et al., 2012). This model also provides a good explanation for the odd phenomenon of binocular rivalry, with the error signal generated by an uncertain stimulus prompting a re-evaluation of the scene, thereby revealing a new percept; the new percept in turn generates an error signal, causing the process to repeat (Hohwy et al., 2008). And so, conscious perception can be understood as "the upshot of prediction error minimization" as the brain continually updates its expectations through perceptual inference (Hohwy, 2012).

Free Energy Principle

One key feature of predictive processing is the continuous reduction of uncertainty achieved by this neural coding strategy. According to the neuroscientist Karl Friston, this 'minimization of surprise' is the guiding force driving the improvement of predictive models (Adams et al., 2013; Friston & Kiebel, 2009). This idea is called the Free Energy Principle, and it is a general approach to understanding how a system will continuously optimize its configuration in response to its environment (Friston et al., 2006). The Free Energy Principle is the practice of applying Bayesian inference to living systems, with the accumulation of new, conflicting information prompting the revision of erroneous priors. This process allows an organism to spontaneously move toward a more ordered, adapted state over time. And just as any phenotypic changes selected by evolutionary processes are paired with changes in the genetic sequence, any changes to a cognitive model are paired with changes to the encoding neural structure, with efficiency dictating the most parsimonious coding is used.

If predictions are accurate, no changes need to be made to the encoding structure or the cognitive model. If predictions are inaccurate, the cognitive model needs to be updated, or the individual needs to take action in the world, so that incoming sensory data matches the prior expectation (Friston, 2010). Of course, ignoring new or conflicting data is an easy, cost-effective way to maintain an existing cognitive model, but this strategy has potentially large costs if the ignored data informs the individual of a survival risk. This approach therefore accommodates some variety in behavior due to individuals have differing priors. An interesting recent addition to the model posits that emotions reflect how well or how poorly the observer is managing errors relative to expectations, as a measure of the discrepancy between expectation and reality (Araya, 2018). In other words, feelings provide ongoing feedback about the reliability of predictions, and the sudden onset of emotion is an indicator that prior expectations are inaccurate in the current context. In this view, emotions provide a contingent measure of having imprecise cognitive models: they provide a feedback modality for the system to assess itself. And so, there are two ways to minimize emotional valence: either update the cognitive model, or enact change in the world to ensure the cognitive model is correct. The result of selecting the correct strategy is a combination of reduced discomfort, more accurate perception, an improved understanding of the world, and an improved ability to achieve goals through behavior.

The Free Energy Principle is a useful framework for contextualizing the Predictive Processing Model within the field of artificial neural network design. While the Decision-To-Engage Model involves accumulating evidence, by taking all data into account prior to making a decision, the Predictive Processing Model asserts that neural networks calculate the difference between expected data and actual incoming data. The Free Energy Principle argues that minimizing this error – reducing the 'surprise' of incoming data – involves a maximization of free energy, because this Bayesian inference process reduces uncertainty, or the total entropy of the system. In this way, the Predictive Processing Model and the Free Energy Principle are two sides of the same coin. Accuracy is the match between predicted and observed data; precision is given by the weight attributed to prediction errors during evidence accumulation. The act of observing the world through the lens of a cognitive model involves calculating the amount of 'surprise', which includes both accuracy – *the quantity of divergence from expectation* – and precision – *the relative importance of that divergence to the observer*. Here, surprisal is a quantitative value, given by accuracy (the expected log likelihood of an event) minus complexity (the informational divergence that emerges during Kalman filtering, adjusted for its relative weight). As the system processes information, the relevant predictive model evolves; the final state of one computational cycle (the posterior probability) defines the starting state of the next computational cycle (the prior probability). This Bayesian inference process permits some measure of ‘surprise’ to amass during the observation. An individual will act to reduce surprise, by resolving that discrepancy with a subsequent action. The more uncertainty is resolved – for example, by fitting novel data into a slightly expanded cognitive model – the more *salient* the event which prompted that adjustment.

The Free Energy Principle, as currently framed, is not a testable hypothesis but rather a general guiding law by which living systems adapt to their environment. Friston has even stated (Friston et al., 2018): “It cannot be falsified. It cannot be disproven. In fact, there’s not much you can do with it, unless you ask whether measurable systems conform to the principle.” The Free Energy Principle asserts that systems take on the most optimal system state in the current environmental context – reducing entropy, disorder, and uncertainty by maximizing predictive precision. The better a system can do that, the more accurate its cognitive models will be and the more effective the behavior resulting from those cognitive models will be. It should be noted that ‘free energy’ in this context is a statistical quantity, as opposed to a thermodynamic quantity. This concept of ‘variational free energy’ has been regularly employed in the machine learning field to solve inference problems with large probability densities, by converting them to optimization problems in search of a local minimum; the concept is helpful in modeling brain processes for the same reason (Friston, 2010). An enormously promising way forward for the predictive processing approach – and for identifying a mechanism underlying the Free Energy Principle – will involve incorporating the concept that ‘predictive value’ is an actual thermodynamic quantity (Still et al., 2012). Indeed, there appears to be a deep connection between energy expenditure and information processing. Identifying the thermodynamic laws linking cortical energy expenditure with information processing will ultimately reveal the relationship between *physically parsing information* and *generating the intangible features of consciousness*, culminating in the thermodynamically irreversible process of causation.

Entropic Brain Theory

Another theoretical framework that employs the concepts of energy and entropy, without necessarily making strict reference to thermodynamics, is Entropic Brain Theory. This framework was proposed by the neuroscientist Robin Carhart-Harris, working in collaboration with the pharmacologist David Nutt (Carhart-Harris et al., 2014). Entropic Brain Theory describes how the brain encodes useful information in the context of perturbation from the internal pharmacological state or the external environment. Under normal waking conditions, the brain is poised just below a critical point, at a transition zone between order and disorder. This threshold is called 'criticality' and in this theoretical framework, it is proposed to be crucial to normal waking consciousness (Tagliazucchi et al., 2012). It is worth noting that a far-from-equilibrium system exhibits interesting properties when energy is constantly inputted into the system – most notably a switching between transiently-stable states, a sensitivity to perturbation from the local environment, and cascading processes across the system which continuously rejuvenate the perturbation. These processes are well-described in the field of non-equilibrium thermodynamics (Glansdorff & Prigogine, 1971; Grandy, 2008; Hillert & Agren, 2006). Here, they are used as a baseline assumption for the operation of the human brain.

In this approach, 'entropy' has a non-standard definition – it is defined as the variance in network synchrony during binned time periods (Carhart-Harris et al., 2014; Tagliazucchi et al., 2012). This measure is posited to be equivalent to Shannon entropy. Notably, this measure is altered in the context of psychedelic drugs. Psilocybin, for example, increases the chaotic nature of the non-equilibrium thermodynamic system, evidenced by increased variance in BOLD signals in both hippocampus and anterior cingulate cortex (Tagliazucchi et al., 2014). LSD, interestingly, weakens the relationship between functional and anatomical connectivity, particularly in the anterior medial prefrontal cortex, with BOLD signals in the pharmacological state seemingly less constrained by anatomical connections (Luppi et al., 2021). The alterations prompted by psychedelic drugs on the functional connectome are proposed to be caused by serotonergic neuromodulation, with increases and decreases in entropy measures across the brain potentially corresponding to 5HT2A receptor activation in key pathways (Brouwer & Carhart-Harris, 2021; Herzog et al., 2020).

The Entropic Brain Theory posits that functional connectivity within and between resting state networks can be gauged by the pattern of BOLD signals, and that variability in this measure is correlated with the level of consciousness. This quantitative connectomics approach describes the changes in cortical-subcortical connectivity from wake to slow-wave sleep quite well (Mitra et al., 2015). In further support of this hypothesis, general anesthetics, like the GABA agonist/GABA-A potentiator propofol, have the opposite effect of psychedelics, triggering a state with high similarity between functional connectivity and anatomical connectivity (Tagliazucchi et al., 2016). This unconscious state, which appears to be less 'disordered' or 'entropic', is associated with lower information integration through decreased gamma synchrony (Lee et al., 2009). And yet, while the 'brain entropy' measure seems to be a relatively clear-cut measure of states of awareness – ranging from unconscious anesthetized states to normal waking consciousness to drug-induced 'higher consciousness' – EEG studies in animals have shown that complexity measures in both gamma and theta range can be dissociated from the level of consciousness in a two-step anesthesia-recovery paradigm (Pal et al., 2020). However, this finding may be a result of the differences in methodology, in which multiple drugs are onboarded and EEG is measured rather than fMRI BOLD signals. Recent studies have demonstrated significant psychological alterations with psychedelics; there must be *some* mechanism underlying this global change to the cognitive outlook (Carhart-Harris et al., 2016).

One shortfall with Entropic Brain Theory, however, is that it does not rigorously hew to the physical definitions of the thermodynamic terms it employs. The authors invoke variance as a measure of Shannon entropy – a 'dimensionless' quantity, not measured on a scale of physical units. Yet it is useful to consider how much Gibbs entropy is created by the system – a quantity of energy, measured in Joules. Indeed, energy must be expended by a physical system to create entropy or information. For this reason, some researchers argue that any discussion of entropy produced by the brain should properly be discussed in terms of the biophysical distribution of mass and energy, not in terms of "ego integrity" (Collell & Fauquet, 2015; Street, 2016). The level of description here is useful as a starting point, but a thermodynamic foundation is needed for the idea to blossom into a full mechanistic theory of consciousness. Calculating the total variance in signal amplitude in fMRI data during psilocybin use, or correlating alpha power to the extent of magical thinking during a trip, is handy in measuring the neural correlates of perceptual richness in psychedelic states, but these approaches do not provide new insight into what information is, what phenomenal consciousness is, the physical processes by which information and perceptual content are produced by the brain, or why we might be equipped with a graded conscious experience at all. In short, this theoretical framework establishes an excellent starting point for thinking about the role of chaos in cortical information processing and yields valuable data on the characteristic effects of psychedelics on cortical activity. But to push this concept further, a more biophysically-informed approach is needed – one which models the mechanistic relationship between energy expenditure, thermodynamic entropy, spontaneous state change, and cortical information processing.

Integrated Information Theory

According to the neuroscientist and psychiatrist Giulio Tononi, consciousness is the total amount of information held by the brain, integrated together (Tononi, 2004; Tononi et al., 2016). This concept addresses 'the binding problem' – the fact that we collect data about our external reality through our individual senses, and process these data in separate regions of the brain, yet we *experience* a bound, cohesive perception of reality (Engel & Singer, 2001; von der Malsburg, 1995). Tononi argues that, to generate a cohesive qualitative experience, information must be bound together, or integrated, across the neural network. Tononi's Integrated Information Theory, or IIT, takes a category theoretic approach to explaining how this integration occurs, using modal logic (Tononi, 2004; Tononi et al., 2016). This quantitative approach asserts that the amount of information held by a neural network is proportional to the total amount of consciousness experienced by that neural network, with this quantity measured in terms of ‘phi’ (Seth, 2011). Any natural or synthetic entity with computing power can potentially be conceived to have some non-zero level of phi (Balduzzi & Tononi, 2009).

Phi is a useful notion for considering the minimum structural and functional requirements for perceptual experience, self-awareness, and other mental functions within the conceptual framework of IIT (Tononi, 2012). The theory also asserts several axioms to explain consciousness. Firstly, IIT posits that consciousness is real – that we each have a personal, private experience of thought. Secondly, IIT asserts that consciousness is exclusive to the entity experiencing it – that is, it is not accessible by outsiders. Thirdly, the theory postulates that consciousness is a conceptual structure, composed of representational objects and events that form cohesive qualitative experiences. As such, it is distinguishable from other experiences, which contain other objects and other events. Fourthly, IIT contends that momentary consciousness (signified by phi) is irreducible, as individual components of the scene are unified to create a single cohesive experience. And finally, IIT asserts that qualitative experience yields a cognitive model of the cause-effect structure of reality. In short, this theory postulates that consciousness has several key characteristics: it is real, exclusive, distinctive, irreducible, and associated with the perception of cause and effect.

There are several major criticisms of IIT. Firstly, the axioms provided are ‘self-evident’, rather than proven from first principles. As a result, they may provide a descriptive account of consciousness, but no explanation of how these criteria come to be. Secondly, while the principles of existence, exclusivity, irreducibility, the establishment of causal relationships within an information set, the capacity to differentiate between objectively different phenomena, and the capacity to integrate temporally-, spatially-, or categorically-linked subjects are *necessary* conditions for consciousness, these processes are not *sufficient* to produce consciousness. There must be physical parameters that facilitate the manifestation of perceptual experience, but these physical processes or building blocks are not specified in the theory, besides stating they could be either neurons or logic gates. Thirdly, the amount of phi does not provide any plausible explanation for the unique *qualitative* nature of consciousness. The theory does not explain how individual conscious states can be distinguishable, and therefore misses something about the nature of consciousness.

Consciousness does seem to involve the integration of all the information held in the brain. However, Integrated Information Theory in its current form cannot explain how cortical neural activity generates a cohesive, streaming, perceptual experience. There is no mechanistic explanation, based on the peculiar characteristics of cortical neuron anatomy and physiology – only a proclamation that consciousness exists, is exclusively tied to a neural network, is not reducible to neural activity, is representative of reality, and implies causality. While this theoretical framework is solidly grounded in mathematical logic, and is therefore likely to be completely accurate, this framework is unable to illuminate how consciousness arises in the physical world, from biological substrates. This theory provides an excellent starting point, but a more complete explanation of the phenomenon is needed, with the generation of information by cortical neural networks explained in terms of biophysics.

Information Closure Theory

Another group of researchers, led by Acer Chang and Ryota Kanai, have made enormous progress toward developing a mechanistic theory of consciousness by focusing on the role of information in producing the perceptual experience (Chang et al., 2020). This approach, called Information Closure Theory, explores how information is generated through the interaction between an encoding system and its surrounding environment. To form a solid basis for the theory, the authors begin with a few key observations about consciousness. They note that consciousness does not consist of a one-to-one mapping from neurons to consciously accessible information, and this critical feature makes the mental representation more robust to noise. They also note that only coarse-grained information is accessible to consciousness, which ensures that consciousness operates at the organism level, not the cellular or molecular level. Both of these features suggest that consciousness is a process of reducing fine-grained information to coarse-grained information – of reducing the tenuous, uncertain details into a broader gist that captures the important, meaningful data. The overall goal of this theoretical framework is to build a model of a computational system interacting with its environment, and encoding the state of its environment, in terms of information closure.

Trivial information closure occurs when a system S has no interaction with its environment E, and remains unperturbed by the state of the environment. *Non-trivial* information closure occurs when a system S is not completely independent of its environment E – the mutual information between the current state of the environment at time *t=0* and the future state of the encoding system at time *t=1* is larger than zero – but nevertheless "the system contains within itself all the information about its own future and the self-predictive information about the environment" (Chang et al., 2020). Interestingly, two distinct processes can achieve informational closure while encoding the state of the environment. With *passive adaptation*, future system states *S(t=1)* are entirely driven by present environmental states. If sensory processes are deterministic, and the system encodes these data faithfully, the system essentially becomes a copy of another informationally closed process, and is therefore also closed. With *active modeling*, future system states *S(t=1)* evolve from the present system states, upon perturbation by the environment. In this scenario, the system collects sensory data about the state of the environment and synchronizes itself with these external dynamics, but without matching them precisely. Engaging in *passive adaptation* allows an organism to fulfil an automated task, completing the sensory-motor loop without any conscious control. Engaging in *active modeling* leads to the emergence of predictive processing, with significant advantages for an organism navigating its way through its environment. In this view, the function of consciousness is modeling both factual and counter-factual states, thereby allowing an organism to consider future actions, learn from fictional scenarios, execute action based on these internally-generated plans, and direct non-reflexive behavior (Kanai et al., 2019).

Information Closure Theory is constructed from a relatively straightforward mathematical model which maps the fine-grained representation of sensory data into a coarse-grained conscious experience. Importantly, the fine-grained details encoded by each neuron, and the molecular processes employed to achieve this encoding, *are not consciously accessible*. And conversely, those fine-grained processes have no knowledge of the bigger picture to which they contribute, because conscious experience is intrinsically a coarse-grained process. In this model, the state space of the system is the coarse-grained conscious experience, and this state space is equivalent to the product of two sets: the fine-grained processes occurring within the system itself, and the fine-grained processes occurring within the environment. This model results in five core implications: Firstly, the brain must generate information, with every conscious percept resolving some uncertainty and thereby *identifying meaning* in a dataset. In this way, consciousness may be considered functionally equivalent to *the process of parsing information*. Secondly, consciousness is a physical process, associated with a physical substrate. And so, every conscious percept corresponds to a physical event. Thirdly, consciousness is a process of encoding the state of the environment into the neural network itself. As a result, consciousness is inherently self-referential; it is the process by which a system defines its own state in relation to its environment. Fourthly, consciousness is comprised of fine-grained data, but is ignorant of the level of detail in the environment and in the system itself which prompts the coarse-grained experience of consciousness. This point arises as a direct result of the assumptions of the model, but it is a useful element of the model. Finally, the model predicts that conscious states are self-determining. Here, the *level* of consciousness corresponds to the degree of perturbance the system permits from the environment and the *contents* of consciousness correspond to both the size of the state space and the input modalities contributing to that state space. Managing the balance between internal drivers and external perturbances yields a stable, continuous conscious experience that is robust to stochasticity at the molecular and cellular level.

There are significant advantages of a theory that differentiates between contents and levels of consciousness (Overgaard & Overgaard, 2010). The *level* of consciousness is associated with graded responses to sensory stimuli and differential gamma frequency amplitudes in the states of wakeful awareness, sleep, sedation, and coma (Casali et al., 2013; Sitt et al., 2014). Meanwhile, the *contents* of consciousness are associated with the qualitative features of the stimulus being observed, with the level of detail limited by any reduction in attentional resources (Belopolsky et al., 2007; Kerlin et al., 2010). By modeling the state space of conscious experience as a combination of environmental stimuli and neural coding capacity, both of these key properties of consciousness are accommodated in Information Closure Theory. Like IIT, the theoretical framework of ICT is solidly grounded in mathematical logic, and is therefore likely to be fully accurate. Yet the theoretical framework in its current form does not explain how a state space is formed from fine-grained processes, in terms of neural structure and function. As such, this theory provides an excellent starting point for a theory of consciousness, but a more complete, mechanistic explanation of the phenomenon is needed.

Conifold Theory

Conifold Theory merges key concepts in computational physics, information thermodynamics, and neurophysiology, and in doing so, it provides a *mechanistic theory* for the emergence of conscious experience from cortical neural network activity. This biophysical process, which relies on probabilistic coding, also achieves exascale computation with high energy efficiency and local memory storage, thereby resolving some key properties of cortical neural networks with the cognitive features that seem to uniquely arise from these systems. Here, probabilistic coding within cortical neural networks naturally yields 1) a cohesive stream of qualitative information content, exclusively accessed by the encoding structure and continuously updated with incoming sensory data; 2) the construction of predictive models of the world *and* the self, based on lived experience; and 3) non-deterministic computational outcomes, built on incoming sensory data and previous experience in that context, which allow the organism to effectively navigate its local environment. This theoretical framework does not focus on describing information processing at the level of whole brain regions, but rather the encoding of upstream signals and electrical noise in cortical neurons.

To understand how cortical neural networks produce qualitative information content, it is useful to model the unique physiological properties of cortical neurons. All neurons pump positively-charged sodium ions outside the cell to attain a negative voltage across the cell membrane. The binding of excitatory neurotransmitters at chemical synapses permits the inward flow of sodium ions, locally increasing the membrane potential (Armstrong & Hille, 1998). If a cortical neuron receives multiple coincident signals within a short spatiotemporal window, and reaches its voltage threshold prior to rectification of the membrane potential, the cell sends a signal – opening voltage-gated ion channels, firing an action potential, and releasing neurotransmitter to post-synaptic neurons (Magee, 2000). Yet signaling outcomes in cortical neurons are not deterministic, like invertebrate neural circuits or spinal reflex circuits, with a single suprathreshold stimulus triggering the action potential (Bialek & Rieke, 1992; Powers & Binder, 1995). Instead, both upstream signals and random electrical noise contribute to signaling outcomes in cortical neurons (Softky & Koch, 1993). Stochastic events are a critical component of the combined input signal, with spontaneous membrane potential fluctuations contributing to the likelihood of the cell firing an action potential (Dorval & White, 2005; Stern et al., 1997). Cortical neurons even maintain a coordinated 'up-state', remaining sensitive to random noise in gating signaling outcomes; this coordinated cellular activity results in the synchronous firing or 'event-related potentials' which accompany the realization of a multi-sensory percept. While simpler neural circuits are robust to noise, cortical neurons allow noise to gate signaling outcomes, and it is impossible to model cortical neural activity without incorporating electrical noise (Faisal et al., 2008; Maoz et al., 2020; Ostojic et al., 2009; Roxin et al., 2011; Stacey et al., 2011).

Random electrical noise reduces both accuracy and energy efficiency in classical circuits. But rather than being energy-inefficient and prone to error, cortical neural circuits direct exceptionally fine motor output and are extraordinarily energy-efficient. Indeed, cortical grey matter exhibits nearly perfect energy efficiency (Engl & Attwell, 2015; Howarth et al., 2012; Zhu et al., 2012), while peripheral neurons and invertebrate neurons are far less energy-efficient (Sengupta et al., 2010).

Cortical neurons also undergo spontaneous remodeling of the physical structure into a more ordered system state, during both development and learning (Hebb, 1949; Jackson et al., 2006; Zarnadze et al., 2016). These two features – extraordinary energy efficiency and spontaneous self-remodeling – are uniquely characteristic of cortical neural networks. Since these two features – which are critical to cortical computation – are not compatible with classical assumptions, these physical processes must be described in terms of non-equilibrium thermodynamics and non-classical computational physics. Conifold Theory aims to do just that.

Conifold Theory discards the classical assumption that cortical neurons are binary computing units, *always in an off-state or an on-state*. Here, cortical neurons compute *the probability of switching* from an off-state to an on-state. This model focuses on the cortical up-state, in which the entire network of neurons resides near action potential threshold, allowing electrical noise to drive synchronous signaling outcomes (Haider et al., 2006). This approach takes into account both upstream signals and electrical noise in prompting the action potential. Here, the voltage state of each neuron is modeled as the mixed sum of all component microstates, or the amount of information that is physically encoded by each computational unit. That quantity of information then compressed, as *consistency* or *meaning* or *predictive value* is extracted from disorder and uncertainty. This model is equivalent to stating that an optimal system state is extracted from a broad probability distribution, as some non-deterministic computational outcome is achieved across the neural network.

The cyclical process of information generation and compression achieved by probabilistic coding can be modeled with wave mechanics (Stoll, a). In this model, the state of each electron is described as some set of probability amplitudes distributed across the *x, y, z,* and *time* axes. The macrostate of the computational unit is given by the membrane potential, with "information" being defined as the mixed sum of all component microstates, or the entire distribution of probability amplitudes. These component pure states are captured by the neuron, which physically encodes the likely state of each local electron into its membrane potential. Here, the polymer membrane structure acts as an ideal holographic recording surface, with the constructive and destructive interference of these complex-valued probability amplitudes *physically encoding information*. Since the probabilistic position and momentum of each electron can contribute to the probabilistic voltage state of multiple neurons, the information content is integrated. In this model, the information that is physically encoded by cortical neurons produces a holographic projection of information content. This bound information content is limited by the range and sensitivity of the sensory apparatus, and is continually updated as new sensory data is encoded in the cortical neural network.

The cyclical process of information generation and compression achieved by probabilistic coding can also be modeled with Hamiltonian mechanics (Stoll, b). The state of each electron in the system is intrinsically uncertain (Born et al., 1926). The *integration* of all possible component particle states creates a cohesive probability density, which is the amount of *information* or *von Neumann entropy* held by the system. The uncertainty is resolved by taking the *derivative* of the Hamiltonian, thereby reducing that probability density into a single reality. This computation yields a non-deterministic outcome for the system, with the position, momentum, and energy state of every electron in the system being defined at a single point in time. But critically, defining the position of each electron causes a shift in the charge distribution across each atom (Esteve et al., 2010; Feynman, 1939). This dipole moment boosts angular momentum, prompting new interactions with other atoms. As a result, any reduction in uncertainty is paired with a movement of ions. In cortical neurons, these fluctuations drive signaling outcomes. And once each component state is defined, those assigned values become the past, and a new probabilistic system state begins to evolve, in accordance with the von Neumann projection postulate (von Neumann, 1932). As a result, the entire computational process repeats, in a cyclical manner, with each non-deterministic computation affecting causation within the system.

The cyclical process of information generation and compression achieved by probabilistic coding can also be modeled with matrix mechanics (Stoll, c). As the non-equilibrium thermodynamic system is perturbed by its environment, the number of possible system states expands to create information, or von Neumann entropy. The system *traps heat to drive computational work*. In accordance with the first law of thermodynamics and the Landauer principle, free energy must be expended to create information and free energy is recovered as information is compressed (Berut et al., 2012; Jun et al., 2014; Landauer, 1961; Yan et al., 2018). Information is compressed by identifying consistencies, or *extracting predictive value* (Still et al., 2012). Here, information is the mixed sum of all component pure states, described by a density matrix. As the encoding thermodynamic system 'A' interacts with its surrounding environment, thermodynamic system 'B', the combined density matrix undergoes a unitary change of basis, and System A identifies a compatible state with its environment, System B, by physically encoding the state of the surrounding environment into its own state. The amount of predictive value that is extracted from that quantity of information during the unitary change of basis is *equivalent* to the amount of entropy reduced and the amount of free energy released during the computation. This process allows the system to acquire *meaning* from a complex dataset, by extracting a signal from the noise. And so, information compression is always paired with the acquisition of predictive value and the release of thermal energy, which is then available to do work. That subsequent work may involve remodeling the system, to encode that predictive value and thermodynamically favor that particular pattern of neural activity to re-occur in similar contexts. A single computational cycle resolves the neural network state in the present moment and yields a semantic statement about the external environment, which is then validated by orienting toward the stimulus in question. Individual semantical statements are held in working memory, then integrated with subsequent neural network states to compute predicted syntactical relationships between events. In this way, the system accumulates sensory data (semantical statements about the world) into cognitive cause-and-effect models (syntactical statements about the structure and operation of the world) by parsing predictive value. Yet information content which may have accurately reflected reality in some context could simply be inaccurate in another context. The ability of the system to recognize the difference is largely determined by how effectively the neural network structure has been remodeled to encode previous information. If the neural network remains adaptable, not completely ordered, more nuanced complexities can be noted and more predictive value can be gained, at some energetic expense.

This process of non-deterministic computation supports an exploration/exploitation dynamic. Any information that is acquired will be parsed for predictive value; the more predictive value contained in the information set, the more free energy is released and made available to do work. Therefore, it is in the interest of the organism to explore its environment, collecting information and extracting predictive value. As the system remodels itself into a more ordered state to encode that predictive value, disorder is reduced, less entropy is produced, and more free energy is available to the system. Therefore, it is also in the interest of the organism to store information gleaned from the environment in a way that thermodynamically favors patterns of activity which have proven useful in the past. As such, it is predicted that thermodynamic computing systems both *reduce* information entropy to maximize efficiency (by favoring previous patterns of neural activity, whenever the environment is familiar and navigable) and *increase* information entropy to maximize potential predictive value (by engaging in exploratory behavior, whenever the context is novel). In combination, this rival energetic cost function leads a cortical neural network to parse the predictive value of incoming data, encode that predictive value, and self-organize into a more ordered state in order to store the information acquired through exploratory behavior. In summary, Conifold Theory asserts that consciousness is a process of neural computation – with information encoded, represented, and parsed to select the most optimal behavior in the present context.

The explanatory power and neuroscientific predictions of Conifold Theory

The three mathematical models, relying on the laws of holography, mechanics, and thermodynamics, each independently demonstrate a cyclical process of information generation and compression. This process is a natural outcome of probabilistic coding, which is observed in cortical neural networks.

As predictive value is extracted from a system-wide probability distribution, the heat-trapping information-encoding system finds an unlikely but compatible state with its local environment, through a mechanistic process of non-deterministic computation. The system then encodes what was learned into its own structure through spontaneous self-remodeling, to favor that pattern of activity to repeat in similar contexts. The resulting theoretical framework has superior explanatory power (Stoll, *in prep-d*). It describes how a cohesive stream of qualitative information content and non-deterministic behavioral outcomes arise from cortical neural networks, but not spinal circuits. It also accounts for the extraordinary energy efficiency and exascale computational power of the cerebral cortex, as well as local memory storage achieved through spontaneous remodeling. This approach, which asserts a deep connection between neurophysiology and information processing, offers the first plausible physical explanation for consciousness as an emergent phenomenon of neural activity.

Conifold Theory makes a number of specific predictions for objective, measurable outcomes which could be tested in the laboratory, to either disprove the theoretical framework or provide additional support for it (Stoll, *in prep -e*). These predictions include estimates for coulomb scattering of sodium ions at the cellular membrane, as well as estimated timescales of sodium ion decoherence. One key prediction ties a novel measurable outcome to neurophysiology: due to the conservation principle, thermal free energy must be released upon information compression. This spontaneous photon emission, characterized by wavelengths in the infrared range, should correlate with event-related potentials in the cerebral cortex during perceptual tasks and should *not* correlate with ictal activity during seizures, when ion channel dysfunction (rather than information processing) leads to highly synchronized neural activity across the cerebral cortex.

The free energy release caused by information compression is predicted to trigger van der Waals interactions between sodium ions and the non-polar lipid membrane of nearby neurons. This force is expected to be observed in cortical neurons directly prior to firing, but not in spinal or peripheral neurons prior to firing. In other words, the result of non-deterministic information processing should affect the neuronal state, prompting a spontaneous action potential. Because this event is expected to reflect a system-wide computation, information compression is expected to result in synchronous firing of sparsely distributed neurons across the network. Such synchronous firing does occur in biological neural networks, and is not well-explained by other theories (Buzsaki & Draguhn, 2004; Stacey et al., 2011). This synchronous firing, in the gamma frequency range, is known to correlate with percept recognition (Csibra et al., 2000; Engel & Singer, 2001; Forseth et al., 2020; Hagiwara et al., 2010). Yet until now, it has not been clear how this neurophysiological signature mechanistically corresponds to conscious information processing.

The result of free energy release into the system is an increase in the availability of energy to do work in the system. As a result, Conifold Theory explains how a cortical neural network could be more energy-efficient than permitted under classical assumptions. Because this energy is directed toward the neurons and individual synapses which experienced the greatest information compression, these neurons and synapses will have the greatest amount of free energy newly available to do work. And because this energy is abruptly directed toward synapses with the greatest reduction in uncertainty, those particular synapses will have the most amount of free energy available for structural remodeling. As a result, neurons which have just 'realized' predictive value should be most active in remodeling the synaptic structure to thermodynamically favor that pattern of activity to repeat in a similar context. This phenomenon has also been observed in cortical neurons, which spontaneously remodel into a more ordered state in an activity-dependent manner (Hebb, 1949; Zarnadze et al., 2016).

Shorter computational cycles, manifesting in gamma frequency events, should prompt compression of semantical information, while longer computational cycles, manifesting in theta frequency events, should permit compression of syntactical information. As a result, quiet restful states characterized by theta rhythm (including sleep) should be associated with the construction or reinforcement of cognitive models of cause-and-effect relationships between events, rather than factual knowledge. This prediction also has strong evidence in the existing literature (Mitra et al., 2015; Steriade et al., 2001). Conifold Theory predicts this memory consolidation should be inhibited by photon absorption at slower frequencies.

In Conifold Theory, the ability to identify truths about the world is an essentially corrective process, which is relatively robust to starting conditions. An abundance of evidence overturning previous beliefs should prompt updating of the cognitive model through plasticity mechanisms, particularly if survival of the organism depends on doing so. However, this process requires *work*, so this option may be rejected, particularly in individuals who are unskilled at regularly updating their cognitive models. The computational process should also be robust to the loss of a kernel, or computational unit, since there is much redundant coding in the system. The overall probability density and the energy efficiency of the system are the key factors in quantifying the amount of content held in conscious awareness and the capacity to affect change, not the number of individual computational units. Indeed, the content of the holographic projection is expected to be reliant on several factors: the recent trajectory of each sodium ion, which contributes to the signaling outcome of each nearby neuron; the kinetics of each ion channel; neuronal morphology and lipid membrane permeability; and notably, the physical location of the individual in space and time, which makes certain observations possible, as well as the ability for the individual to notice these sensory stimuli, given their contextual expectations and the amount of energy they are devoting to attending a stimulus.

In Conifold Theory, thermoregulatory control is a critical requirement of conscious information processing. For this reason, higher temperatures – such as those generated by fever – should lead to an increase in probable states for the neural network, correlated with an increase in richness and diversity in the information content. A raised temperature in the central nervous system is therefore predicted to cause increased vividness in perception, with external heat adding noise to the system, resulting in increased error rate and fatigue. Conversely, a lowered temperature in the central nervous system should lead to a decrease in perceptual richness, reduced memory formation, and less ability to direct voluntary movement – simply because there is less energy available for these computational tasks. Therefore, older individuals with a reduced metabolic efficiency, who have trouble maintaining brain temperature, are predicted to benefit from warmer ambient temperatures. This straightforward intervention should help to promote neural plasticity and stave off dementia.

In this theoretical framework, probabilistic coding by cortical neurons naturally generates a holographic projection of information content, corresponding to the quality and quantity of data received by the sensory apparatus. This process is utterly dependent on ion channel function, with perceptual experience expected to be impaired by any drugs or genetic conditions which reduce the probabilistic nature of cortical neuron signaling. For this reason, either increasing *excitation*, through potentiation of glutamatergic signaling, or increasing *inhibition,* through potentiation of GABAergic signaling, should lead to reduced perceptual content and diminished motor control. These effects are known to occur. Pharmacological or genetic glutamatergic over-excitation leads to the stereotyped movement and ictal activity characteristic of an epileptic fit, paired with simple, sharp, and inaccurate perceptual content (Hanada, 2020; Parrish et al., 2019). Benzodiazepine-, barbiturate-, or alcohol-induced GABA signaling leads to reduced perceptual content and impaired motor control (Campbell et al., 2014; Olsen et al., 1986). By contrast, any pharmacological or genetic intervention that *increases* the probabilistic nature of cortical neuron signaling, by requiring a greater number of EPSCs to reach action potential threshold, is predicted to increase the richness of perceptual content.

Reconciling previously-proposed theoretical frameworks with Conifold Theory

Conifold Theory asserts that consciousness can be fully explained in terms of physical processes. In this view, the mind is an emergent property of a neural network that encodes information. Our minds, like our bodies, are natural products of evolutionary processes, following the laws of mechanics and thermodynamics. Consciousness allows us to perceive reality, store the information gained through observation, and plan effective actions which aid survival. In this view, consciousness is a useful computational process. It is made possible by the unique anatomical and physiological characteristics of cortical neural networks, which parse incoming sensory information for predictive value and maximize energy efficiency by identifying a compatible state with the surrounding environment. Over time, this physical computational process remodels the encoding structure into a more ordered state, thermodynamically favoring previously-learned patterns of activity. The encoding structure is paired with information content, providing an easily-accessible representation of reality which helps us to navigate effectively. In short, consciousness allows us to perceive our world, grow a better understanding of our world, and choose our own actions in the world.

In this view, phenomenal consciousness is real and worth explaining, but is not an epiphenomenon. The process of acquiring data from the local environment and encoding these data into the neural network drives a categorically new emergent property – the representation of bound information content, reflecting the state of external reality. In Conifold Theory, the holographic projection of information content is *not* equivalent to reality itself, nor is it equivalent to neural coding; it is something categorically separate, worth explaining on its own terms. That continuously-updated stream of perceptual content, bound across sensory modalities, is a representation of the information encoded by the neural network, indicative of a fundamentally different kind of computation than that observed in classical systems. In this inherently probabilistic form of thermodynamic computing, representation is part of the procedure – it provides a state space to perceive the information being processed and a method for the system itself to actively participate in information compression.

Conifold Theory agrees with the first-order theorists, that perceptual awareness is possible without higher-order access or reflection, and this type of phenomenal consciousness may occur in children and animals. Indeed, Conifold Theory argues that perceptual experiences accumulate over time to form higher-order cognitive models. As such, Conifold Theory allows for naïve perception, with these percepts being temporally sequenced into cognitive models of the world. The emergence of cognitive models, occurring in tandem with the development of prefrontal cortical function, provides a 'lens' through which perceptual data can be accessed, thereby sharpening percepts and speeding up categorization tasks. In this view, cognitive constructs make previously-experienced percepts and behaviors easier to call into access, because synaptic remodeling has thermodynamically favored the associated patterns of activity. Conversely, existing cognitive constructs may reduce the ability of an individual to accept new or conflicting information, because the associated patterns of neural activity are thermodynamically unfavored. In this view, existing cognitive models fit a lens onto raw percepts, for better or for worse. The theory is therefore compatible with Higher-Order Theories.

Conifold Theory is certainly compatible with Global Workspace Theory, the concept of the neuronal avalanche, and the idea that incoming sensory information is interpreted through the lens of existent internal representations; it does not oppose these cognitive descriptions of consciousness. Rather, this new perspective expands on the conceptual framework of the Global Workspace to detail exactly how a cognitive structure might physically arise from neural activity. For example, the "theater stage" described in Global Workspace could be construed as equivalent to the holographic projection described by Conifold Theory, where perceptual inputs merge together and play out over time. Meanwhile the "spotlight of attention" proposed in Global Workspace Theory may be equivalent to the quantity of energy expended to orient toward a stimulus. These independently-derived theories may not represent completely different frameworks for understanding consciousness, but mere linguistic differences in discussing what consciousness is – either a cognitive event or a physical process of thermodynamic computation. It should be said, Conifold Theory is also enormously compatible with Recurrent Processing Theory, in recognizing that widespread signaling events across the cerebral cortex are required for conscious experience to occur. Resolving the role of feedforward processes described by Global Workspace Theory, the role of recursive cortico-cortical connectivity highlighted by Recurrent Processing Theory, and the molecular mechanisms described by Conifold Theory, should prove useful. Understanding these systems-level processes in terms of the critical timing coordination facilitated by cortical up-states (which coordinate the signaling threshold of neurons at the network level) and ion channel kinetics (which coordinate the electrochemical potential of neurons at the cellular level) should prove a worthwhile approach. Both the pruning of neural connections during learning and the active signaling mechanisms underpinning a cortical up-state are expected to play a role in achieving the efficient coordination of timing needed for a unitary conscious experience.

Conifold Theory also fits well with the Somatic Marker Hypothesis. Neuroscience certainly accommodates proprioceptive cues contributing to decision-making as well as sensory cues arising from the local environment. It is likely that the truth lies in the middle ground, that our mental states are driven by a mix of internal and external cues. Indeed, information about the internal state of the body arrives in the brain just like any other sensory modality and is similarly incorporated into the activity of sparsely-distributed cells across neural network. And just like any other sensory modality, incoming data can be ignored or rejected or fixated upon, depending on how well prepared the brain is to deal with incoming data in the moment.

Conifold Theory essentially agrees with the Self-Organizing Meta-Representational Account in that neural information processing itself is unconscious. The neural membrane is not ‘choosing’ to receive signals. It just so happens that probabilistic coding mechanisms give rise to a cohesive stream of information content; it is this emergent property which is equivalent to perceptual consciousness. Both theories agree that strong incoming signals can be handled automatically, without any need for much attention, while weak signals are either amplified or discarded in the course of processing; intermediate signals therefore provide the thrust of conscious attention. Intermediate signals are parsed for predictive value, in light of subsequent events, and used to construct cognitive models, which in turn provide valuable input in future situations. In SOMA, this time-dependent chain of events is solely a cognitive process; in Conifold Theory, it is a physical process of thermodynamic computing. Conifold Theory is certainly compatible with SOMA in considering the ‘self’ as the brain learning about itself by modeling its own changes over time, the relationship between self and other, and the relationship between perception and action. However, Conifold Theory differs from SOMA in explaining exactly how streaming perceptual awareness naturally emerges from neural computation and how conscious information processing drives the formation of cognitive models and the selection of behavior. In short, Conifold Theory asserts that consciousness is an active process of distributing energy towards attention, cognitive modeling, and initiation of action, while SOMA presents these processes as passive events which happen *to* the system.

Conifold Theory is certainly compatible with the view of the self as a central cognitive construct; this new theoretical framework only aims to explain the physical processes by which the self-model and other cognitive constructs are formed. This view complements Metzinger's Phenomenal Self Model by explaining how the self-model is constructed by the physical distribution of matter, energy, and information across time. This work reinforces the key idea of a self being something capable of suffering, which is central to Metzinger's philosophy. Indeed, a real existence that is threatened by the risk of energy dissipation (obliteration, or the erasure of thermodynamic work) means that suffering itself is real. The existence of a self that is capable of acting in the world, but is prevented from doing so, experiences a perceived obliteration of structural integrity or capacity for agency. Therefore, in complete agreement with Metzinger's philosophy, the two primary markers of suffering are *the loss of control* and *the disintegration of the self* (Metzinger, 2017). There is a profound implication here. If the self and the capacity for suffering are both physically real, then there is meaning or predictive value to be gleaned from *being* and from *suffering*. With an awareness of existence (metaphysical understanding), and a cognitive model of this conscious existence as a tool for gaining information about the world (epistemological understanding), there arises a certain responsibility for considering what we should do in this world, given the abilities we have. A *moral* imperative arises. The sensory-motor loop must be completed – because that is what we exist to do. It is a waste of energy to perceive ourselves and our capacity to suffer, and to grow a cognitive model of how this physical process works, without using this knowledge to fuel a conscious decision on *how* to act. By this reasoning, Conifold Theory resolves not only some metaphysical issues raised by existentialist philosophers, but also a central moral problem raised by the fact of conscious existence. If we are thermodynamic computers, and we are capable of both exploring our world and exploiting the knowledge gained, and we understand all conscious beings to suffer when this goal is thwarted, then surely we should use this knowledge to choose our actions. Only in doing so can we alleviate our existential anxiety and find some purpose in our existence.

Conifold Theory presents consciousness as a natural emergent property of neural activity. But it provides a far more nuanced explanatory framework than CEMI theory – it connects the distinctive features of cortical neural networks with the key features of conscious experience, it demonstrates how probabilistic coding in cortical neural networks gives rise to both perceptual experience and non-deterministic behavioral outcomes, and it usefully excludes the possibility of consciousness in systems without the notable behavioral indicator of spontaneous goal-directed action. Simply put, consciousness does have something to do with the electrical processes taking place in the cerebral cortex, but the best theory is the one that details these processes in an accurate, mechanistic manner. It should be said, however, that a field must start out with a vague idea of the key components needed to achieve a phenomenon before one can clearly articulate the specific mechanisms involved.

Conifold Theory agrees with Orch-OR theory that quantum physics is the most appropriate level at which to study neural activity, rather than holding onto untenable classical assumptions that cannot account for the emergent properties of cortical neural networks. Conifold Theory and Orch-OR have a compatible view of a geometrical vector space emerging from quantum mechanical interactions occurring within neurons, which is spontaneously resolved through solely physical processes, in a periodic cycle. However, Conifold Theory asserts that any model of neural computation should focus on the neural membrane, rather than on the cytoskeleton. Cytoskeletal dynamics, from the binding of calcium to the placement of tau on the microtubule lattice to the transport of kinesin and dynein, are relevant to every single cell in the body, and therefore cannot possibly explain how consciousness emerges from the brain rather than any other organ. To be viable, any theory of consciousness must relate the physical processes of encoding, representing, and parsing information to the anatomical and physiological characteristics of cortical neurons. Simply put, consciousness is not an emergent property of microtubule structures; it is an emergent property of probabilistic coding by macro-scale computational units whose membrane voltage encodes the state of the local environment.

Conifold Theory presents consciousness as a natural emergent property of cortical neuron structure and function, made possible by the probabilistic coding of macro-scale computational units. This model finds itself largely in agreement with Holonomic Brain Theory, elaborating and formalizing the previous effort into a far more rigorous framework. Early proponents of Holonomic Brain Theory demonstrated that taking a Fourier transform of electrical noise in cortical neurons naturally reproduces the inter-spike interval, proving that noise materially contributes to signaling outcomes in the cerebral cortex (Gabor, 1968; Longuet-Higgins, 1968). Recent studies continue to demonstrate that electrical noise is a significant contributor to cortical firing patterns (Dorval & White, 2005; Echeveste et al., 2020; Maoz et al., 2020; Stern et al., 1997). As a result, cortical neuron activity cannot be deterministic. Conifold Theory agrees with the temporal component of Holonomic Theory, but formalizes the theory by adding a spatial component as well. In other words, holography is not just a hand-waving metaphor for the representation of information in cortical neurons, it is an actual physical process. In Conifold Theory, information is captured by the polymer membrane surface of the neural network, and holographically projected into a cohesive stream of information content, which is continually updated with incoming data. And so, Conifold Theory goes a full step further than Holonomic Brain Theory, arguing that perceptual awareness is truly holographic – and that perceivable information content is a natural emergent property of neural computation, which allows the organism to perceive its environment and navigate that environment effectively.

Conifold Theory asserts that reality exists, and consciousness exists to perceive that reality and navigate within it. For this reason, Conifold Theory is in strict opposition to the views espoused by the proponents of Interface Theory. Interface Theory asserts that objective reality may not even exist at all, and if it does, we cannot perceive it in any accurate sense. Oddly enough, the mathematical models underlying this theoretical framework do not even support its conclusions. Conifold Theory is in agreement with neuroscientific findings that combinatorial coding and elaborate processing of sensory data permits an organism to approximate the state of the environment, such as the location and certain qualitative features of an object. It is not only the survival value of the object, in relation to the organism, that is observed; the actual features of the object are observed and its general characteristics can be agreed between individuals. In Conifold Theory, examples of non-veridical perception are theorized to offer evidence of a trade-off between accurate representation of the current situation and accurate construction of a longer-term cognitive model. In other words, developing a general understanding that is robust to anomalous data may occasionally require obscuring the fine details of a scene. Thus, non-veridical perception does not imply there is no reality. Certainly, our perceptual abilities may be flawed and subject to thermodynamic constraints, but they are still incredibly valuable in helping us navigate our world. If Interface Theory is correct, then we have no way of knowing what is true and we may as well give up on the scientific process completely. Alternatively, we can set aside this epistemological despair and put our brains to work, focusing on using our senses and our capacity for reason to bypass our cognitive blind-spots, in order to better understand our world.

Conifold Theory asserts that consciousness exists, and for this reason is dramatically in contrast to both the Multiple Drafts Model and Attention Schema Theory. These illusionist approaches contend that consciousness is merely a process of parsing data, or distributing attention to incoming data; in either case, qualia are not considered ‘real’. Meanwhile, Conifold Theory contends that consciousness is a physical process of neural computation, and qualia are a physical representation of information content encoded by the system. In Conifold Theory, qualia are a critical mechanistic component of the computational process, not some dodge or illusion. It should be said that sometimes, when people have no explanation for something, it is easier to assert that thing does not exist. A theoretical framework which offers a detailed mechanistic explanation for streaming perceptual experience, in terms of the physical processes of neural computation, may go some way toward relieving that discomforting state of uncertainty, thereby allowing skeptics to entertain the idea that qualia are real – at least in a tentative and conditional manner, as the predictive power of the theory is tested.

Conifold Theory is *not* in opposition to the field of neuroscience. It agrees there is a neuroanatomical and neurophysiological basis for perceptual awareness, attention, memory, predictive processing, decision making, and voluntary behavior. Our mental processes are grounded in physical reality, and we cannot continue to acquire sensory data or act purposefully in the world without the structural and functional integrity of the central nervous system. Conifold Theory does not deny or invalidate this concept, it only describes the biophysical characteristics of cortical neural networks which generate the cohesive stream of perceptual experience, allocate attention toward salient stimuli, guide the formation of mental models, drive spontaneous remodeling of the system into a more optimal state, and initiate voluntary, goal-directed behavior. Furthermore, there is no doubt, given the thousands of studies on the subject, that cortical rhythms underpin conscious processes. In Conifold Theory, oscillations are the outcome of information processing, and their frequency reflects the information being processed. Semantic statements about the state of the environment are realized in a single computational cycle, within the periodicity of gamma rhythms. Syntactical statements, which predict cause and effect relationships between events, require a sequence of neural network states to be parsed for predictive value, so the timescales for compressing syntactical information will naturally be longer. Because these signatures of cortical neuron networks are inevitable and characteristic consequences of this thermodynamic computing process, these various rhythms are likely to manifest in any biological or engineered system that *has* conscious experience.

Conifold Theory is also compatible with Unlimited Associative Learning. The concept of tying consciousness into information storage and information sharing, posited by the latter framework, is exceptionally apt. The idea of information storage and information sharing being evolutionary transition markers, making it possible for a temporary arrangement of matter and energy to achieve far more possible states than would be conceivable otherwise, also posited by Ginsburg and Jablonka, is useful as well. Paradoxically, reducing information entropy yields a sufficiently stable state to permit exploration and gain further entropy – as long as the predictive value gleaned from that initial experience is somehow saved within the system structure. As such, Conifold Theory demonstrates a thermodynamic basis for the evolutionary transition marker, and an intuitive explanation for why genetic and cognitive storage are such powerful drivers of adaptation during natural selection. Here, *information storage, information processing,* and *information sharing* provide superior energetic efficiency in navigating the local environment, making further exploration of the world possible.

Conifold Theory asserts that cortical neuron computation is fundamentally different than classical computation, and cannot be effectively simulated on von Neumann architecture. This is because energy cannot be physically redistributed to electrons upon collapse of the probability density with a transistor-based network. In other words, encoding probabilities as weights in an artificial layered neural network certainly simulates the non-deterministic computation, but this simulation cannot physically implement a redistribution of the Hamiltonian. The construction of engineered systems with similar anatomy and physiology to mammalian cortical neurons would be required to achieve actual consciousness, taking the form of an ambient-temperature, hardware-instantiated quantum computer which allows itself to be perturbed by its environment. The null hypothesis to Conifold Theory is essentially the models presented by Adaptive Resonance Theory and Semantic Pointer Architecture, which suggest that all intelligent processes can be simulated on classical hardware.

Conifold Theory is well in agreement with recent neuroscientific approaches like Decision to Engage and Predictive Processing. Conifold Theory absolutely vindicates the Decision to Engage model, in that noise (within both the stimulus and the neural coding of it) is critical to understanding decision-making. And Conifold Theory agrees that decision-making is the central process and very purpose of consciousness, with probabilistic neural coding driving awareness, confidence, and resultant action.

Conifold Theory also agrees with the Predictive Processing Model, vindicating the Bayesian inference approach to thinking about consciousness as an in-lifetime error-correction mechanism.

According to the Free Energy Principle, the minimization of 'surprise' is the guiding force driving any improvement to the accuracy of predictive models. Conifold Theory asserts this is not just an information theoretic quantity; it is a thermodynamic quantity. In other words, the minimization of energetic expenditure is the guiding force driving any improvement to the accuracy of predictive models. Here, a simple energetic cost function prompts the organism to *decide* between adjusting the belief system, with the goal of better matching reality, or instigating some behavior, with the goal of adjusting reality to meet the prior expectation. Whichever course of action is deemed to be most energy efficient will be selected. This approach solidly grounds the Free Energy Principle in a broadly-applicable process of thermodynamic computation. Like the Free Energy Principle, Conifold Theory provides a general approach to understanding how a system will optimize its configuration in response to its environment, through cognition and behavior. Unlike the Free Energy Principle, Conifold Theory provides a testable, mechanistic framework which ties neural structure and function to conscious processes, including mental representation and non-deterministic output. In this way, Conifold Theory vindicates the Free Energy Principle, and elaborates upon it, with the continuous accumulation of new, conflicting information triggering erroneous priors to be updated, *if and only if* it would be a greater waste of energy to keep them.

According to Entropic Brain Theory, noise and variance are essential signatures of normal wakeful awareness; this information entropy *increases* with psychedelic drugs and *decreases* with drugs that potentiate GABA activity. Once again, Conifold Theory asserts that entropy is not just an information theoretic quantity; it is a thermodynamic quantity. Conifold Theory is in agreement with Entropic Brain Theory, in that increased probability densities are associated with augmented perceptual richness and the occupation of 'unlikely' neural network states. Conifold Theory just goes a step further than Entropic Brain Theory in applying the laws of thermodynamics to identify a *mechanism* by which probabilistic coding gives rise to perceptual content and a *mechanism* by which this perceptual content is parsed for meaning, to gain predictive value from messy, detailed, and often conflicting data.

Conifold Theory is also highly compatible with Integrated Information Theory. While IIT states several axioms, Conifold Theory demonstrates how these features naturally arise in systems that obey certain mathematical and physical laws. Firstly, Conifold Theory shows how a neural network with probabilistic gating mechanisms creates information content in the form of probability densities which exist in higher-dimensional space. That is, consciousness is *real*. Secondly, it shows how this information content is tied to the neural network encoding it. That is, conscious states are *exclusive*. Thirdly, it shows how mental states can be uniquely defined by probabilistic coding of incoming sensory data. That is, *consciousness represents reality*. Fourthly, Conifold Theory specifies that consciously-represented information content is *irreducible*, as it is mathematically and physically equivalent to integrated information – a system-wide wavefunction or integrated probability density existing in higher-dimensional space. And finally, Conifold Theory specifies how consciousness *gives rise to the cause-effect structure of reality* by demonstrating how wavefunction collapse physically reduces the total set of probable neural network states into a single reality which in turn influences subsequent behavior, participating in causation. In short, IIT asserts the necessary features of consciousness, but does not explain the mechanisms underpinning it. Conifold Theory accomplishes both of these tasks, and therefore provides a more complete theory of consciousness than IIT.

Conifold Theory is also highly compatible with Information Closure Theory. Again, while ICT builds a model of consciousness from analytical logic, Conifold Theory demonstrates how this process occurs, based on the physical properties of cortical neural networks. The mathematical model underpinning ICT is built on the interaction of the system S and its environment E, with a state space emerging from the intertwining of fine-grained processes occurring in S and E, respectively. Conifold Theory agrees completely with this premise, and goes a step further to show the mechanical process underlying the computational cycle of information generation and compression, which occurs as a system interacts with its environment over some time evolution. Conifold Theory also details the thermodynamic constraints of this process of ambient-temperature quantum computation, as well as the material requirements for the encoded information to be perceivable by the system as it is undergoing non-deterministic computation. The theoretical framework of ICT yields five core implications, and these are vindicated by Conifold Theory. Firstly, *the brain generates information and compresses that information to glean meaning from it, by physically extracting predictive value from correlated probability amplitudes*. Therefore, both theories agree that consciousness may be considered functionally equivalent to the process of parsing information. Secondly, *consciousness is expected to be a physical process, associated with a physical substrate*. Both theories agree that every conscious percept corresponds to a physical event, whether it is an event in the external world, an event in the body, or a stochastic event arising in the neural network. Conifold Theory has the added advantage of showing how the intangible aspects of consciousness arise from neural structure and function, with information being captured and encoded by the neural membrane surface. Thirdly, *consciousness is a process of encoding the state of the environment into the neural network itself*. Both theories agree that consciousness is inherently self-referential; it is the process by which a system defines its own state in relation to its environment. Conifold Theory has the added advantage here of explaining the process as a discrete process by which the state space is generated and compressed – in terms of holography (with constructive and destructive interference between complex waves yielding a non-deterministic computational outcome), in terms of quantum mechanics (with the integration and derivation of a complex wavefunction yielding eigenstates or observables on the surface boundary region of a high-dimensional probability density), or in terms of thermodynamics (with a unitary change of basis over some time evolution yielding a zero determinant, thereby reducing von Neumann entropy into available free energy). Fourthly, *consciousness is comprised of fine-grained data, but is ignorant of the level of detail in the environment and the system itself which* *generates the coarse-grained experience of consciousness*. Conifold Theory shows how the computing process produces a coarse-grained conscious experience from a fine-grained coding process, due to the unification of information achieved by holography, the unification of information achieved by probabilistic mechanics, and the unification of information achieved by system-wide thermodynamic computing. Finally, ICT predicts that conscious states are self-determining, with the *level* of consciousness corresponding to the degree of perturbance the system permits from the environment and the *contents* of consciousness corresponding to both the size of the state space and the type of input modalities contributing to that state space. Conifold Theory proves this model correct, by demonstrating exactly how non-deterministic computational outcomes are achieved in a cortical neural network undergoing ambient-temperature quantum computation.

Conclusions

The study of consciousness has taken a long and bumpy path, marked by abrupt turns in the road and foggy conditions. Yet progress was made every step of the way. A dialectic between philosophy and biology has usefully led to the articulation of how qualitative perceptual experience differs from contemplation *about* that experience, and how both differ from the neural structures and neural activity underlying these conscious events. Meanwhile, concurrent progress in philosophy and physics led to the general appreciation that *even partially free will* is likely to be incompatible with causal determinism. And researchers from all three fields contributed ideas toward the question of why consciousness, characterized by both perceptual experience and non-deterministic behavior, arose at all. Over the decades, the field of cognitive sciences has evolved to productively merge neuroscience and psychology, identifying the neuroanatomical and neurophysiological correlates of perception, awareness, attention, memory, and decision-making. The discovery of Bayesian inference methods provided critical conceptual advances which tied cognition to statistics, thermodynamics, and information theory. And the development of artificial neural networks revealed both the advantages and the limitations of a reductionist approach. There is now little doubt across the field of cognitive sciences that consciousness is a natural emergent property of neural computation – and yet, a full mechanistic explanation has remained elusive.

Conifold Theory builds on the progress made by philosophers, mathematicians, and neuroscientists who have so ably articulated the problem. This new approach usefully describes consciousness in terms of thermodynamic computing, with each cortical neuron harnessing probabilistic particle states to gate a state change in the macro-scale computational unit. During sensory processing, neural networks cyclically generate and compress system-wide integrated information sets as they physically encode what is happening in the immediate environment of the body. This information is then combined with previously acquired information, stored in memory, to construct cause-effect models of the world. These models or beliefs about the structure and operation of the world are then employed by the organism to select appropriate behaviors, in the context of incoming sensory data. This computational process yields both a cohesive stream of qualitative information content and a non-deterministic signaling outcome for each computational unit in the neural network. A rival energetic cost function, which naturally emerges from the laws of non-equilibrium thermodynamics, guides the organism to both *expend energy* to explore its environment and *save energy* by exploiting previously gained knowledge. By undergoing thermodynamic cycles of information generation and compression, an organism achieves exascale computing power with extraordinarily high energy efficiency and local memory storage, all while parsing *meaning* from incoming sensory information and *choosing what actions to take* based on that information. Conifold Theory therefore asserts the key features of consciousness – perceptual content, predictive models, and non-deterministic behavioral outcomes – are natural emergent properties of cortical neural network computation. Here, consciousness is a useful computational process, which permits an organism to perceive reality, build an understanding of reality, and act effectively within reality.

The goal of modern consciousness studies is to explain consciousness in mechanistic terms, with neural activity naturally giving rise to streaming perceptual experience, attention, awareness, predictive processes, cognitive models of the world, and non-deterministic behavior suited to the present context. While previous approaches have tended to focus on either the cognitive aspects of consciousness, or the physical aspects of neural activity, Conifold Theory successfully links the two. In doing so, it validates and elaborates a number of previous approaches, offering an opportunity to reconcile multiple views into a more comprehensive theoretical framework. The next steps are bound to be exciting, with valuable opportunities to test out the predictions of the theory and connect the field of cognitive sciences more deeply with physics, philosophy, mathematical logic, computer engineering, information theory, neuroscience, and psychology.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzman machines *Cogn Sci, 9*, 147–169.

Adams, R. A., Shipp, S., & Friston, K. J. (2013, May). Predictions not commands: active inference in the motor system. *Brain Struct Funct, 218*(3), 611-643. https://doi.org/10.1007/s00429-012-0475-5

Aitchison, L., & Lengyel, M. (2017, Oct). With or without you: predictive coding and Bayesian inference in the brain. *Curr Opin Neurobiol, 46*, 219-227. https://doi.org/10.1016/j.conb.2017.08.010

Anastassiou, C. A. P., R.; Markram, H.; Koch, C. (2011). Ephatic coupling of cortical neurons. *Nat Neurosci, 14*(2), 217-223.

Araya, J. M. (2018). Emotion and predictive processing. *University of Edinburgh, Dissertation*.

Armstrong, C. M., & Hille, B. (1998, Mar). Voltage-gated ion channels and electrical excitability. *Neuron, 20*(3), 371-380. https://doi.org/10.1016/s0896-6273(00)80981-2

Asadi-Pooya, A. A., Sharan, A., Nei, M., & Sperling, M. R. (2008, Aug). Corpus callosotomy. *Epilepsy Behav, 13*(2), 271-278. https://doi.org/10.1016/j.yebeh.2008.04.020

Asfour, Y. R., Carpenter, G. A., Grossberg, S., & Lesher, G. W. (1993). Fusion ARTMAP: an adaptive fuzzy network for multi-channel classification. *Proceedings of the Third International Conference on Industrial Fuzzy Control and Intelligent Systems*.

Auksztulewicz, R., Schwiedrzik, C. M., Thesen, T., Doyle, W., Devinsky, O., Nobre, A. C., Schroeder, C. E., Friston, K. J., & Melloni, L. (2018, Oct 3). Not All Predictions Are Equal: "What" and "When"

Predictions Modulate Activity in Auditory Cortex through Different Mechanisms. *J Neurosci, 38*(40), 8680-8693. https://doi.org/10.1523/JNEUROSCI.0369-18.2018

Baars, B. J. (1997). In the Theater of Consciousness: Global Workspace Theory. *J Consciousness Studies, 4*, 292-309.

Balduzzi, D., & Tononi, G. (2009). Qualia: The Geometry of Integrated Information. *PLOS Computational Biol*, https://doi.org/10.1371/journal.pcbi.1000462.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. R. (1996, Mar-Apr). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cereb Cortex, 6*(2), 215-225. https://doi.org/10.1093/cercor/6.2.215

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008, Dec 26). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142-1152. https://doi.org/10.1016/j.neuron.2008.09.021

Beckers, G., & Zeki, S. (1995, Feb). The consequences of inactivating areas V1 and V5 on visual motion perception. *Brain, 118 (Pt 1)*, 49-60. https://doi.org/10.1093/brain/118.1.49

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020, Jul 17). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nat Commun, 11*(1), 3625. https://doi.org/10.1038/s41467-020-17236-y

Belopolsky, A. V., Zwaan, L., Theeuwes, J., & Kramer, A. F. (2007, Oct). The size of an attentional window modulates attentional capture by color singletons. *Psychon Bull Rev, 14*(5), 934-938. https://doi.org/10.3758/bf03194124

Berut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., & Lutz, E. (2012, Mar 7). Experimental verification of Landauer's principle linking information and thermodynamics. *Nature, 483*(7388), 187-189. https://doi.org/10.1038/nature10872

Bialek, W., & Rieke, F. (1992, Nov). Reliability and information transmission in spiking neurons. *Trends Neurosci, 15*(11), 428-434. https://doi.org/10.1016/0166-2236(92)90005-s

Birch, J., Ginsburg, S., & Jablonka, E. (2020, Dec 3). Unlimited Associative Learning and the origins of consciousness: a primer and some predictions. *Biol Philos, 35*, 56. https://doi.org/10.1007/s10539-020-09772-0

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227-247.

Block, N. (1997). Biology versus computation in the study of consciousness. *Behavioral and Brain Sciences, 20*(1), 159-166.

Block, N. (2011a). The higher-order approach to consciousness is defunct. *Analysis, 71*(3), 419–431.

Block, N. (2011b). Perceptual consciousness overflows cognitive access. *Trends Cogn Sci, 15*(12), 567–575.

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. (2016). Concepts as semantic pointers: A framework and computational model. *Cognitive Science, 40*(5), 1128-1162.

Bobier, B., Stewart, T. C., & Eliasmith, C. (2014, Jun). A unifying mechanistic model of selective attention in spiking neurons. *PLoS Comput Biol, 10*(6), e1003577. https://doi.org/10.1371/journal.pcbi.1003577

Born, M., Heisenberg, W., & Jordan, P. (1926). Zur Quantenmechanik. II. . *Zeitschrift für Physik, 35*(8-9), 557–615. https://link.springer.com/article/10.1007%2FBF01379806

Brouwer, A., & Carhart-Harris, R. L. (2021, Apr). Pivotal mental states. *J Psychopharmacol, 35*(4), 319-352. https://doi.org/10.1177/0269881120959637

Brown, R. (2015). The HOROR theory of phenomenal consciousness. *Philosophical Studies, 172*(7), 1783–1794.

Buzsaki, G., & Draguhn, A. (2004, Jun 25). Neuronal oscillations in cortical networks. *Science, 304*(5679), 1926-1929. https://doi.org/10.1126/science.1099745

Campbell, A. E., Sumner, P., Singh, K. D., & Muthukumaraswamy, S. D. (2014, Aug). Acute effects of alcohol on stimulus-induced gamma oscillations in human primary visual and motor cortices. *Neuropsychopharmacology, 39*(9), 2104-2113. https://doi.org/10.1038/npp.2014.58

Carhart-Harris, R. L., Bolstridge, M., Rucker, J., Day, C. M., Erritzoe, D., Kaelen, M., Bloomfield, M., Rickard, J. A., Forbes, B., Feilding, A., Taylor, D., Pilling, S., Curran, V. H., & Nutt, D. J. (2016, Jul). Psilocybin with psychological support for treatment-resistant depression: an open-label feasibility study. *Lancet Psychiatry, 3*(7), 619-627. https://doi.org/10.1016/S2215-0366(16)30065-7

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front Hum Neurosci, 8*, 20. https://doi.org/10.3389/fnhum.2014.00020

Carpenter, G. A. (2001, Mar 1). Neural-network models of learning and memory: leading questions and an emerging framework. *Trends Cogn Sci, 5*(3), 114-118. https://doi.org/10.1016/s1364-6613(00)01591-6

Carpenter, G. A., & Grossberg, S. (1987, Dec 1). ART 2: self-organization of stable category recognition codes for analog input patterns. *Appl Opt, 26*(23), 4919-4930. https://doi.org/10.1364/AO.26.004919

Carpenter, G. A., & Grossberg, S. (1990). ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recogntion architectures. *Neural Networks, 3*, 129-152.

Carpenter, G. A., & Grossberg, S. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks, 4*, 565-588.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans Neural Netw, 3*(5), 698-713. https://doi.org/10.1109/72.159059

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1995). A fuzzy ARTMAP nonparametric probability estimator for nonstationary pattern recognition problems. *IEEE Trans Neural Netw, 6*(6), 1330-1336. https://doi.org/10.1109/72.471374

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991). ART2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks, 4*, 493-504.

Carpenter, G. A., & Ross, W. D. (1995). ART-EMAP: A neural network architecture for object recognition by evidence accumulation. *IEEE Trans Neural Netw, 6*(4), 805-818. https://doi.org/10.1109/72.392245

Carrubba, S. F., C.; Chesson, A.L.; Webber, C.L.; Zbilut, J.P.; Marino, A.A. (2008). Magnetosensory evoked potentials: consistent nonlinear phenomena. *Neurosci Res, 60*(1), 95-105.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013, Aug 14). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med, 5*(198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Chalmers, D. (1997). Availability: The cognitive basis of experience. *Behavioral and Brain Sciences, 20*(1), 148-149.

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies, 2*(3), 200-219.

Chambers, J. D., Elgueda, D., Fritz, J. B., Shamma, S. A., Burkitt, A. N., & Grayden, D. B. (2019). Computational Neural Modeling of Auditory Cortical Receptive Fields. *Front Comput Neurosci, 13*, 28. https://doi.org/10.3389/fncom.2019.00028

Chang, A. Y. C., Biehl, M., Yu, Y., & Kanai, R. (2020). Information Closure Theory of Consciousness. *Front Psychol, 11*, 1504. https://doi.org/10.3389/fpsyg.2020.01504

Chater, N. (2018). The Mind is Flat: The Illusion of Mental Depth and The Improvised Mind. *Allen Lane Publishing*.

Chrobak, J. J., & Buzsaki, G. (1998, Jan 1). Gamma oscillations in the entorhinal cortex of the freely behaving rat. *J Neurosci, 18*(1), 388-398. https://www.ncbi.nlm.nih.gov/pubmed/9412515

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X. J., Pouget, A., & Shadlen, M. N. (2011, Feb 24). Variance as a signature of neural computations during decision making. *Neuron, 69*(4), 818-831. https://doi.org/10.1016/j.neuron.2010.12.037

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008, Jun). Decision-making with multiple alternatives. *Nat Neurosci, 11*(6), 693-702. https://doi.org/10.1038/nn.2123

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. *Prog Brain Res, 168*, 19-33. https://doi.org/10.1016/S0079-6123(07)68003-0

Cleeremans, A. (2011). The Radical Plasticity Thesis: How the Brain Learns to be Conscious. *Front Psychol, 2*, 86. https://doi.org/10.3389/fpsyg.2011.00086

Cleeremans, A. (2019). Consciousness Designs Itself. *Journal of Consciousness Studies, 26*(3-4), 88-111.

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J. R., Munoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020, Feb). Learning to Be Conscious. *Trends Cogn Sci, 24*(2), 112-123. https://doi.org/10.1016/j.tics.2019.11.011

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

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Semin Neurosci, 2*, 263-275.

Csibra, G., Davis, G., Spratling, M. W., & Johnson, M. H. (2000, Nov 24). Gamma oscillations and object processing in the infant brain. *Science, 290*(5496), 1582-1585. https://www.ncbi.nlm.nih.gov/pubmed/11090357

Damásio, A. (2000). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. *Mariner Books*.

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

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998, Nov 24). A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A, 95*(24), 14529-14534. https://doi.org/10.1073/pnas.95.24.14529

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

Descartes, R. (1641). Meditation II: On the Nature of the Human Mind, Which Is Better Known Than the Body. *From "Meditations on First Philosophy: With Selections from the Objections and Replies" edited by John Cottingham, 1996, Cambridge: Cambridge University Press.* .

DeWolf, T., Stewart, T. C., Slotine, J. J., & Eliasmith, C. (2016, Nov 30). A spiking neural model of adaptive arm control. *Proc Biol Sci, 283*(1843). https://doi.org/10.1098/rspb.2016.2134

Dorval, A. D., & White, J. A. (2005, Oct 26). Channel noise is essential for perithreshold oscillations in entorhinal stellate neurons. *J Neurosci, 25*(43), 10025-10028. https://doi.org/10.1523/JNEUROSCI.3557-05.2005

Echeveste, R., Aitchison, L., Hennequin, G., & Lengyel, M. (2020, Sep). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nat Neurosci, 23*(9), 1138-1149. https://doi.org/10.1038/s41593-020-0671-1

Edelman, G. (2004). Wider Than the Sky: The Phenomenal Gift of Consciousness. *Yale University Press*.

Elgueda, D., Duque, D., Radtke-Schuller, S., Yin, P., David, S. V., Shamma, S. A., & Fritz, J. B. (2019, Mar). State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. *Nat Neurosci, 22*(3), 447-459. https://doi.org/10.1038/s41593-018-0317-8

Eliasmith, C. (2005, Jun). A unified approach to building and controlling spiking attractor networks. *Neural Comput, 17*(6), 1276-1314. https://doi.org/10.1162/0899766053630332

Engel, A. K., & Singer, W. (2001, Jan 1). Temporal binding and the neural correlates of sensory awareness. *Trends Cogn Sci, 5*(1), 16-25. https://www.ncbi.nlm.nih.gov/pubmed/11164732

Engl, E., & Attwell, D. (2015, Aug 15). Non-signalling energy use in the brain. *J Physiol, 593*(16), 3417-3429. https://doi.org/10.1113/jphysiol.2014.282517

Esteve, J. G., Falceto, F., & Garcia Canal, C. (2010). Generalization of the Hellmann-Feynman theorem. *Physics Letters A, 374*(6), 819-822. https://arxiv.org/pdf/0912.4153

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008, Apr). Noise in the nervous system. *Nat Rev Neurosci, 9*(4), 292-303. https://doi.org/10.1038/nrn2258

Faraji, M. R., & Qi, X. (2016). Face recognition under varying illuminations using logarithmic fractal dimension-based complete eight local directional patterns. *Neurocomputing, 199*, 16-30. https://doi.org/10.1016/J.NEUCOM.2016.01.094

Feynman, R. P. (1939). Forces in Molecules. *Physical Review, 56*, 340.

Forseth, K. J., Hickok, G., Rollo, P. S., & Tandon, N. (2020, Oct 16). Language prediction mechanisms in human auditory cortex. *Nat Commun, 11*(1), 5240. https://doi.org/10.1038/s41467-020-19010-6

Frankish, K. (2016). Illusionism as a theory of consciousness. *J Consciousness Studies, 23*(11-12), 11-39.

Franz, E. A., Waldie, K. E., & Smith, M. J. (2000, Jan). The effect of callosotomy on novel versus familiar bimanual actions: a neural dissociation between controlled and automatic processes? *Psychol Sci, 11*(1), 82-85. https://doi.org/10.1111/1467-9280.00220

Friston, K. (2010, Feb). The free-energy principle: a unified brain theory? *Nat Rev Neurosci, 11*(2), 127-138. https://doi.org/10.1038/nrn2787

Friston, K., Fortier, M., & Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. . *ALIUS Bulletin, 2*, 17-43.

Friston, K., & Kiebel, S. (2009, May 12). Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B Biol Sci, 364*(1521), 1211-1221. https://doi.org/10.1098/rstb.2008.0300

Friston, K., Kilner, J., & Harrison, L. (2006, Jul-Sep). A free energy principle for the brain. *J Physiol Paris, 100*(1-3), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Fröhlich, F. M., D.A. (2010). Endogenous Electric Fields May Guide Neocortical Network Activity. *Neuron, 67*(1), 129–143.

Gabor, D. (1948). A new microscopic principle. *Nature 161*, 777.

Gabor, D. (1968, Feb 10). Holographic model of temporal recall. *Nature, 217*(5128), 584. https://doi.org/10.1038/217584a0

Gallea, M. A. (2017). A brief reflection on the not-so-brief history of the lobotomy. *BC Medical Journal, 59*(6), 302-304.

Gamerman, D., & Lopes, H. F. (2006). Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (Second Edition). *Chapman & Hall*.

Gazzaniga, M. S. (2011). Who's in Charge?: Free Will and the Science of the Brain. *Ecco Press, New York*.

Ginsburg, S., & Jablonka, E. (2010, Sep 7). The evolution of associative learning: A factor in the Cambrian explosion. *J Theor Biol, 266*(1), 11-20. https://doi.org/10.1016/j.jtbi.2010.06.017

Glansdorff, P., & Prigogine, I. (1971). Thermodynamic theory of structure, stability, and fluctuations. *London: Wiley-Interscience*.

Goddard, C. A., Sridharan, D., Huguenard, J. R., & Knudsen, E. I. (2012, Feb 9). Gamma oscillations are generated locally in an attention-related midbrain network. *Neuron, 73*(3), 567-580. https://doi.org/10.1016/j.neuron.2011.11.028

Goodale, M. A., & Milner, A. D. (1992, Jan). Separate visual pathways for perception and action. *Trends Neurosci, 15*(1), 20-25. https://doi.org/10.1016/0166-2236(92)90344-8

Gosmann, J., & Eliasmith, C. (2016). Optimizing Semantic Pointer Representations for Symbol-Like Processing in Spiking Neural Networks. *PLoS One, 11*(2), e0149928. https://doi.org/10.1371/journal.pone.0149928

Grandy, W. T. (2008). Entropy and the time evolution of macroscopic systems. *Oxford: Oxford University Press*.

Graziano, M. S. A. (2021, Jan-Jan). What makes us so certain that we're conscious? *Cogn Neurosci, 12*(2), 67-68. https://doi.org/10.1080/17588928.2020.1838468

Graziano, M. S. A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cogn Neurosci, 2*(2), 98-113.

Griffiths, T. D., Rees, G., Rees, A., Green, G. G., Witton, C., Rowe, D., Buchel, C., Turner, R., & Frackowiak, R. S. (1998, May). Right parietal cortex is involved in the perception of sound movement in humans. *Nat Neurosci, 1*(1), 74-79. https://doi.org/10.1038/276

Grion, N., Akrami, A., Zuo, Y., Stella, F., & Diamond, M. E. (2016, Feb). Coherence between Rat Sensorimotor System and Hippocampus Is Enhanced during Tactile Discrimination. *PLoS Biol, 14*(2), e1002384. https://doi.org/10.1371/journal.pbio.1002384

Groen, I. I. A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2018, Dec). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Comput Biol, 14*(12), e1006690. https://doi.org/10.1371/journal.pcbi.1006690

Grossmann, T., Johnson, M. H., Farroni, T., & Csibra, G. (2007, Dec). Social perception in the infant brain: gamma oscillatory activity in response to eye gaze. *Soc Cogn Affect Neurosci, 2*(4), 284-291. https://doi.org/10.1093/scan/nsm025

Hagiwara, K., Okamoto, T., Shigeto, H., Ogata, K., Somehara, Y., Matsushita, T., Kira, J., & Tobimatsu, S. (2010, May 15). Oscillatory gamma synchronization binds the primary and secondary somatosensory areas in humans. *Neuroimage, 51*(1), 412-420. https://doi.org/10.1016/j.neuroimage.2010.02.001

Haider, B., Duque, A., Hasenstaub, A. R., & McCormick, D. A. (2006, Apr 26). Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *J Neurosci, 26*(17), 4535-4545. https://doi.org/10.1523/JNEUROSCI.5297-05.2006

Hameroff, S. R., & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness. . *Toward a science of consciousness; the first Tucson discussions and debates. S.R. Hameroff, A.W. Kaszniak, A.C. Scott (Eds.) MIT Press*.

Hameroff, S. R., & Penrose, R. (2014). Consciousness in the universe: A review of the 'OrchOR' theory. *Physics of Life Reviews, 11*(1), 39-78.

Hanada, T. (2020). Ionotropic glutamate receptors in epilepsy: A review focusing on AMPA and NMDA receptors. *Biomolecules, 10*(3), 464.

Hanks, T. D. M., M.E.; Kiani, R.; Hopp, E.; Shadlen, M.N. . (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. . *J Neurosci, 31*(7), 6339-6352.

Harel, A., Kravitz, D. J., & Baker, C. I. (2013, Apr). Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cereb Cortex, 23*(4), 947-957. https://doi.org/10.1093/cercor/bhs091

Harris, A. Z., & Gordon, J. A. (2015, Jul 8). Long-range neural synchrony in behavior. *Annu Rev Neurosci, 38*, 171-194. https://doi.org/10.1146/annurev-neuro-071714-034111

Harris, H. D., Murphy, G. L., & Rehder, B. (2008, Oct). Prior knowledge and exemplar frequency. *Mem Cognit, 36*(7), 1335-1350. https://doi.org/10.3758/MC.36.7.1335

Hayes, B. K., & Taplin, J. E. (1993a, Sep). Development of conceptual knowledge in children with mental retardation. *Am J Ment Retard, 98*(2), 293-303. https://www.ncbi.nlm.nih.gov/pubmed/8398088

Hayes, B. K., & Taplin, J. E. (1993b, Jun). Developmental differences in the use of prototype and exemplar-specific information. *J Exp Child Psychol, 55*(3), 329-352. https://doi.org/10.1006/jecp.1993.1019

Hebb, D. O. (1949). The Organization of Behavior. . *New York: Wiley & Sons*.

Hegde, J., & Van Essen, D. C. (2000, Mar 1). Selectivity for complex shapes in primate visual area V2. *J Neurosci, 20*(5), RC61. https://www.ncbi.nlm.nih.gov/pubmed/10684908

Herrmann, C. S., Munk, M. H., & Engel, A. K. (2004, Aug). Cognitive functions of gamma-band activity: memory match and utilization. *Trends Cogn Sci, 8*(8), 347-355. https://doi.org/10.1016/j.tics.2004.06.006

Herweg, N. A., Apitz, T., Leicht, G., Mulert, C., Fuentemilla, L., & Bunzeck, N. (2016, Mar 23). Theta-Alpha Oscillations Bind the Hippocampus, Prefrontal Cortex, and Striatum during Recollection: Evidence from Simultaneous EEG-fMRI. *J Neurosci, 36*(12), 3579-3587. https://doi.org/10.1523/JNEUROSCI.3629-15.2016

Herzog, R., Mediano, P. A. M., Rosas, F. E., Carhart-Harris, R., Perl, Y. S., Tagliazucchi, E., & Cofre, R. (2020, Oct 20). A mechanistic model of the neural entropy increase elicited by psychedelic drugs. *Sci Rep, 10*(1), 17725. https://doi.org/10.1038/s41598-020-74060-6

Hesselmann, G., Flandin, G., & Dehaene, S. (2011, Jun 1). Probing the cortical network underlying the psychological refractory period: a combined EEG-fMRI study. *Neuroimage, 56*(3), 1608-1621. https://doi.org/10.1016/j.neuroimage.2011.03.017

Heusser, K. T., D.; Thoss, F. (1997). Influence of an alternating 3 Hz magnetic field with an induction of 0.1 millitesla on chosen parameters of the human occipital EEG. *Neurosci Letters, 239*(2-3), 57-60.

Hillert, M., & Agren, J. (2006). Extremum principles for irreversible processes. *Acta Materialia, 54*(8), 2063-2066.

Hoffman, D. (2019). The case against reality: Why evolution hid the truth from our eyes. . *W.W. Norton & Company*.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Front Psychol, 5*, 577. https://doi.org/10.3389/fpsyg.2014.00577

Hoffman, D. D., & Singh, M. (2012). Computational evolutionary perception. *Perception, 41*(9), 1073-1091. https://doi.org/10.1068/p7275

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Front Psychol, 3*, 96. https://doi.org/10.3389/fpsyg.2012.00096

Hohwy, J. (2017, Jan). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Conscious Cogn, 47*, 75-85. https://doi.org/10.1016/j.concog.2016.09.004

Hohwy, J., Roepstorff, A., & Friston, K. (2008, Sep). Predictive coding explains binocular rivalry: an epistemological review. *Cognition, 108*(3), 687-701. https://doi.org/10.1016/j.cognition.2008.05.010

Hopfield, J. J. (1982, Apr). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A, 79*(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554

Hopfield, J. J. (1984, May). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Natl Acad Sci U S A, 81*(10), 3088-3092. https://doi.org/10.1073/pnas.81.10.3088

Hopfield, J. J. (1987, Dec). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proc Natl Acad Sci U S A, 84*(23), 8429-8433. https://doi.org/10.1073/pnas.84.23.8429

Howarth, C., Gleeson, P., & Attwell, D. (2012, Jul). Updated energy budgets for neural computation in the neocortex and cerebellum. *J Cereb Blood Flow Metab, 32*(7), 1222-1232. https://doi.org/10.1038/jcbfm.2012.35

Hubel, D. H., & Wiesel, T. N. (1959, Oct). Receptive fields of single neurones in the cat's striate cortex. *J Physiol, 148*, 574-591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hunsberger, E., Scott, M., & Eliasmith, C. (2014, Aug). The competing benefits of noise and heterogeneity in neural coding. *Neural Comput, 26*(8), 1600-1623. https://doi.org/10.1162/NECO_a_00621

Jablonka, E., & Szathmary, E. (1995, May). The evolution of information storage and heredity. *Trends Ecol Evol, 10*(5), 206-211. https://doi.org/10.1016/s0169-5347(00)89060-6

Jackson, A., Mavoori, J., & Fetz, E. E. (2006). Long-term motor cortex plasticity induced by an electronic neural implant. *Nature, 444*(7115), 56-60.

John, E. R. (2002). The neurophysics of consciousness. *Brain Research Reviews, 39*, 1-28.

Jun, Y., Gavrilov, M., & Beckhoefer, J. (2014). High-precision test of Landauer's principle in a feedback trap. *Phys Rev Lett, 113*, 190601.

Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neurosci Conscious, 2019*(1), niz016. https://doi.org/10.1093/nc/niz016

Kanwisher, N., McDermott, J., & Chun, M. M. (1997, Jun 1). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci, 17*(11), 4302-4311. https://www.ncbi.nlm.nih.gov/pubmed/9151747

Kean, S. (2014). Phineas Gage, Neuroscience's Most Famous Patient. *Slate*.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010, Jan 13). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J Neurosci, 30*(2), 620-628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Kiani, R., Corthell, L., & Shadlen, M. N. (2014, Dec 17). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329-1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008, Mar 19). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *J Neurosci, 28*(12), 3017-3029. https://doi.org/10.1523/JNEUROSCI.4761-07.2008

Kim, J., Ricci, M., & Serre, T. (2018, Aug 6). Not-So-CLEVR: learning same-different relations strains feedforward neural networks. *Interface Focus, 8*(4), 20180011. https://doi.org/10.1098/rsfs.2018.0011

Kita, S., & Lausberg, H. (2008, Feb). Generation of co-speech gestures based on spatial imagery from the right-hemisphere: evidence from split-brain patients. *Cortex, 44*(2), 131-139. https://doi.org/10.1016/j.cortex.2006.04.001

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neurosci Conscious, 2021*(1), niab001. https://doi.org/10.1093/nc/niab001

Koch, C. (2014). Neuronal "Superhub" Might Generate Consciousness. *Scientific American*.

Köhler, W. (1947). Gestalt Psychology: An introduction to new concepts in modern psychology. *New York: Liveright Publishing Corporation*.

Kok, P., Jehee, J. F., & de Lange, F. P. (2012, Jul 26). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron, 75*(2), 265-270. https://doi.org/10.1016/j.neuron.2012.04.034

Kok, P., Rahnev, D., Jehee, J. F., Lau, H. C., & de Lange, F. P. (2012, Sep). Attention reverses the effect of prediction in silencing sensory signals. *Cereb Cortex, 22*(9), 2197-2206. https://doi.org/10.1093/cercor/bhr310

Kok, P., van Lieshout, L. L., & de Lange, F. P. (2016, Nov 22). Local expectation violations result in global activity gain in primary visual cortex. *Sci Rep, 6*, 37706. https://doi.org/10.1038/srep37706

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013, Jan). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn Sci, 17*(1), 26-49. https://doi.org/10.1016/j.tics.2012.10.011

Kreiman, G., Koch, C., & Fried, I. (2000, Sep). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci, 3*(9), 946-953. https://doi.org/10.1038/78868

Lachmann, M., Sell, G., & Jablonka, E. (2000, Jul 7). On the advantages of information sharing. *Proc Biol Sci, 267*(1450), 1287-1293. https://doi.org/10.1098/rspb.2000.1140

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends Cogn Sci, 10*(11), 494-501.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development, 5*, 183-191.

Lau, H., & Rosenthal, D. M. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci, 15*(8), 365–373.

LeDoux, J. E., & Brown, R. (2017, Mar 7). A higher-order theory of emotional consciousness. *Proc Natl Acad Sci U S A, 114*(10), E2016-E2025. https://doi.org/10.1073/pnas.1619316114

Lee, U., Mashour, G. A., Kim, S., Noh, G. J., & Choi, B. M. (2009, Mar). Propofol induction reduces the capacity for neural information integration: implications for the mechanism of consciousness and general anesthesia. *Conscious Cogn, 18*(1), 56-64. https://doi.org/10.1016/j.concog.2008.10.005

Leopold, D. A., & Logothetis, N. K. (1996, Feb 8). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature, 379*(6565), 549-553. https://doi.org/10.1038/379549a0

Locke, J. (1689). An essay concerning human understanding. *London: Penguin Group, Reprinted Edition 1997 by Roger Woolhouse*.

Logothetis, N. K., & Schall, J. D. (1989, Aug 18). Neuronal correlates of subjective visual perception. *Science, 245*(4919), 761-763. https://doi.org/10.1126/science.2772635

Logothetis, N. K., & Schall, J. D. (1990). Binocular motion rivalry in macaque monkeys: eye dominance and tracking eye movements. *Vision Res, 30*(10), 1409-1419. https://doi.org/10.1016/0042-6989(90)90022-d

Longuet-Higgins, H. C. (1968, Jan 6). Holographic model of temporal recall. *Nature, 217*(5123), 104. https://doi.org/10.1038/217104a0

Luppi, A. I., Carhart-Harris, R. L., Roseman, L., Pappas, I., Menon, D. K., & Stamatakis, E. A. (2021, Feb 15). LSD alters dynamic integration and segregation in the human brain. *Neuroimage, 227*, 117653. https://doi.org/10.1016/j.neuroimage.2020.117653

MacNeil, D., & Eliasmith, C. (2011). Fine-tuning and the stability of recurrent neural networks. *PLoS One, 6*(9), e22885. https://doi.org/10.1371/journal.pone.0022885

Malach, R. (2011). Conscious perception and the frontal lobes: comment on Hau and Rosenthal. *Trends Cogn Sci, 15*(11), P507.

Malnic, B., Hirono, J., Sato, T., & Buck, L. B. (1999). Combinatorial receptor codes for odors. *Cell, 5*(5), 713-723. https://doi.org/10.1016/S0092-8674(00)80581-4

Maoz, O., Tkacik, G., Esteki, M. S., Kiani, R., & Schneidman, E. (2020, Oct 6). Learning probabilistic neural representations with randomly connected circuits. *Proc Natl Acad Sci U S A, 117*(40), 25066-25073. https://doi.org/10.1073/pnas.1912804117

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1987). The Appeal of Parallel Distributed Processing. *MIT Press*.

McFadden, J. (2002). The Conscious Electromagnetic Information (Cemi) Field Theory: The Hard Problem Made Easy? *J Consciousness Studies, 9*(8), 45-60.

McFadden, J. (2013). The CEMI Field Theory Gestalt Information and the Meaning of Meaning. *J Consciousness Studies, 20*(3-4), 3-4.

McKemmish, L. K., Reimers, J. R., McKenzie, R. H., Mark, A. E., & Hush, N. S. (2009, Aug). Penrose-Hameroff orchestrated objective-reduction proposal for human consciousness is not biologically feasible.

Phys Rev E Stat Nonlin Soft Matter Phys, 80(2 Pt 1), 021912. https://doi.org/10.1103/PhysRevE.80.021912

Meister, M. (1996). Multineuronal codes in retinal signaling. *Proc Natl Acad Sci U S A, 93*, 609-614.

Melloni, L., Schwiedrzik, C. M., Mueller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *J Neurosci, 31*(4), 1386-1396.

Merleau-Ponty, M. (1945). Phenomenology of Perception. *Editions Gallimard (Paris). Translated into English by Colin Smith and published by Routledge & Kegan Paul (London) in 1962.*

Mestre, D. R., Brouchon, M., Ceccaldi, M., & Poncet, M. (1992, Sep). Perception of optical flow in cortical blindness: a case report. *Neuropsychologia, 30*(9), 783-795. https://doi.org/10.1016/0028-3932(92)90082-w

Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. *MIT Press*.

Metzinger, T. (2007). Self Models. *Scholarpedia, 2*(10), 4174.

Metzinger, T. (2017). Suffering. *The Return of Consciousness: A New Science on Old Questions, Edited by Kurt Almquist and Anders Haag*.

Mitra, A., Snyder, A. Z., Tagliazucchi, E., Laufs, H., & Raichle, M. E. (2015, Nov 9). Propagated infra-slow intrinsic brain activity reorganizes across wake and slow wave sleep. *Elife, 4*. https://doi.org/10.7554/eLife.10781

Mostert, P., Kok, P., & de Lange, F. P. (2015, Dec 15). Dissociating sensory from decision processes in human perceptual decision making. *Sci Rep, 5*, 18253. https://doi.org/10.1038/srep18253

Naundorf, B., Wolf, F., & Volgushev, M. (2006, Apr 20). Unique features of action potential initiation in cortical neurons. *Nature, 440*(7087), 1060-1063. https://doi.org/10.1038/nature04610

Newman, J., & Grace, A. A. (1999). Binding across Time: The Selective Gating of Frontal and Hippocampal Systems Modulating Working Memory and Attentional States. *Consciousness and Cognition, 8*(2), 196-212.

O'Connor, T., & Franklin, C. (2020). Free Will. *The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta*.

Olsen, R. W., Yang, J., King, R. G., Dilber, A., Stauber, G. B., & Ransom, R. W. (1986, Nov 24). Barbiturate and benzodiazepine modulation of GABA receptor binding and function. *Life Sci, 39*(21), 1969-1976. https://doi.org/10.1016/0024-3205(86)90320-6

Olthof, B. M. J., Rees, A., & Gartside, S. E. (2019, Nov 6). Multiple Nonauditory Cortical Regions Innervate the Auditory Midbrain. *J Neurosci, 39*(45), 8916-8928. https://doi.org/10.1523/JNEUROSCI.1436-19.2019

Ostojic, S., Brunel, N., & Hakim, V. (2009, Aug 19). How connectivity, background activity, and synaptic properties shape the cross-correlation between spike trains. *J Neurosci, 29*(33), 10234-10253. https://doi.org/10.1523/JNEUROSCI.1275-09.2009

Overgaard, M., & Overgaard, R. (2010). Neural correlates of contents and levels of consciousness. *Front Psychol, 1*, 164. https://doi.org/10.3389/fpsyg.2010.00164

Pal, D., Li, D., Dean, J. G., Brito, M. A., Liu, T., Fryzel, A. M., Hudetz, A. G., & Mashour, G. A. (2020, Jan 15). Level of Consciousness Is Dissociable from Electroencephalographic Measures of Cortical Connectivity, Slow Oscillations, and Complexity. *J Neurosci, 40*(3), 605-618. https://doi.org/10.1523/JNEUROSCI.1910-19.2019

Paradis, A. L., Cornilleau-Peres, V., Droulez, J., Van De Moortele, P. F., Lobel, E., Berthoz, A., Le Bihan, D., & Poline, J. B. (2000, Aug). Visual perception of motion and 3-D structure from motion: an fMRI study. *Cereb Cortex, 10*(8), 772-783. https://doi.org/10.1093/cercor/10.8.772

Parrish, R. R., Codadu, N. K., Mackenzie-Gray Scott, C., & Trevelyan, A. J. (2019, Apr). Feedforward inhibition ahead of ictal wavefronts is provided by both parvalbumin- and somatostatin-expressing interneurons. *J Physiol, 597*(8), 2297-2314. https://doi.org/10.1113/JP277749

Penrose, R. (1979). Singularities and time-asymmetry, in General Relativity: An Einstein Centennary Volume, edited by S. W. Hawking and W. Israel. *Cambridge University Press*.

Penrose, R. (1989). Shadows of the Mind: A Search for the Missing Science of Consciousness. *Oxford University Press*.

Penrose, R. (2008). Causality, quantum theory and cosmology, in On Space And Time, edited by S. Majid. *Cambridge University Press*.

Pins, D., & Ffytche, D. (2003, May). The neural correlates of conscious vision. *Cereb Cortex, 13*(5), 461-474. https://doi.org/10.1093/cercor/13.5.461

Pinto, Y., Neville, D. A., Otten, M., Corballis, P. M., Lamme, V. A. F., de Haan, E. H. F., Foschi, N., & Fabri, M. (2017, May 1). Split brain: divided perception but undivided consciousness. *Brain, 140*(5), 1231-1237. https://doi.org/10.1093/brain/aww358

Pinto, Y., Vandenbroucke, A. R., Otten, M., Sligte, I. G., Seth, A. K., & Lamme, V. A. F. (2017, Feb). Conscious visual memory with minimal attention. *J Exp Psychol Gen, 146*(2), 214-226. https://doi.org/10.1037/xge0000255

Pistohl, T., Joshi, D., Ganesh, G., Jackson, A., & Nazarpour, K. (2015, May). Artificial proprioceptive feedback for myoelectric control. *IEEE Trans Neural Syst Rehabil Eng, 23*(3), 498-507. https://doi.org/10.1109/TNSRE.2014.2355856

Plant, G. T., Laxer, K. D., Barbaro, N. M., Schiffman, J. S., & Nakayama, K. (1993, Dec). Impaired visual motion perception in the contralateral hemifield following unilateral posterior cerebral lesions in humans. *Brain, 116 (Pt 6)*, 1303-1335. https://doi.org/10.1093/brain/116.6.1303

Pockett, S. (2000). The Nature of Consciousness: A Hypothesis. *iUniverse Press*.

Powers, R. K., & Binder, M. D. (1995, Aug). Effective synaptic current and motoneuron firing rate modulation. *J Neurophysiol, 74*(2), 793-801. https://doi.org/10.1152/jn.1995.74.2.793

Pribram, K. H., & Meade, S. D. (1999). Conscious awareness: Processing in the synaptodendritic web. *New Ideas in Psychology, 17*(3), 205-214.

Pulvermuller, F., Eulitz, C., Pantev, C., Mohr, B., Feige, B., Lutzenberger, W., Elbert, T., & Birbaumer, N. (1996, Jan). High-frequency cortical responses reflect lexical processing: an MEG study. *Electroencephalogr Clin Neurophysiol, 98*(1), 76-85. https://www.ncbi.nlm.nih.gov/pubmed/8689998

Ramachandran, V. S. (2011). The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human. *W. W. Norton Company*.

Rees, G., Kreiman, G., & Koch, C. (2002, Apr). Neural correlates of consciousness in humans. *Nat Rev Neurosci, 3*(4), 261-270. https://doi.org/10.1038/nrn783

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009, Sep 10). Changes of mind in decision-making. *Nature, 461*(7261), 263-266. https://doi.org/10.1038/nature08275

Reuter, F., Del Cul, A., Malikova, I., Naccache, L., Confort-Gouny, S., Cohen, L., Cherif, A. A., Cozzone, P. J., Pelletier, J., Ranjeva, J. P., Dehaene, S., & Audoin, B. (2009, Jan 15). White matter damage impairs access to consciousness in multiple sclerosis. *Neuroimage, 44*(2), 590-599. https://doi.org/10.1016/j.neuroimage.2008.08.024

Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B., & Varela, F. J. (1999, Feb 4). Perception's shadow: long-distance synchronization of human brain activity. *Nature, 397*(6718), 430-433. https://doi.org/10.1038/17120

Rosenthal, D. M. (2005). Consciousness and Mind. *Clarendon Press*

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia, 46*, 829-840.

Roxin, A., Brunel, N., Hansel, D., Mongillo, G., & van Vreeswijk, C. (2011, Nov 9). On the distribution of firing rates in networks of cortical neurons. *J Neurosci, 31*(45), 16217-16226. https://doi.org/10.1523/JNEUROSCI.1677-11.2011

Sabokrou, M., Fayyaz, M., Fathy, M., & Klette, R. (2017, Apr). Deep-cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans Image Process, 26*(4), 1992-2004. https://doi.org/10.1109/TIP.2017.2670780

Sacks, O. (1985). The Man Who Mistook His Wife For A Hat & Other Clinical Tales. *Summit Books*.

Scellier, B., & Bengio, Y. (2019, Feb). Equivalence of equilibrium propagation and recurrent backpropagation. *Neural Comput, 31*(2), 312-329. https://doi.org/10.1162/neco_a_01160

Schiff, N. D., Nauvel, T., & Victor, J. D. (2014, Apr). Large-scale brain dynamics in disorders of consciousness. *Curr Opin Neurobiol, 25*, 7-14. https://doi.org/10.1016/j.conb.2013.10.007

Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W., & Melloni, L. (2014, May). Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. *Cereb Cortex, 24*(5), 1152-1164. https://doi.org/10.1093/cercor/bhs396

Searle, J., Dennett, D., & Chalmers, D. (1997). The Mystery of Consciousness. *New York: The New York Review of Books*.

Sengupta, B., Stemmler, M., Laughlin, S. B., & Niven, J. E. (2010, Jul 1). Action potential energy efficiency varies among neuron types in vertebrates and invertebrates. *PLoS Comput Biol, 6*, e1000840. https://doi.org/10.1371/journal.pcbi.1000840

Seth, A. K., & Hohwy, J. (2021, Jan-Jan). Predictive processing as an empirical theory for consciousness science. *Cogn Neurosci, 12*(2), 89-90. https://doi.org/10.1080/17588928.2020.1838467

Seth, A. K. B., A.B., Barnett, L. (2011). Causal density and integrated information as measures of conscious level. *Philos Trans A Math Phys Eng Sci., 369*(1952), 3748-3767.

Shadlen, M. N., & Kiani, R. (2013, Oct 30). Decision making as a window on cognition. *Neuron, 80*(3), 791-806. https://doi.org/10.1016/j.neuron.2013.10.047

Sheinberg, D. L., & Logothetis, N. K. (1997, Apr 1). The role of temporal cortical areas in perceptual organization. *Proc Natl Acad Sci U S A, 94*(7), 3408-3413. https://doi.org/10.1073/pnas.94.7.3408

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016, Jan 28). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489. https://doi.org/10.1038/nature16961

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

Singh, R., & Eliasmith, C. (2006, Apr 5). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *J Neurosci, 26*(14), 3667-3678. https://doi.org/10.1523/JNEUROSCI.4864-05.2006

Sitt, J. D., King, J. R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., & Naccache, L. (2014, Aug). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain, 137*(Pt 8), 2258-2270. https://doi.org/10.1093/brain/awu141

Snyder, J. S., Schwiedrzik, C. M., Vitela, A. D., & Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Front Hum Neurosci, 9*, 594. https://doi.org/10.3389/fnhum.2015.00594

Softky, W. R., & Koch, C. (1993, Jan). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci, 13*(1), 334-350. https://www.ncbi.nlm.nih.gov/pubmed/8423479

Solomon, S. G., & Lennie, P. (2007, Apr). The machinery of colour vision. *Nat Rev Neurosci, 8*(4), 276-286. https://doi.org/10.1038/nrn2094

Sperling, M. R., Feldman, H., Kinman, J., Liporace, J. D., & O'Connor, M. J. (1999, Jul). Seizure control and mortality in epilepsy. *Ann Neurol, 46*(1), 45-50. https://doi.org/10.1002/1531-8249(199907)46:1<45::aid-ana8>3.0.co;2-i

Stacey, W. C., Krieger, A., & Litt, B. (2011, Apr). Network recruitment to coherent oscillations in a hippocampal computer model. *J Neurophysiol, 105*(4), 1464-1481. https://doi.org/10.1152/jn.00643.2010

Steriade, M., Timofeev, I., & Grenier, F. (2001, May). Natural waking and sleep states: a view from inside neocortical neurons. *J Neurophysiol, 85*(5), 1969-1985. https://doi.org/10.1152/jn.2001.85.5.1969

Stern, E. A., Kincaid, A. E., & Wilson, C. J. (1997, Apr). Spontaneous subthreshold membrane potential fluctuations and action potential variability of rat corticostriatal and striatal neurons in vivo. *J Neurophysiol, 77*(4), 1697-1715. https://doi.org/10.1152/jn.1997.77.4.1697

Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of prediction. *Physical Review Letters, 109*(12), 120604.

Stoll, E. A. (under review-a). Modeling the probabilistic behavior of electrons at the neural membrane yields a holographic projection of information content.

Stoll, E. A. (under review-b). The mechanisms underpinning non-deterministic computation in cortical neural networks.

Stoll, E. A. (under review-c). Random electrical noise drives non-deterministic computation in cortical neural networks.

Stoll, E. A. (in prep-d). The explanatory power of Conifold Theory.

Stoll, E. A. (in prep-e). The neuroscientific predictions of Conifold Theory.

Stockel, A., & Eliasmith, C. (2021, Jan). Passive Nonlinear Dendritic Interactions as a Computational Resource in Spiking Neural Networks. *Neural Comput, 33*(1), 96-128. https://doi.org/10.1162/neco_a_01338

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

Stuss, D. T. (1991). Disturbance of self-awareness after frontal system damage, in Awareness of Deficit After Brain Injury: Clinical & Theoretical Issues, edited by G.P. Prigatano & D.L. Schacter. *Oxford University Press*, 63-83.

Taberner, A. M., & Liberman, M. C. (2005). Response properties of single auditory nerve fibers in the mouse. *J Neurophysiol, 93*(1), 557-569. https://doi.org/10.1152/jn.00574.2004

Tagliazucchi, E., Balenzuela, P., Fraiman, D., & Chialvo, D. R. (2012). Criticality in large-scale brain FMRI dynamics unveiled by a novel point process analysis. *Front Physiol, 3*, 15. https://doi.org/10.3389/fphys.2012.00015

Tagliazucchi, E., Carhart-Harris, R., Leech, R., Nutt, D., & Chialvo, D. R. (2014, Nov). Enhanced repertoire of brain dynamical states during the psychedelic experience. *Hum Brain Mapp, 35*(11), 5442-5456. https://doi.org/10.1002/hbm.22562

Tagliazucchi, E., Chialvo, D. R., Siniatchkin, M., Amico, E., Brichant, J. F., Bonhomme, V., Noirhomme, Q., Laufs, H., & Laureys, S. (2016, Jan). Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J R Soc Interface, 13*(114), 20151027. https://doi.org/10.1098/rsif.2015.1027

Tan, A. H., Lu, N., & Xiao, D. (2008). Integrating temporal difference methods and self-organizing neural networks for reinforcement learning with delayed evaluative feedback. *IEEE Trans Neural Netw, 9*(2), 230-244.

Tegmark, M. (2000). The importance of quantum decoherence in brain processes. *Physical Review, E61*, 4194-4206.

Thibaut, J. P., Gelaes, S., & Murphy, G. L. (2018, May). Does practice in category learning increase rule use or exemplar use-or both? *Mem Cognit, 46*(4), 530-543. https://doi.org/10.3758/s13421-017-0782-4

Thiele, A., & Stoner, G. (2003, Jan 23). Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature, 421*(6921), 366-370. https://doi.org/10.1038/nature01285

Tomb, I., Hauser, M., Deldin, P., & Caramazza, A. (2002, Nov). Do somatic markers mediate decisions on the gambling task? *Nat Neurosci, 5*(11), 1103-1104; author reply 1104. https://doi.org/10.1038/nn1102-1103

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 2*(5), 42.

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch Ital Biol. , 150*(2-3), 56-90.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci, 17*, 450–461.

Tootell, R. B., Hadjikhani, N. K., Vanduffel, W., Liu, A. K., Mendola, J. D., Sereno, M. I., & Dale, A. M. (1998, Feb 3). Functional analysis of primary visual cortex (V1) in humans. *Proc Natl Acad Sci U S A, 95*(3), 811-817. https://doi.org/10.1073/pnas.95.3.811

Tseng, P., Chang, Y. T., Chang, C. F., Liang, W. K., & Juan, C. H. (2016, Aug 30). The critical role of phase difference in gamma oscillation within the temporoparietal network for binding visual working memory. *Sci Rep, 6*, 32138. https://doi.org/10.1038/srep32138

Tsuchiya, N., Wilke, M., Frassle, S., & Lamme, V. A. F. (2015, Dec). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends Cogn Sci, 19*(12), 757-770. https://doi.org/10.1016/j.tics.2015.10.002

van Gulick, R. (1989). What difference does consciousness make? *Philosophical Topics, 17*(1), 211-230.

Vandenbroucke, A. R. E., Fahrenfort, J. J., Meuwese, J. D. I., Scholte, H. S., & Lamme, V. A. F. (2016, Apr). Prior Knowledge about Objects Determines Neural Color Representation in Human Visual Cortex. *Cereb Cortex, 26*(4), 1401-1408. https://doi.org/10.1093/cercor/bhu224

Voelker, A. R., Blouw, P., Choo, X., Dumont, N. S., Stewart, T. C., & Eliasmith, C. (2021, Jul 26). Simulating and Predicting Dynamical Systems With Spatial Semantic Pointers. *Neural Comput, 33*(8), 2033-2067. https://doi.org/10.1162/neco_a_01410

von der Malsburg, C. (1995, Aug). Binding in models of perception and brain function. *Curr Opin Neurobiol, 5*(4), 520-526. https://doi.org/10.1016/0959-4388(95)80014-x

von Neumann, J. (1932). Mathematische Grundlagen der Quantenmechanik. *Berlin: Springer*.

Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017, Jan). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol, 13*(1), e1005324. https://doi.org/10.1371/journal.pcbi.1005324

Webb, T. W., & Graziano, M. S. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Front Psychol, 6*, 500. https://doi.org/10.3389/fpsyg.2015.00500

Webb, T. W., Kean, H. H., & Graziano, M. S. (2016, Jun). Effects of Awareness on the Control of Attention. *J Cogn Neurosci, 28*(6), 842-851. https://doi.org/10.1162/jocn_a_00931

Whittington, M. A., Cunningham, M. O., LeBeau, F. E. N., Racca, C., & Traub, R. D. (2010). Multiple origins of the cortical gamma rhythm. *Dev Neurobiol, 71*(1), 92-106. https://doi.org/10.1002/dneu.20814

Wilterson, A. I., Kemper, C. M., Kim, N., Webb, T. W., Reblando, A. M. W., & Graziano, M. S. A. (2020, Dec). Attention control and the attention schema theory of consciousness. *Prog Neurobiol, 195*, 101844. https://doi.org/10.1016/j.pneurobio.2020.101844

Yan, L. L., Xiong, T. P., Rehan, K., Zhou, F., Liang, D. F., Chen, L., Zhang, J. Q., Yang, W. L., Ma, Z. H., & Feng, M. (2018). Single-atom demonstration of the quantum Landauer principle. *Phys Rev Lett, 120*, 210601.

Zarnadze, S., Bauerle, P., Santos-Torres, J., Bohm, C., Schmitz, D., Geiger, J. R., Dugladze, T., & Gloveli, T. (2016, May 24). Cell-specific synaptic plasticity induced by network oscillations. *Elife, 5*. https://doi.org/10.7554/eLife.14912

Zhang, W., Itoh, K., Tanida, J., & Ichioka, Y. (1990, Nov 10). Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Appl Opt, 29*(32), 4790-4797. https://doi.org/10.1364/AO.29.004790

Zhu, X. H., Qiao, H., Du, F., Xiong, Q., Liu, X., Zhang, X., Ugurbil, K., & Chen, W. (2012). Quantitative imaging of energy expenditure in human brain. *Neuroimage, 60*(4), 2107-2117.