Consciousness and Cognition 30 (2014) 73–90

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Two theories of consciousness: Semantic pointer competition vs. information integration

Paul Thagard [a,*], Terrence C. Stewart [b]

[a] Department of Philosophy, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[b] Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Article history:
Received 27 November 2013

Keywords:
Binding
Competition
Consciousness
Information integration
Neural engineering framework
Neural mechanisms
Qualia
Representation
Semantic pointers

ABSTRACT

Consciousness results from three mechanisms: representation by firing patterns in neural populations, binding of representations into more complex representations called semantic pointers, and competition among semantic pointers to capture the most important aspects of an organism's current state. We contrast the semantic pointer competition (SPC) theory of consciousness with the hypothesis that consciousness is the capacity of a system to integrate information (IIT). We describe computer simulations to show that SPC surpasses IIT in providing better explanations of key aspects of consciousness: qualitative features, onset and cessation, shifts in experiences, differences in kinds across different organisms, unity and diversity, and storage and retrieval.

1. Introduction

Everyone has conscious experiences such as sensing pain, having to urinate, seeing blue, tasting chocolate, hearing music, touching wool, smelling grass, and feeling happy or sad. Consciousness also often accompanies high-level cognitive processes such as memory, learning, problem solving, decision making, and language use. Explaining consciousness is one of the most challenging problems in contemporary science, and only recently have neuroscientists dared to tackle it. The most audacious current proposal is Guilio Tononi's hypothesis that consciousness is the capacity of a system to integrate information. We will argue, however, that the information integration theory (IIT) faces serious mathematical and empirical problems.

As an alternative, we propose that consciousness is a neural process resulting from three mechanisms: representation by firing patterns in neural populations, binding of representations into more complex representations called semantic pointers, and competition among semantic pointers to capture the most important aspects of an organism's current state. Whereas IIT assumes that consciousness is a quantity that can be possessed by non-organisms such as photodiodes, countries, and the Internet, the theory of semantic pointer competition (SPC) restricts consciousness to organisms that possess sufficiently complex kinds of neural processes. We will show that SPC surpasses IIT in providing better explanations of key aspects of consciousness, including: qualitative features, onset and cessation, shifts in experiences, differences in kinds across different

---

[*] Corresponding author.
E-mail addresses: pthagard@uwaterloo.ca (P. Thagard), terry.stewart@gmail.com (T.C. Stewart).
URLs: http://cogsci.uwaterloo.ca/Biographies/pault.html (P. Thagard), http://terrystewart.ca (T.C. Stewart).

http://dx.doi.org/10.1016/j.concog.2014.07.001

organisms (e.g. self consciousness in humans versus mere feeling in simpler animals), unity and diversity, and storage and retrieval.

First we give general descriptions of IIT and SPC, leaving mathematical details for appendices. Then we outline how SPC explains key phenomena about consciousness using the three mechanisms of representation, binding, and competition. We argue that these explanations are far more empirically plausible and conceptually coherent than the ones provide by IIT. In order to substantiate the claim that SPC can rigorously explain the phenomena, we present computer simulations that show that the three mechanisms can approximate the relevant results. Finally, our general discussion addresses several issues relevant to evaluating IIT and SPC, including the possibility of consciousness in non-biological systems.

2. Two theories

2.1. Information integration theory

According to IIT, consciousness is integrated information generated by a complex of elements (Balduzzi & Tononi, 2009; Koch, 2012; Tononi, 2004, 2008, 2010, 2012; Tononi & Koch, 2008). In order to ensure that we are not misinterpreting IIT, we will report it using a series of quotes from Tononi's writings.

Tononi (2010, p. 299): "Consciousness has to do with a system's capacity for information integration. In this approach, every causal mechanism capable of choosing among alternatives generates information, and information is integrated to the extent that it is generated by a system above and beyond its parts. The set of integrated informational relationships generated by a complex of mechanisms – its quale – specify both the quantity and the quality of experience. As argued below, depending on the causal structure of a system, information integration can reach a maximum value at a particular spatial and temporal grain size. It is also argued that changes in information integration reflect a system's ability to match the causal structure of the world, both on the input and the output side."

Tononi, 2010, p. 300: "High *information* means that a system's causal mechanisms can specify precisely which out of a large repertoire of potential states could have caused its current state. High *integration* means that the information generated by the system as a whole is much higher than the information generated by its parts taken independently. In other words, integrated information reflects how much information a system's mechanisms generate above and beyond its parts."

Tononi, 2012, p. 172: "Integrated information measures how much can be distinguished by the whole above and beyond its parts, and Φ is its symbol."

Tononi, 2008, p. 224: "The IIT claims that, just as the *quantity* of consciousness generated by a complex of elements is determined by the amount of integrated information it generates above and beyond its parts, the *quality* of consciousness is determined by the set of all the informational relationships its mechanisms generate. That is, *how* integrated information is generated within a complex determines not only the amount of consciousness it has, but also what kind of consciousness."

Tononi, 2008, p. 233: "If consciousness is integrated information, then integrated information exists. Moreover, according to the IIT, it exists as a fundamental quantity—as fundamental as mass, charge, or energy."

Tononi, 2008, p. 236: "The IIT implies that many entities, as long as they include some functional mechanisms that can make choices between alternatives, have some degree of consciousness."

We find Tononi's mathematical definition of Φ hard to follow, but we will attempt to clarify and assess it in Appendix A. We take his basic claim to be that consciousness is a quantity, Φ, possessed by any system (regardless of its specific causal mechanisms) that is able to generate more information (in the technical sense of reduction of uncertainty) than is generated by its parts working independently. Then consciousness in humans, other organisms, and non-biological entities is explained by their having the capacity to generate integrated information.

According to Tononi, conscious experiences arise from an organism's ability to distinguish between stimuli. Simple entities like photodiodes can only discriminate between basic features like light and dark, so their consciousness is limited. In contrast, animals can see many features and integrate them into much richer experiences resulting from choices among alternatives. We find Tononi's explanations implausible for many reasons that will be provided after we outline an alternative theory of consciousness.

2.2. Semantic pointer competition

Our theory of consciousness employs Eliasmith (2013) new idea of semantic pointers, which are representations that can function as symbols while retaining connections to sensory and motor representations. We propose that consciousness results from formation of semantic pointers and competition among them. Semantic pointers have already been useful for explaining many psychological phenomena, including recognizing patterns, serial memory, controlling motor actions, and inference (Eliasmith, 2013; Eliasmith et al., 2012), behavioral priming (Schröder & Thagard, 2013, in press); intention (Schröder, Stewart, & Thagard, 2014), emotion (Schröder & Thagard, in press), creativity (Thagard, 2014a), and concepts (Blouw, Solodkin, Thagard, & Eliasmith, 2014; Thagard, 2012). We will state the SPC theory of consciousness as a concise set of hypotheses, expound those hypotheses in more detail, and then compare SPC with IIT.

SPC, the semantic pointer competition theory of consciousness, consists of the following hypotheses:

H1. Consciousness is a brain process resulting from neural mechanisms.

H2. The crucial mechanisms for consciousness are: representation by patterns of firing in neural populations, binding of these representations into semantic pointers, and competition among semantic pointers.

H3. Qualitative experiences result from the competition won by semantic pointers that unpack into neural representations of sensory, motor, emotional, and verbal activity.

H1 reveals a stark contrast between SPC and IIT. The restriction of consciousness to brains is not an a priori stipulation, but merely a recognition that all the entities so far known to be conscious have brains. Just as a theory in physics is not expected to explain all possible universes, and just as a biological theory is not expected to explain all possible life forms, so a theory of consciousness should not be expected to apply to all possible conscious entities. Explaining actual phenomena is difficult enough. Our evidence that people are conscious includes individual introspections and reports from other people of being conscious, as well as kinds of behavioral evidence such as shifts in attention that fit well with the view that people are conscious. The extension of attribution of consciousness to other animals who are not capable of self-reports is justified by their behavioral and anatomical similarity to people. For example, it is reasonable to infer that rats feel fear because their fear behaviors and neural circuitry for fear are similar to those in humans (LeDoux, 1996). Attributing even a little consciousness to entities like photodiodes and countries is dubious in the absence of any behavioral evidence or anatomical analogies.

Of course, it is possible that someday there will be robots or space aliens whose behavior, including self-reports, may provide some reason for supposing that non-humans also possess something like consciousness. But at that point there will need to be a serious investigation of whether consciousness in these entities really is much like consciousness in humans. Until then, a theory of consciousness need only apply to entities that actually display behaviors that are indicative of consciousness, which currently are just animals with sufficiently complex brains.

Hypotheses H2 breaks down into three claims about neural representation, semantic pointers, and competition. The first of these is relatively uncontroversial. Neural populations represent the world because neurons that interact with the world and each other become tuned to regularities in the environment (e. g. Dayan & Abbot, 2001; Eliasmith & Anderson, 2003; O'Reilly & Munakata, 2000). The second claim about semantic pointers is too new to have become controversial. A semantic pointer is a special kind of neural representation – pattern of firing in a population of neurons – that is capable of operating both as a symbol and as a compressed version of sensory and motor representations. For example, people's neural concept of chocolate unpacks (decompresses) into sensory representations of sweetness, texture, and so on, while allowing the semantic pointer to figure in inferences such as that you should not eat much chocolate because it is a kind of candy. Semantic pointers are formed by binding together simpler representations, where binding is a neural process that compresses information into a more compact form suitable for manipulation.

Fig. 1 pictures how binding takes several representations and combines them into a semantic pointer. Fig. 2 pictures the functioning of semantic pointers as symbols and their unpacking into the lower-level representations from which they were formed by binding. In simulated neural networks, binding can be performed by a mathematical procedure called circular convolution that is described in Appendix B. Semantic pointers are a specific proposal about representation of representations, which has been claimed to be important for explaining consciousness in terms of higher-order representations (Carruthers, 2011) and meta-representations (Cleeremans, 2011).

According to H2, the third mechanism of consciousness is competition among semantic pointers. Semantic pointers do not by themselves explain consciousness, because there are countless neural representations being formed all the time, most

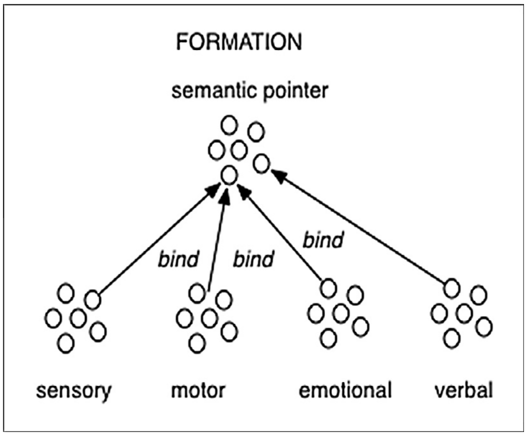

Fig. 1. Semantic pointers are neural representations (patterns of firing) formed by binding sensory, motor, emotional, and/or verbal representations, which are also patterns of firing in neural populations. The circles indicate representations accomplished by populations of thousands or millions of spiking neurons.

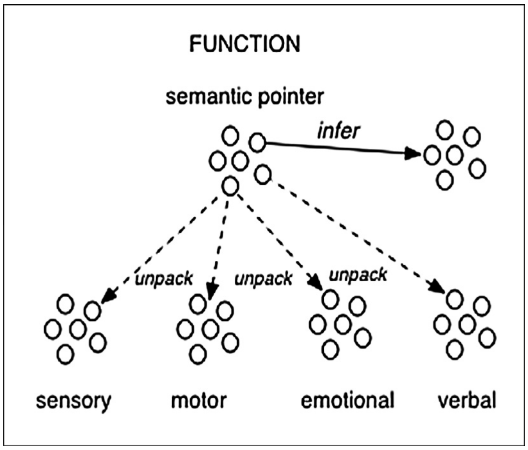

Fig. 2. Semantic pointers function to provide inferences by virtue of relations to other semantic pointers and can also unpack (decompress) into the sensory, motor, emotional, and/or verbal representations whose bindings formed the semantic pointer.

of which do not break through into consciousness. For example, context enables you to interpret the sentence "The pen is in the bank" as meaning that a writing device is in a financial institution rather than that a pig enclosure is by a river. Doing this requires the construction of various complex representations, only some of which become conscious. Attention is a psychological process that selects a small subset of candidate representations as worthy to enter consciousness (Braun, 2009). Many cognitive scientists have maintained that attention functions by means of competition among representations (e.g. Crick & Koch, 2003; Desimone & Duncan, 1995; Maia & Cleeremans, 2005; Slagter, Johnstone, Beets, & Davidson, 2010; Smith & Kosslyn, 2007). Selfridge (1958) explained pattern recognition as an interactive process among competing demons. We propose that the specific kind of representations engaged in the competition for attention are semantic pointers.

One key feature of our approach is that semantic pointers can compete with each other even when they are being represented by the same neurons, as shown in Fig. 3. That is, since a population of neurons forms a distributed representation of a concept (i.e. each concept is ideally some pattern of activity across all these neurons), those same neurons can represent a combination of semantic pointers via a pattern of activity that is a combination of the two patterns of activity for the two concepts. In order to have them compete, we add recurrent connections between all of the neurons. The weights of these connections are set using the Neural Engineering Framework, as described in Appendix B.

Hypothesis H3 proposes that qualitative experiences – all the sensations, feelings, and thoughts that people are aware of – result from competition among semantic pointers. How this works is described in more detail in the next section.

The differences between IIT and SPC are shown in Table 1. The two theories are similar in that they both propose that consciousness depends on combinations that require competitive choices. But SPC specifies neural mechanisms for combination and competition, including a particular kind of representation tied in part to sensory-motor processes.

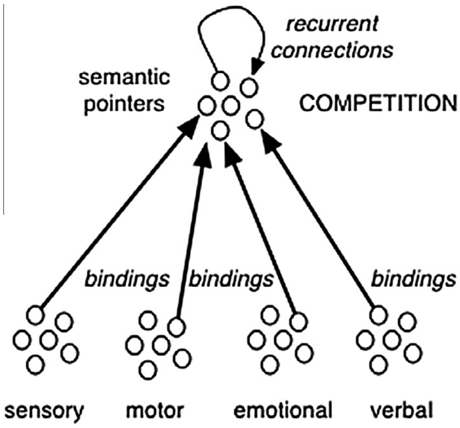

Fig. 3. Competition between semantic pointers. Two or more semantic pointers are produced in the same neural population by bindings of multimodal representations, shown by dark arrows to indicate that multiple semantic pointers are produced. Recurrent connections among the neurons in a population that encodes the semantic pointers generates competition among them.

Table 1
Differences between information integration and semantic pointer competition theories of consciousness.

	Information Integration	Semantic Pointer Competition
Consciousness is a:	Mathematical quantity, different from mass and energy	Neural process
Consciousness is a property of:	Many kinds of entities, including photodiodes and countries as well as brains	Organisms with complex neural networks
Mechanisms:	Various	Neural representation, binding into semantic pointers, and competition
Qualitative experiences are:	Shapes in a mathematical space	Neural processes dependent on specific kinds of semantic pointers

Although SPC challenges IIT, the idea that consciousness results from semantic pointer competition is compatible with theories according to which non-conscious processes operating in parallel compete for access to a global workspace that is distributed across many brain areas (Baars, 2005; Dehaene, 2009, 2014; Dehaene & Changeux, 2011). SPC is also compatible with the ideas of Damasio (2012) about how mental maps are formed in convergence-divergence zones in the brain. SPC goes beyond these neural theories in being more specific about the representations and mechanisms that constitute the global workspace. For further discussion of the relation between SPC and the idea of a neuronal workspace, see Section 5.

3. Competing explanations

Whether the IIT, SPC, or some other theory of consciousness is most plausible needs to be decided on the basis of which theory provides the best explanation of the full range of known facts about consciousness. The classes of facts to be explained include the qualitative experiences mentioned in hypothesis H3 and also the following: the onset and cessation of consciousness that occurs in sleep, anesthesia, injury, and seizures; attentional shifts in consciousness in response to external and internal stimuli; different kinds of consciousness, ranging from pain sensation in fish to self-consciousness in humans; and the unity and disunity of consciousness.

3.1. Qualitative experience

The biggest challenge for theories of consciousness is to explain the occurrence of subjective, qualitative experiences such as the pain of a headache, the taste of a beer, and the joy of a victory. Tononi follows philosophers in calling such experiences qualia (singular: quale). Balduzzi and Tononi (2009, p. 1) introduce the notion of a “quale as a shape that embodies the entire set of informational relationships generated by interactions in the system.” We propose instead that qualia are neural processes that result from the formation and competition of semantic pointers.

A theory of qualia must answer two basic questions (cf. Damasio, 2012, p. 269). 1. Why are there different qualitative experiences such as pains, tastes, sounds, and emotions? 2. Why do these experiences feel like anything at all? The first of these is answerable because of the different neural representations that get bound into a semantic pointer. Consider pain, which typically results from stimulation of sensory neurons (nociceptors) that respond to crushing, tearing, heat, cold, or chemical irritation. In humans, the conscious experience of pain usually arises from such stimulation plus bindings in the brain that integrate the nociceptive signals with locations, emotions, and self-conceptions, producing in humans complex semantic pointers like “I have an annoying pain in my toe.” Other perceptual experiences such as taste, touch, sound, hearing, smell, kinesthesia, balance, temperature, time, and bladder fullness result from different patterns of neural firing generated from different receptors that get combined into different semantic pointers. Perceptual experiences do not always result directly from perceptions, because the brain can store neural patterns and reactivate them in memory, imagination, and dreaming.

Emotions such as happiness, sadness, fear, and anger result in part from sensory inputs, especially ones for internal bodily states, but also from appraisals of the implications of the current state of an organism for its goals such as surviving and reproducing (Thagard & Aubie, 2008; Thagard & Schröder, in press). Semantic pointers bind together neural representations of a situation, physiological changes, and cognitive appraisals to produce a combined representation. Different combinations produce different qualitative experiences because of the different neural firings that contribute to the resulting semantic pointer. Situations can be represented purely perceptually, or by verbal sentences, which can be neurally constructed by the same kinds of binding operations that produce semantic pointers (Eliasmith, 2013; Eliasmith & Thagard, 2001). Semantic pointers explain why we can talk about our conscious experiences, because semantic pointers can function as symbols that can be bound into verbal reports such as “My toe hurts” or “I see a blue sky.” In non-human animals, however, semantic pointers can have symbolic functions in inference without being encoded verbally.

Semantic pointers provide an explanation of “why the neural basis of a conscious state with a specific conscious quality is the neural basis of that conscious quality rather than another” (Block, 2009, p. 1114). Specific conscious qualities result from specific patterns of neural firing and binding, making it easy to differentiate among the many different kinds of conscious experience.

Incorporating the self into representations of pain and emotions requires further binding that we discuss below in the section on kinds of consciousness. The question of what it is like to be conscious is too obscure to pursue, but differences in neural representations easily handle the more particular question of why there are differences in what it is like to be in pain, in love, and in the desert.

Binding of different kinds of inputs into semantic pointers explains the differences among various sensations and feelings, but why do these experiences feel like anything at all? The best available explanation is that in general qualitative experience is an emergent property of the three mechanisms that operate in organisms capable of neural representation, binding, and semantic pointer competition. There is nothing mysterious about emergent properties, which belong to wholes, do not belong to any of their parts, are not aggregates of properties of the parts, and result from the interactions of the parts (cf. Bunge, 2003; Wimsatt, 2007). Such emergence is common in complex systems (Findlay & Thagard, 2012), and representation, binding, and competition are each emergent properties of the interactions of many neurons, not the aggregate of bits of representation, binding, or competition that are properties of individual neurons. Then consciousness is an emergent property of the interaction of several mechanisms with emergent properties.

There are several available explanations of why consciousness has emerged. It may be an adaptation that increases the ability of organisms to survive and reproduce (like emotions), a side effect of other processes that are adaptive (like mathematical ability), a chance result (like diseases caused by mutations), or a cultural artifact (like rock music). Given the apparent commonality of consciousness in humans and other animals, the chance and cultural explanations are implausible, but we see no evidence that currently decides between adaptation and side effects explanations. It is possible that consciousness makes attention, learning, and action more effective (Damasio & Carvalho, 2013), but it also possible that consciousness resulted as a side effect from sufficiently complex processes of neural representation, binding, and semantic pointer competition. Saying that conscious experience is emergent does not imply that it is epiphenomenal, with no causal role, because the neural processes of consciousness can sometimes have causal effects on action, in accord with experimental findings (Baumeister, Masicampo, & Vohs, 2011). Conscious experience can be stored in a semantic pointer, enabling it to have an ongoing causal role in mental processes such as recollection, reenactment, imagination, and empathy.

Our explanation of qualitative experience is very different from Tononi's. We view qualia as tied to particular neural mechanisms, not a completely abstract mathematical shape in an unspecified information space. Our account can explain both the differences among various experiences and also how they emerge. Moreover, this account can be stated sufficiently exactly to inspire computer simulations of different sorts of mental states that generate different sensations and feelings, described in section 4.

3.2. Onset and cessation of consciousness

There are more straightforward facts than qualia that must be explained by any plausible theory of consciousness. Humans and other animals experience the cessation and onset of consciousness every day when they go to sleep and wake up. Cessation of consciousness can also result from unusual occurrences such as anesthetics, concussions, strokes, seizures, and death. Tononi vaguely associates these changes with information integration, but we now propose much more specific mechanisms that produce termination and revival of consciousness.

Most kinds of consciousness cessation involve increases in neural inhibition, decreases in neural excitation, or both, interfering with processes of representing, binding, and competition that all depend on a balance among excitation and inhibition. In sleep, the accumulation of the neurotransmitter adenosine (which can be blocked by caffeine) acts as a direct negative-feedback inhibitor of neuronal activity (Porkka-Heiskanen & Kalinchuk, 2011; Steriade & McCarley, 2005). Anesthetics work by enhancing inhibitory transmission and/or blocking excitatory conduction (Villars, Kanusky, & Dougherty, 2004). Concussions cause many kinds of metabolic disturbances including lactate buildup that hinders neurons (Tripoli & Torg, 2011). Death brings total elimination of neural firing and hence the end of consciousness, and less severe events such as strokes and heart attacks can also dramatically reduce excitation. In sum, decreases in excitation and/or increases in inhibition produce disruptions in representation, binding, and competition that remove consciousness in sleep, anesthesia, injury, disease, and death.

It is easy to see how the mechanisms of SPC can explain loss of consciousness from increases in neural inhibition and decreases in excitation. If neural firing decreases, there will be less formation of representations and less binding of representations into semantic pointers. Moreover, in the semantic pointer competition, there will be fewer semantic pointers that achieve a minimal threshold required for conscious experience. Once loss of consciousness occurs because of sleep, anesthetics, or concussion, a drop in sensory input further decreases neural activity required for the formation, binding, and competition of representations. Dreaming revises consciousness by replacing sensory inputs by internally generated neural activity arising from memory.

Seizures and epilepsy are more complicated in that they seem to involve hyperactivity of neurons rather than inhibition. (Bromfeld, Cavaso, & Sirven, 2006). How can an excess of excitation disrupt representation, binding, and competition to an extent that eliminates consciousness? We conjecture that, when undampened excitation floods a neural network, it saturates the network of competing semantic pointers so that none are able to exceed a threshold that corresponds to entry into consciousness.

The operation of a threshold explains why neural firing, binding, and competition among semantic pointers can all operate in unconscious processes. Section 4.3 describes a simulation in which a lack of excitation or excess of inhibition can keep neural firing below the threshold that is required for conscious experience.

In contrast to our mechanistic explanation of the onset and cessation of consciousness, Tononi associates phenomena such as sleep and seizures with loss of information integration. Such loss, however, is more plausibly an effect rather than a cause of the cessation of consciousness that results from increased neural inhibition, decreased excitation, or hyperexcitability. For a causal explanation of why fatigue, drugs, concussions, seizures, and death can produce unconsciousness, IIT needs to provide a precise mathematical account of how changes in inhibition and excitation reduce Φ.

3.3. Shifts in consciousness

As people go through their days, their consciousness naturally moves among many different kinds of experiences, for example when a doorbell rings to shift attention away from working. SPC inherits from other competition theories of attention the natural ability to explain these shifts. At one moment, some representations are winning the battle for the limited resources of consciousness, but incursion of a new stimuli such as the doorbell has two effects. First, new semantic pointers are formed, binding the stimuli together and connecting them with other representations such as the verbal representation that a parcel is being delivered and the emotion of excitement. Second, if the resulting semantic pointers develop more powerful patterns of firing than those already dominating consciousness, then the new semantic pointers will suppress and replace the old ones, unless there is sufficient mental capacity that the new and old can coexist.

How such competition works is most clear in simple, localist neural networks that associate one artificial neuron with one representation. If a neuron for doorbell receives activation from input neurons and thereby inhibits a neuron for coffee, then attention shifts from coffee to neuron. The situation is much more complicated in more biologically realistic networks where doorbell and coffee are each represented by patterns of firing in large numbers of neurons with a great many excitatory and inhibitory connections. Nevertheless, the neural networks described in our simulations in the next section can accomplish the more complicated kind of activation and suppression needed to produce competition between representations. Moreover, the competing semantic pointers have the huge advantage over localist representations that they point to sensory, motor, emotional, and verbal information by virtue of their unpacking functions shown in Fig. 2. Localist models cannot explain why attentional shifts bring different kinds of experience because they lack the connections to sensory and motor processes that are integral to semantic pointers. In contrast, semantic pointer competition can readily explain shifts in conscious experiences of different kinds, an important phenomenon not addressed by information integration. In the terminology of Block (2005), SPC can explain both access consciousness (attention that results from competition) and phenomenal consciousness (experience that results from the binding of sensory representations).

Competition among semantic pointers is useful for explaining many important psychological phenomena such as binocular rivalry and inattentional blindness. Binocular rivalry occurs when different visual images are presented to each eye and the brain fails to integrate them, so that the brain switches between the interpretations suggested by each eye as the result of competition between incompatible construals. Inattentional blindness occurs when people fail to notice obvious events as in the famous invisible gorilla experiment of Chabris and Simons (2010). We conjecture that such failures result from the ignored events being outcompeted by other representations encoded as semantic pointers.

3.4. Kinds of consciousness

Although the semantic pointer approach rejects the idea that consciousness comes in many grades of information integration, it can accommodate the view that there are various kinds of consciousness associated with organisms of different levels of neural complexity. For example, Damasio (2012, p. 179) distinguishes between (1) a minimal kind of consciousness that involves a sense of the here and now, and (2) an extended sense that includes autobiographical representations of self. We propose that different kinds of consciousness all require neural representation and semantic pointer competition, but differ with respect to the extent of binding.

First, consider fish, the simplest organism for which behavioral evidence currently supports the attribution of consciousness. Braithwaite (2010) argues that it is plausible that fish feel pain because their anatomy includes detectors for tissue damage whose stimulation is transmitted to the trigeminal nerve and alters behavior such as avoiding novel objects. Such alteration of behavior can be eliminated by a dose of morphine that affects the fish's opioid system. It therefore seems that the attention of fish is affected by pain, just like the attention of humans, suggesting that for fish pain is a conscious experience. SPC can explain these results by supposing that the fish have a representation of pain that competes with their representations of novel objects such as a tower. Fish do not have a lot of neurons to work with, perhaps only a million, so we expect that are they are unable to produce very complicated semantic pointers and incapable of holding many in consciousness. Nevertheless, the pain caused by a noxious substance such as vinegar is enough to distract them, even though their semantic pointers may be built out of no more complex representations than "pain" or "new object".

Mammals have a lot more neurons, approximately 50 million in a rat and hundreds of millions in a cat, enabling them to produce much more complex semantic pointers. A rat may be able to bind together complexes of sensory representations for smell, taste, and pain, enabling it to perform one-trial learning as occurs in taste aversion resulting from radiation. Mammals show many of the same pain behaviors as humans, such as eye closing, aggression, and writhing.

Finally, *self*-consciousness requires millions or billions of neurons to enable representations of self. One behavioral method for determining whether an organism has self-consciousness is to see whether it can detect itself in a mirror, a test so far only passed by eight species: humans, gorillas, chimpanzees, orangutans, elephants, dolphins, pigs, and magpies (Broom, Sena, & Moynihan, 2009; Prior, Schwarz, & Güntürkün, 2008). These organisms have enough neurons and neural processing to produce bindings that include the self. For example, the semantic pointer representing that you have an unpleasant pain in your toe can require binding together representations of pain, toe, unpleasantness, and yourself. Unpleasantness is an emotion that binds physiological information with cognitive appraisals, and the representation of yourself binds representations of your body, your memories, self-concepts, and current experiences (Thagard & Schröder, in press). The self can be viewed as a multilevel system of social, psychological, neural, and molecular mechanisms (Thagard, 2014c).

Hence SPC can easily explain how consciousness operates at different levels of complexity corresponding to different capacities for producing semantic pointers by binding. Information integration is at best a consequence of these capacities, not a factor that explains why consciousness has qualitatively different kinds.

3.5. Unity and disunity of consciousness

Since Immanuel Kant (1929), attempts have been made to explain the unity of consciousness, for example when drinking a cup of coffee produces an integrated experience of taste, smell, touch, and vision. Tononi (2004) thinks that unity results from information integration, but we maintain that it results from mechanisms of binding and competition which can also explain the frequent disunity or disjointedness of competition. As Brook and Raymont (2010) summarize, there are disorders that disrupt the unity of consciousness such as hemi-neglect, asognosia, and schizophrenia. Even in everyday life, consciousness can be fragmented when one is attending to multiple stimuli, so a theory of consciousness needs to be able to explain disunity as well as unity.

According to SPC, unity of consciousness results from the binding of representations into semantic pointers. Consciousness overcomes confusion of unconnected experiences because neural mechanisms bind separate perceptual representations into ones that combine different modalities. For example, representation of a cup of coffee can consist of a pattern of firing in a neural population that draws on patterns of firing for coffee's color, taste, and smell, and feel, and also the weight of the cup, the pleasure associated with drinking the coffee, and even the representation of the self as the owner of all these experiences.

Although a high degree of unity is a common property of consciousness, SPC can also explain why consciousness is sometimes fragmented. Binding usually produces multiple semantic pointers that compete for consciousness, but that process need not result in exactly one winner. Semantic pointer competition, like lotteries, can have multiple winners, with several unified representations entering consciousness. For example, you can be conscious of your cup of coffee, your computer screen, and radio music all playing at once. The computational model described in the next section is not a winner-take-all process, but rather one that allows for a small number of semantic pointers to outcompete other representations to surpass a threshold. Hence consciousness is sometimes disorganized even in normal brains, and can become even more so when brain functioning is disrupted by drugs or disease. For example, drugs like LSD (which disrupts serotonin) and diseases like schizophrenia (with excessive dopamine activity) interfere with normal binding and competition operations and hence produce abnormally disorganized conscious experiences. Hence for SPC, the unity and disunity of consciousness result from the functioning and malfunctioning of specific brain mechanisms, not from abstract quantities of information integration.

It remains an open question why human consciousness is limited to a relatively small number of representations, perhaps even fewer than George Miller's (1956) magical number seven plus or minus two. The small capacity of consciousness may be an adaptive feature that encourages effective action, but it may also be an unfortunate bug arising as a side effect of insufficient numbers of neurons to produce large numbers of simultaneously active semantic pointers. In our simulations, the mechanisms of neural representation, binding into semantic pointers, and semantic pointer competition all require large numbers of neurons. Forming more complex semantic pointers by recursive binding and allowing more of them to enter consciousness would require more neurons. Why has the brain not evolved to produce richer kinds of representations, for example relations among more than just three or four objects, and to allow more representations to be active in consciousness? The answer probably involves many factors, such as limitations on metabolism resulting from the energy needed to run many neurons, limitations on head size resulting from the challenges of childbirth, and lack of advantage for survival and reproduction of larger brains during the period around 100,000 years ago when modern humans evolved.

Tononi (e.g. 2004, p.20) marks as one of the advantages of IIT that it provides necessary and sufficient conditions that determine the quantity of consciousness generated by a system. Even if his definition of Φ can overcome the mathematical problems described in Appendix A, we view the enterprise of defining quantity of consciousness as fundamentally misguided. Compare the enterprise of trying to define quantity of *life*. There are many different kinds of life, ranging from single-celled bacteria to plants to insects to fish to reptiles to mammals. It would be pointless to try to assign a number to how much life each of these kinds of living entities possesses. Rather, the explanation of life relies on the discoveries in the past hundred years of key mechanisms for producing life, including genetics, metabolism, motion, signaling, and reproduction. Life results from the interaction of multiple mechanisms, not from some mysterious quantity. Analogously, a theory of consciousness should attempt to identify the mechanisms that produce it rather than concoct an abstract quantity.

3.6. Storage and retrieval

Many conscious experiences result from memory, when storage of an event leads later to recall and a similar experience. Hence semantic pointers should be capable of storage and retrieval, including filling in of missing information. Showing this capability, as well the other aspects of consciousness just described, can best be performed by computer simulations.

4. Simulations

In cognitive science, computer simulations play several valuable roles (Thagard, 2012, ch. 1). They show that hypotheses can be made mathematically precise enough to be simulated, and the simulations ensure a basic level of computational tractability. Most importantly, successful simulations show a principled connection between hypothetical mechanisms and facts to be explained, when the program that implements the mechanisms performs in the same way as people. We will first describe how the three mechanisms of neural representation, binding into semantic pointers, and semantic pointer computation are computationally implemented. Then we will present six simulations corresponding to the explanations of six kinds of evidence mentioned in the last section.

4.1. Mechanisms

In accord with standard practice (e.g. Rumelhart & McClelland, 1986), our neural representations are distributed, in that many neurons are required for each mental representation such as a concept. Specifically, we produce distributed representations using the encoding techniques of the Neural Engineering Framework (Eliasmith, 2013; Eliasmith & Anderson, 2003). The neural simulator Nengo makes it easy to define populations of neurons in terms of what they represent and then form connections between neural populations in terms of the computations to be performed on these representations (http://www.nengo.ca). All the simulations described below are run in Nengo; for the underlying mathematics, see Appendix B. We interpret a concept as a pattern of firing in a population of neurons that results from the synaptic connections (excitatory and inhibitory links) between the neurons.

Nengo is able to take multiple concepts and bind them together into semantic pointers by means of neural computations. Concepts as patterns cannot simply be added together to get a combined pattern. That is, if the concept DOG has a verbal representation and the concept BLACK has a sensory representation, then we cannot capture the combination by building an additive representation BLACK + DOG, which could not tell whether it was a sensory BLACK and a verbal DOG or vice versa. Instead, we need the neurons to compute a new pattern for this unique combination. Furthermore, the neurons must be able to create these novel combined patterns and extract out the original patterns for completely novel combinations of stimuli without changing the connections between the neurons. People are capable of experiencing completely novel combinations of stimuli in hundreds of milliseconds, much faster than neurons can form new connections.

The solution to this problem is to build a neural network that can *bind* two patterns together (Stewart & Eliasmith, 2012; Thagard & Stewart, 2011). For binding, we use circular convolution, a mathematical operation that takes in two patterns and produces a third, novel pattern (Eliasmith & Thagard, 2001; Plate, 2003). This function is approximately invertible: given the novel pattern and one of the two original patterns, the other pattern can be recovered. Using techniques from the Neural Engineering Framework, the simulator Nengo implements convolution in biologically realistic spiking neurons. The neurons receiving the verbal pattern *X* are set up to produce the pattern for VERBAL*X, where * is circular convolution. The same production happens for the sensory neurons. If the verbal system represents DOG and the sensory system represents BLACK, then the resulting semantic pointer will be VERBAL*DOG + SENSORY*BLACK. Importantly, this combination of convolution and addition will work even if the system has never seen that arrangement of inputs before.

Our simulations use convolution as a binding mechanism rather than neural synchrony which has also been proposed to explain binding in consciousness (e.g. Crick, 1994; Engel, Fries, König, Brecht, & Singer, 1999; Goldfarb & Treisman, 2013; Prinz, 2012). Arguments that convolution has computational and empirical advantages over synchrony have been provided elsewhere (Stewart & Eliasmith, 2012; Thagard & Stewart, 2011; see also Canales, Gomez, & Maffet, 2007). Synchrony accounts do not provide a plausible neural mechanism for maintaining or changing bound representations, and they lack an account of how attention works.

The final mechanism we require is *competition* among semantic pointers, which applies to incompatible concepts. For example, we may want VERBAL*DOG to be positively associated with SENSORY*DOG (since they tend to occur together), but negatively associated with SENSORY*CAT. That is, whenever we have the VERBAL*DOG pattern in the semantic pointer, we want it to suppress any patterns that look like SENSORY*CAT, while enhancing any patterns that look like SENSORY*DOG. Suppression and enhancement are implemented by recurrent connections among the neurons in the semantic pointers that correspond to each concept, providing a mechanism for shifts in attention. The connection strengths for these neurons are computed using the Neural Engineering Framework by optimizing the function that enhances and suppresses the set of patterns that are in competition. The result is a distributed model that that behaves like a localist network in which competition between nodes is implemented by symmetric inhibitory links between them. Implemented in Nengo, the mechanisms of competition, binding, and representation suffice to simulate many interesting aspects of consciousness.

4.2. Simulation 1: Qualitative differences

The first simulation shows that the model is capable of binding together a wide variety of completely different inputs, including novel combinations that have never been experienced before. Fig. 4 shows a neural system where the patterns stored in two different brain areas, carrying sensory and verbal information, are combined into a single semantic pointer. This is done by a set of neural connections, found using the Neural Engineering Framework, that closely approximate the binding operation of circular convolution for any possible input pattern. To show that the pattern is successfully stored, we also include connections to extract out the original pattern from the semantic pointer. To demonstrate this system's capabilities, we present a sequence of novel inputs, using randomly generated patterns that the model has never seen before. These new patterns are successfully bound together and extracted, showing that it works with any qualitatively new situation.

The graph in Fig. 4 shows that the extracted patterns are not exactly the same as the original inputs, but are similar to the original patterns. To evaluate the capacity of this model to handle a larger number of different experiences, we can measure this similarity and compute how likely the extraction is to be correct. A correct extracted pattern must be more similar to the original pattern than to any other known pattern. As shown by Plate (2003), the memory capacity of these systems increases exponentially, and this model is capable of correctly extracting patterns with a million different sensory and verbal patterns (with 95% accuracy). Because any sensory input can be paired with any verbal input, the model can handle a trillion qualitatively different combinations. Even more combinations are possible if there are multiple sensory regions for vision, audition, emotion, and so on. Hence our Nengo simulation based on the Neural Engineering Framework can accommodate the broad range of qualitatively different states that occur in human consciousness.

4.3. Simulation 2: Onset and cessation

We described in Section 3.2 how loss of consciousness is associated with inhibition of neural firing and the blocking of neural excitation. We model this association by adding a population of neurons that inhibits semantic pointers, not the whole brain. By selectively inhibiting the neural populations devoted to binding different stimuli, the system loses the ability to extract individual patterns from the semantic pointer. Fig. 5 shows the model from Fig. 4 with the addition of this inhibitory population. As the inhibitory population increases its activity, the extracted patterns are diminished. Thus around .3 s, the system becomes unconscious of the meowing cat because the semantic pointer is deactivated. When inhibition is diminished (or excitation is increased), then consciousness can be regained, as for the mooing cow around .5 s.

4.4. Simulation 3: Shifts of consciousness

To model shifts of consciousness, we need a method for incompatible neural patterns to compete with each other. Given an ambiguous input that is the combination of two patterns by simple addition, the pattern present in a semantic pointer should bias the system's interpretation of the pattern. For example, if the ambiguous pattern for DOG + CAT is presented to a system that is has just recently been given the pattern for DOG, then the extracted pattern should remain DOG. If the CAT pattern is increased in strength, then the extracted pattern should suddenly shift to CAT.

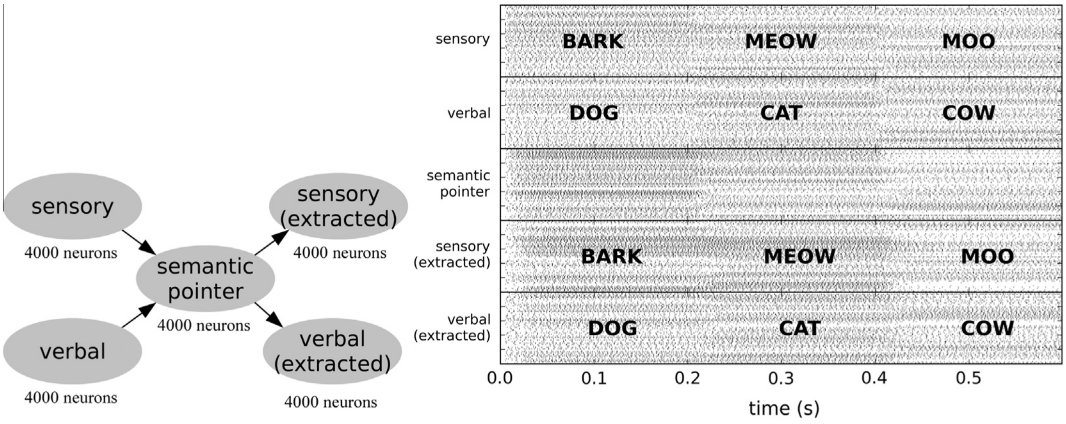

Fig. 4. A neural model capable of representing novel combinations of qualitatively different stimuli. The graph shows spiking behavior of each neural population as different patterns are input to the sensory and verbal neurons. The text labels indicate the pattern that most closely matches the neural firing of that group of neurons. The graph on the right shows the firing patterns of large groups of neurons, with a line for each neuron and a mark indicating that the neuron is firing. Hence darker shading indicates that more neurons are firing as the system is serially presented with a barking dog, a meowing cat, and a mooing cow.

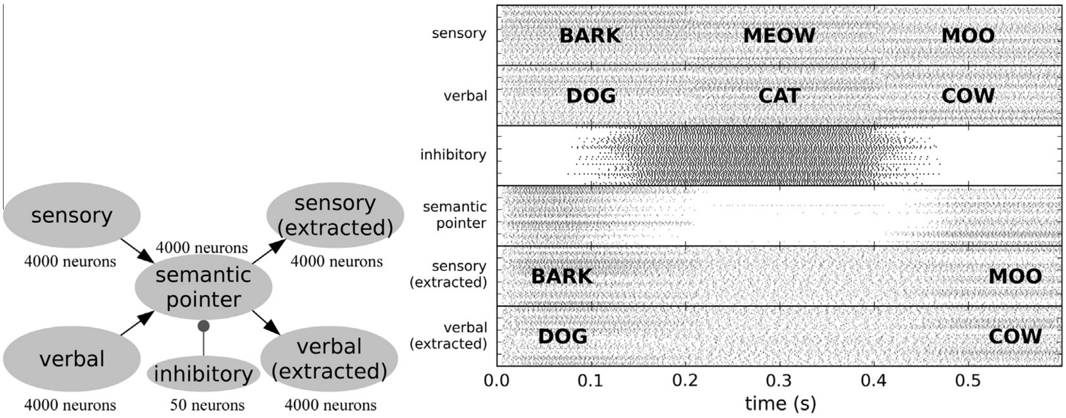

Fig. 5. A neural model of cessation and onset of consciousness. The firing of the inhibitory population disrupts the formation of the semantic pointer, making the extracted patterns disappear. Darkness of shading displays firing of neurons.

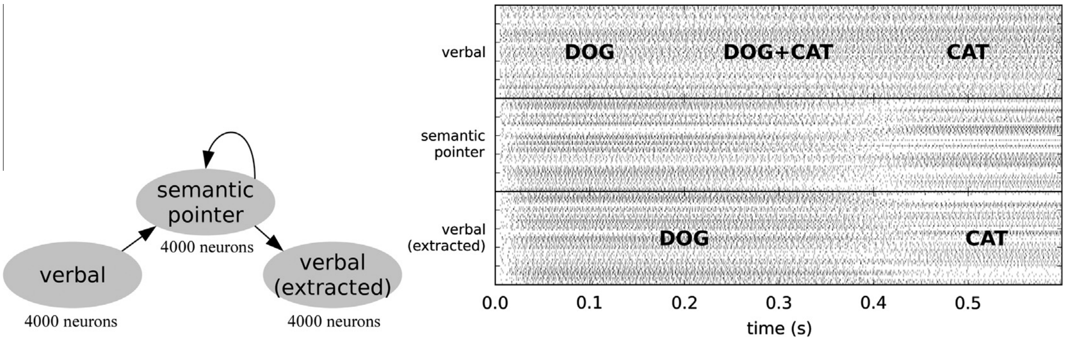

Fig. 6. A neural model showing shifts of consciousness. The pattern for DOG and CAT are both present in the verbal neurons, at differing degrees of strength. Initially the DOG pattern is very strong and the CAT pattern is very weak, and the extracted pattern is DOG. The strength of the CAT pattern is increased gradually, causing the extracted pattern to suddenly switch to CAT near time = 0.4 s.

Fig. 6 shows the addition to the computational model needed to achieve this shift: a recurrent connection from the neurons representing the semantic pointer back to themselves. The Neural Engineering Framework is used to find the connection weights such that the CAT pattern will cause a decrease in strength for the DOG pattern, and vice versa. The result is the shift in activation shown in Fig. 5, with the neural system shifting from the DOG interpretation to the CAT interpretation as the result of different verbal inputs. Hence semantic pointer competition can explain shifts in consciousness.

4.5. Simulation 4: Kinds of consciousness

Modeling different kinds of consciousness depends on the ability of the neural system to scale up to many different kinds of representations, including representations of self. We have not developed a full model of the neural representation of self-knowledge, which would have to cover more than eighty different phenomena (Thagard, 2014c). For example, in Fig. 7 we display a model that combines verbal and sensory information with emotional and self-knowledge information. For the emotional and self-knowledge information, the input patterns are not randomly generated, but are built up using the same process as we use to make one, global semantic pointer. We construct the emotional representation from basic emotional components as in our model of intention (Schröder, Stewart, and Thagard, 2014). As a placeholder for a more complete theory of self-representation, this model also includes location and vocation information as part of its "self" input. The system works equally well for combining any information that can be represented as spike patterns. Because the inputs to a semantic pointer are anything that can be represented by spiking patterns in neurons, and because the semantic pointer itself is a spiking pattern in neurons, it too can be treated as another input. This nesting allows for representations about previously bound pattern combinations, including representations of the self.

4.6. Simulation 5: Unity and disunity

Next, we can use the same competition system used to do shifts in consciousness (simulation 3) to form consistent relationships across different input patterns. In addition to using the competition mechanism to cause one pattern to inhibit

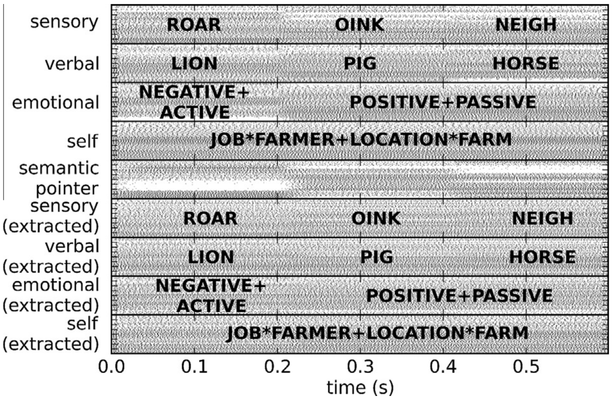

Fig. 7. A neural model showing a wider variety of representations being bound together. Multiple patterns can be combined, and the input patterns can themselves be semantic pointers.

another, we can also use it to cause one pattern to enhance another. Because this recurrent connection occurs in the semantic pointer neurons, we can have the verbal pattern DOG enhance a consistent sensory pattern such as BARK. This interaction makes the system more robust to noisy inputs. When presented with inputs that do not go together (such as DOG and MEOW), accuracy is reduced. If the inputs are not extremely strong patterns, one or the other may be ignored. Hence our computational model can cover both unity, when mutually enhancing semantic pointers are activated together, and disunity, when competition prevents coherence among the most active semantic pointers. It would be interesting to model the disunity of consciousness that occurs in schizophrenia and hallucinogenic drug use, presumably because of excess excitation or lack of inhibition.

Fig. 8 shows that unity allows different aspects of a semantic pointer to influence each other. Here, we form recurrent connections for the semantic pointer in such a way as to unify CAT and MEOW. That is, when the verbal input of CAT comes in at the same time as a weak version of the BARK pattern, the combined representation will actually be CAT and MEOW. This embedding of associations between elements of a semantic pointer allows for complex filling in of related elements.

4.7. Simulation 6: Storage and retrieval

Finally, we demonstrate that semantic pointers can be stored and reactivated, supporting conscious experience resulting from memory. By combining the separate representations from different modalities into one pattern of neural activity, we can store that information and recall it after the event has passed. This activity can be accomplished by connecting the neurons representing the semantic pointer back to themselves, using the NEF to compute the weights that will cause the neurons to maintain the activity pattern with which they are currently firing. This computation of the simple identity function $f(x) = x$ is the basis of the working memory model in Spaun (Eliasmith et al., 2012).

Fig. 9a shows this model when it is presented with an input of the pattern for the first 0.2 s. As in the first simulation (Fig. 4), the network can combine these into a single semantic pointer, and then accurately extract that information.

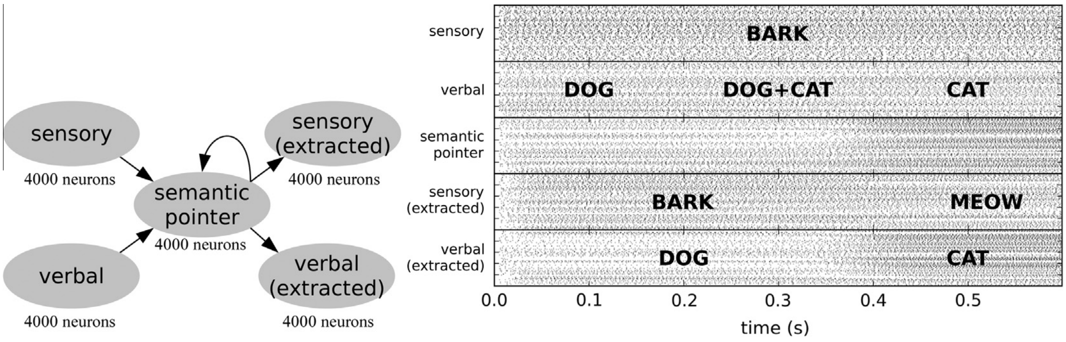

Fig. 8. A neural model showing interaction between inputs, encouraging unified patterns and only allowing conflicting patterns if they are extremely strong (or unrelated). Here, the sensory input is a weak version of the BARK pattern while the verbal input changes smoothly from DOG to CAT. The output for the first 0.4 s remains as BARK and DOG, indicating that the weak CAT input is successfully being ignored. However, as the CAT verbal input becomes stronger, it suddenly wins over the DOG pattern around $t = 0.4$. At the same time, the strong unification between a CAT and MEOW put enough influence on the semantic pointer that MEOW is extracted, even though BARK was input. Thus, one aspect of the semantic pointer can influence another.

However, in this case, after 0.2 s has elapsed, all input to the model is removed, leaving the input neurons firing with their background activity. However, the neurons in the semantic pointer population maintain their firing sufficiently to extract the information after the stimulus has been removed.

Furthermore, this reactivation process can be used to intensify the representation, adding new features to the representation using the same mechanism in the previous simulation (Fig. 8). The idea is similar to that seen in a standard Hopfield network (Hopfield, 1982): the neurons are connected such that if they are presented with a pattern that is similar to one they have encountered before, then the neural activity will change to become like that previous pattern. In the case of semantic pointers, this means that the stored representation will be "filled in" based on the typical features encountered by the system. Fig. 9b shows the same model from Fig. 9a presented with just a portion of the previous sensory stimulus. It not only stores that representation but also fills in the missing part.

We have described six simulations that show the capacity of the theory of semantic pointer competition to explain qualitative difference in consciousness, onset and cessation, shifts in consciousness, kinds of consciousness, unity and disunity, and storage and retrieval. Ideally, it would be useful to produce a computational model of IIT to allow a detailed comparison with our SPC simulations. Appendix A describes mathematical and computational problems with IIT that make it difficult to simulate.

5. Discussion

Tononi deserves commendation for offering a bold hypothesis concerning the nature of consciousness, but his hypothesis needs to be evaluated according to whether it provides better explanations of the evidence than alternative theories. We have proposed a very different theory according to which consciousness is a neural process resulting from three specific mechanisms, not an abstract quantity. We have shown how neural representation, binding into semantic pointers, and semantic pointer computation can all be stated sufficiently precisely to be implemented in a computer model that approximates the performance of people on tasks relevant to consciousness.

In the nineteenth century, before biological mechanisms such as genetics, cell metabolism, and cell division were understood, it was natural to hypothesize a vital force that distinguished living things from non-living ones. This hypothesis became superfluous with the realization that biological mechanisms arise from chemical and physical processes. Similarly, consciousness does not need to be treated as a special kind of entity in addition to matter and energy, but rather can be explained as a biological process fully compatible with explanations in physics and chemistry. The information integration theory of consciousness is ontologically less parsimonious than neural theories of consciousness, because it postulates consciousness as a different kind of entity from those used in the rest of science (compare Chalmers, 1996). Such postulation would be justified if there were facts that could not be explained by conventional means, but we have shown that biological mechanisms can explain the most important aspects of consciousness, making information integration ontologically excessive. Moreover, IIT's idiosyncratic attribution of consciousness to objects like photodiodes and countries that have no behavioral or anatomical signs of consciousness is unwarranted given the availability of alternative neural theories that apply well to entities from fish to humans that actually behave in ways consistent with consciousness. Unlike Tononi's (2010, p. 300) attempt to use a thought experiment to establish the "essence" of consciousness, we eschew thought experiments as frequently misleading about the nature of mental processes (Thagard, 2014d).

Perhaps Tononi would contend that semantic pointer competition is just a special instance of a causal mechanism capable of integrating information by choosing among alternatives, so that IIT is more general than SPC. However, in the absence of empirical facts explained by IIT but not by SPC, we see IIT as eliminable.

The explanations provided by the semantic pointer competition theory are mechanistic in that they concern a system of parts whose interactions produce regular changes (Bechtel, 2008; Craver & Darden, 2013; Findlay & Thagard, 2012). In

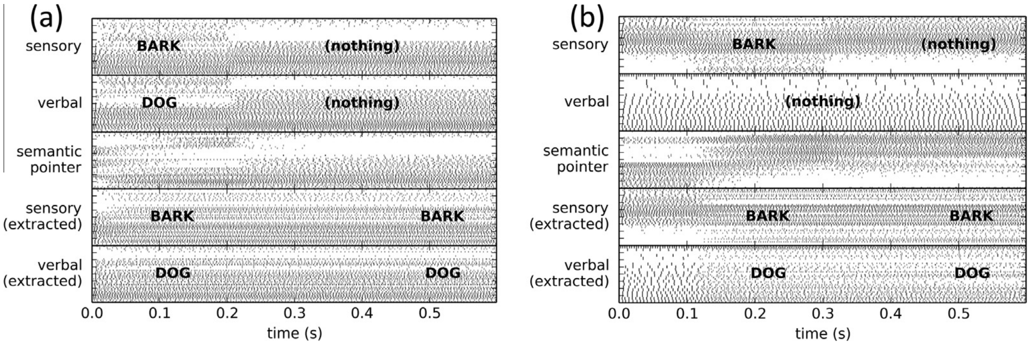

Fig. 9. A neural model capable of storing combined representations over time, and of retrieving previously stored representations. In (a) the patterns for BARK and DOG are presented for only the first 0.2 s. However, the system is able to successfully extract those patterns after they have been removed. In (b) the simulation is only given BARK as input. Again it can store and retrieve BARK after it has been removed, but it is also able to retrieve DOG as well, since it has been associated with BARK in the past.

particular, brain mechanisms involve systems of neurons with complex interactions that are capable of producing representations, bindings, and competition. Our enterprise is consistent with experimental attempts to find neural correlates of consciousness (e.g. Haynes, 2009), but is more theoretical in that we want to explain those correlations, not just describe them. We view it as a virtue rather than a defect that our account of consciousness applies just to biological organisms with complex brains rather than to imaginary entities such as future computers and space aliens. We are prepared to revise our theory if such entities appear with anything like human consciousness.

Dehaene (2014) has presented an impressive array of evidence to support his view that consciousness is the global availability of information encoded and broadcast in a neuronal workspace. This view appears complementary with our SPC theory, which could be expanded to explain how information is broadcast across numerous brain areas. SPC goes beyond Dehaene's account in specifying mechanisms for encoding information as semantic pointers and for using competition to determine what information is important enough to broadcast.

This paper has not addressed philosophical arguments that neural theories of consciousness are necessarily wrong (for responses, see Thagard, 2010, 2014b, 2014d). Nor have we attempted to give necessary and sufficient conditions for consciousness, because our aim is for a mechanistic theory that explains the facts of consciousness, not a definition or conceptual analysis. Sensibly, biologists have not attempted to state a set of necessary and sufficient conditions for life, but instead have developed richly explanatory theories of many aspects of life. Different kinds of life employ different kinds of important mechanisms, for example photosynthesis in plants and locomotion in animals. The scientific task is not to define consciousness, or merely to discover its neural correlates, but rather to discern the most important mechanisms for producing it.

6. Conclusion

To sum up, the semantic pointer competition theory of consciousness is superior to the information integration theory in the following respects. First, it provides more detailed explanations of a broad range of important facts about consciousness, including qualitative experience, onset and cessation of consciousness, shifts in consciousness, kinds of consciousness, unity and disunity, and storage and retrieval. Of course, there are many other phenomena to be explained by a theory of consciousness such as blindsight and visual agnosia (Lamme, 2006), and future work should explore whether they can be modeled by semantic pointer competition. We have provided only a few examples of contrastive analysis of what differentiates conscious and unconscious phenomena (Baars, 2009). But our explanations of the onset and cessation of consciousness and of shifts in consciousness are readily extendible to other kinds of phenomena. The general contrast between conscious and unconscious processes results from the formation of semantic pointers that achieve a threshold level of activity; and the more specific contrast between what is conscious and what is unconscious at a particular time results from semantic pointer competition. Winning a competition does not suffice to make a semantic pointer conscious unless the neurons in the relevant population are firing rapidly as the result of a good balance of excitation over inhibition.

Second, SPC is precise enough to be implemented in computer simulations, unlike the computationally difficult calculation of Φ. Third, SPC is ontologically simpler than IIT, postulating only biological processes rather than a mysterious quantity of information integration that goes beyond mass and energy. Some people will find it sad that consciousness is not sufficiently abstract to survive the death of brains, but wishful thinking couched in mathematics is no better than the theological version concerning souls. Fourth, SPC does not attribute consciousness to entities such as photodiodes and countries that show no behavioral signs of consciousness. We have argued that there is no specifiable quantity of consciousness, any more than there is a quantity of life. Undoubtedly, SPC will need expansion and revision with the growth of empirical and theoretical knowledge about the brain, but for now it stands as far preferable to information integration as an explanation of consciousness.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada. For comments on previous drafts, we are grateful to Ned Block, Bruce Bridgeman, Chris Eliasmith, and an anonymous reviewer.

Appendix A. Mathematical exposition and critique of information integration

Tononi attempts to define mathematically the amount of integrated information in a system, but we find his exposition puzzling. In his most recent mathematical treatment (Tononi, 2010, p. 303) he defines Φ, the integrated information in a system X, by this equation:

$$\Phi(X_{i,t1}) = H[X_{i,t1} \| X_{i,t1}/\mathrm{MIP}] \tag{1}$$

First, the left side of the equation of the equation is odd, for why does Φ apply to a particular state $X_{i,\ t1}$ rather than the whole system X? Second, the symbol H is usually used in information theory for the entropy of a random variable, but $[X_{i,\ t1}||X_{i,\ t1}/\mathrm{MIP}]$ is obviously not a random variable, so what does H mean here? Third, the symbol "||" is supposed to stand for the relative entropy (Kullback–Leibler divergence) of two probability distributions, but $X_{i,\ t1}$ is a state, not a probability distribution, so relative entropy does not apply. Fourth, the expression "$X_{i,\ t1}$/MIP" is misleading because it looks like a quantity divided by another quantity MIP, but MIP actually means something like the partition of the system into parts that independently generate information. Whatever the meaning of "$X_{i,\ t1}$/MIP", it does not seem to be a probability distribution of the sort that could feed into a calculation of relative entropy (K–L divergence), standardly defined on probability distributions P and Q as:

$$D_{KL}(P\|Q) = \sum_{i} \ln(P(i)/Q(i))P(i) \tag{2}$$

For these 4 reasons, Tononi's equation for Φ is mathematically uninterpretable, so that no calculation of Φ is possible.

Rather than concluding that Φ is unintelligible, however, it would be more charitable to try it interpret it as a quantity that might actually be computed. Our best effort is to define two probability distributions, one for the system X as an interconnected whole and one for the X divided up into independent parts. We can use X^* for the probability distribution for X and X^{**} for the probability distribution of X as a set of minimally independent parts. Then relative entropy (K–L divergence) applies and we might get an interpretable definition of the information integration of a system X as:

$$\Phi(X) = X^*\|X^{**} \tag{3}$$

A consequence of this definition is that it makes it clear how difficult it is to actually calculate Φ. For example, let X be a system consisting of N neurons. If the neurons are rate neurons, with the capability of firing 100 times per second, then knowing the probability distribution would require knowing the probabilities of N^{100} outcomes, an extraordinarily large quantity for a realistic number of neurons in the millions or billions. The situation is even worse for spiking neurons, whose outcomes require taking into account the temporal patterns of firing, not just the rate of firing. In that case the number of outcomes is $2^{N\hat{}100}$. Hence there is no reasonable way to estimate X^*. The situation is even worse for X^{**}, because for any system X with N elements there are 2^N possible subsets to be evaluated, and determination of a minimal partition looks computationally intractable because partition problems are typically NP-complete (Garey & Johnson, 1979). If X^* and X^{**} are incalculable, then so is $X^*\|X^{**}$. We conclude that making the mathematical definition of Φ more mathematically precise has the effect of demonstrating how unrealistic it is. Even if, contrary to our arguments in earlier sections, there is something like a quantity of consciousness, it would be not be illuminating to have it as incalculable as Φ.

Appendix B. Mathematical exposition of semantic pointer competition

This section is adapted from the appendix to Schröder et al. (2014).

To construct the computational models shown in this paper, we make use of the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). In this approach, we specify a type of distributed representation for each group of neurons, and we analytically solve for the connection weights between neurons that will produce the desired computations between groups of neurons. While this approach does encompass neural learning techniques, we do not use any learning in the models presented here.

More formally, the "patterns" for the various different inputs (e.g. CAT, DOG, MEOW), outputs (e.g. CAT, DOG), and internal concepts (e.g. VERBAL, SENSORY) are all defined as randomly chosen 64-dimensional unit vectors. This gives a unique randomly-generated vector for each concept. To use these patterns in a neural model, we must define how a group of neurons can store a vector using spiking activity, and how this spiking activity can be decoded back into a vector.

To define this neural encoding, the NEF generalizes standard results from sensory and motor cortices (e.g. Georgopoulos, Schwartz, & Kettner, 1986) that in order to represent a vector, each neuron in a population has a random "preferred direction vector" – a particular vector for which that neuron fires most strongly. The more different the current vector is from that preferred vector, the less quickly the neuron will fire. In particular, Eq. (4) gives the amount of current J that should enter a neuron, given a represented vector x, a preferred direction vector e, a neuron gain α, and a background current b. The parameters α and b are randomly chosen, and adjusting their statistical distribution produces neurons that give realistic background firing rates and maximum firing rates (Eliasmith & Anderson, 2003). These parameters also impact the model itself; for example, having an overall lower average firing rate means that the model will require more neurons to produce the same level of accuracy.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + b \tag{4}$$

This current can then be provided as input to any existing model of an individual neuron, to determine the exact spike pattern for a particular input vector x. For this paper, we used the standard Leaky Integrate-and-Fire neuron model, which is a simple model that captures the behavior of a wide variety of observed neurons (Koch, 1999, ch. 14). Input current causes the membrane voltage V to increase as per Eq. (5), with neuron membrane resistance R and time constant τ_{RC}. For the models presented here, τ_{RC} was fixed at 20 ms (Isokawa, 1997). When the voltage reaches a certain threshold, the neuron fires (emits a spike), and then resets its membrane voltage for a fixed refractory period. For simplicity, we normalize the voltage range such that the reset voltage to 0, the firing threshold is 1, and R is also 1.

$$\frac{dV}{dt} = \frac{JR - V}{\tau_{rc}} \tag{5}$$

Given Eqs. (4) and (5), we can covert any vector x into a spiking pattern across a group of realistically heterogenous neurons. Furthermore, we can use Eqs. (6) and (7) to convert that spiking pattern back into an estimate of the original x value. This lets us determine how accurately the neurons are representing given values. More neurons leads to higher accuracy. The idea behind Eq. (3) is that we can take the average activity a of each neuron i, and estimate x by finding a fixed weighting factor d for each neuron. Eq. (4) shows how to solve for the optimal d as a least-squared error minimization problem, where the sum is over a random sampling of the possible x values.

$$\hat{x} = \sum a_i d_i \tag{6}$$

$$d = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \sum_x a_i a_j \quad \Upsilon_j = \sum_x a_j \mathbf{x} \tag{7}$$

These two equations allow us to interpret the spiking data coming from our models. In Fig. 4 through Fig. 8, we take the spike pattern, decode it to an estimate of x, and compare that to the ideal vectors for the various concepts in the model. If these vectors are close, then we add the text labels (e.g. CAT, DOG, MEOW) to the graphs, indicating that the pattern is very similar to the expected pattern for those terms.

It should be noted that this produces a generic method for extracting x from a spiking pattern without requiring a specific set of x values to optimize over. That is, we can accurately use d to determine if a particular pattern of activity means CAT even though we do not use the CAT vector to compute d. The sums used to compute d in Eq. (4) are over a random sampling of x. Since x covers a 64-dimensional vector space and since we use only 5000 samples in that space (increasing this number does not affect performance), it is highly unlikely that the sampling includes exactly the vector for CAT (or any other semantic pointer), but as shown in the Fig. 4 through Fig. 8, we can still use d to identify the presence of those semantic pointers (or any others).

Importantly, we also use Eq. (7) to compute the connection weights between groups of neurons. In contrast to other neural modelling methods which rely on learning, the NEF allows us to directly compute connection weights that will cause neural models to behave in certain ways. For example, given two groups of neurons, we can form connections between them that will pass whatever vector is represented by one group to the next group by using the connection weights given in Eq. (8) (see Eliasmith & Anderson, 2003 for the proof).

$$\omega_{ij} = \alpha_j \mathbf{e}_j \cdot \mathbf{d}_i \tag{8}$$

However, simply passing information from one group to another is insufficient to implement simulations presented here. Fortunately, the NEF shows that you can find alternate d values to estimate complex nonlinear functions. That is, instead of simply passing a value from one group to another, we can define an arbitrary function $f(x)$ and compute d^f as per Eq. (9). Now, if synaptic connections are formed via Eq. (8), if the first neural population fires with the pattern for x, then the connections will cause the second population to fire with a pattern representing the result of $f(x)$.

$$\mathbf{d^f} = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \sum_x a_i a_j \quad \Upsilon_j = \sum_x a_j \mathbf{f}(\mathbf{x}) \tag{9}$$

This approach allows us to define the transformations described in this paper. For example, the connections from the sensory population population to the semantic pointer population are found by using Eq. (9) with the function being $f(x) = x^{*}$SENSORY, where * is circular convolution (see below) and SENSORY is a randomly chosen 64-dimensional vector. This same approach is used for other inputs. For the extracted outputs, we use $f(x) = x/$SENSORY, where the / indicates circular correlation (see below). These functions are used to compute d^f (Eq. (9)), which is then used to compute the synaptic connection weights (Eq. (8)). The model is then run. To provide input to the model, we generate input current into the sensory neurons for the particular sensory stimuli (Eq. (4)). To analyze and interpret the spiking patterns, we convert the spikes back into a vector (Eq. (6)) and compare it to the ideal vectors for each concept.

The binding function used here is *circular convolution*. This takes two vectors (x and y) and produces a third vector z as per Eq. (10). This vector z can be thought of as a compressed representation of x and y, forming the basis of our semantic pointers. Importantly, given z and y (or x) we can recover an approximation of x (or y) by computing the circular correlation (Eq. (11)). This is how semantic pointers can be decompressed into their constituents.

$$\mathbf{z}_i = \sum_j \mathbf{x}_j \mathbf{y}_{i-j} \tag{10}$$

$$\hat{\mathbf{x}}_i = \sum_j \mathbf{z}_j \mathbf{y}_{i+j} \tag{11}$$

In general, it is possible to use the Neural Engineering Framework to build a network where there are two input populations (one for x and one for y) and one output population (z) such that you can input any two arbitrary vectors and get out their convolution. Importantly, this will work for any input vectors, not just the randomly chosen ones used in the optimization (Eq. (9)). However, for the simulations described here, we use a simpler method where a particular neural connection always convolves its input vector x with a fixed vector, such as $f(x) = x^{*}$VERBAL. The synaptic connection weights computed using this function and Eqs. (8) and (9) result in a spiking neural network that accurately combines information into a single memory semantic pointer regardless of what particular vector x is provided to the sensory system. A similar function is defined for the other connections into the memory system, resulting in a final semantic pointer of x^*SENSORY + y^*VERBAL. To decompress this semantic pointer, we use a circular correlation instead (Eq. (11)).

To implement competition, we again use Eqs. (8) and (9), forming connections from the semantic pointer population back to itself. In this case, the function we want to compute is a combination of all the semantic pointer competition effects. For example, if we want DOG and CAT to compete, we can use Eq. (12), where the result is the outer product between the two vectors. With this function, if x is DOG, the result is approximately −CAT, and if x is CAT, the result is −DOG. Thus, whenever the pattern for CAT is present, it will reduce the pattern for DOG, and vice versa. We can also adjust the strength of this competition by adjusting the scaling value used (here set to −1). For more complex situations, more interactions among semantic pointers can be added simply by adding them into this equation.

$$f(x) = \{-1(\mathrm{DOG} \times \mathrm{CAT}) - 1(\mathrm{CAT} \times \mathrm{DOG})\}x \tag{12}$$

References

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*, 45–53.

Baars, B. J. (2009). Contrastive analysis. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 195–198). Oxford: Oxford University Press.

Balduzzi, D., & Tononi, G. (2009). Qualia: The geometry of integrated information. *PLoS Computational Biology, 5*, e1000462.

Baumeister, R. F., Masicampo, E. J., & Vohs, K. D. (2011). Do conscious thoughts cause behavior? *Annual Review of Psychology, 62*, 331–361.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences, 9*, 46–52.

Block, N. (2009). Comparing the major theories of consciousness. In M. Gazzaniga (Ed.), *The cognitive neuroscience IV*. Cambridge, MA: MIT Press.

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. (2014). Concepts as semantic pointers: A theory and computational model. *Unpublished manuscript, University of Waterloo.*

Braithwaite, V. (2010). *Do fish feel pain?* Oxford: Oxford University Press.

Braun, J. (2009). Attention and awareness. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 72–77). Oxford: Oxford University Press.

Bromfield, E. B., Cavazos, J. E., & Sirven, J. I. (2006). An introduction to epilepsy [Internet] <http://www.ncbi.nlm.nih.gov/books/NBK2510/>.

Brook, A., & Raymont, P. (2010). The unity of consciousness. *Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/consciousness-unity/>.

Broom, D. M., Sena, H., & Moynihan, K. L. (2009). Pigs learn what a mirror image represents and use it to obtain information. *Animal Behavior, 78*, 1037–1041.

Bunge, M. (2003). *Emergence and convergence: Qualitative novelty and the unity of knowledge*. Toronto: University of Toronto Press.

Canales, A. F., Gomez, D. M., & Maffet, C. R. (2007). A critical assessment of the consciousness by synchrony hypothesis. *Biological Research, 40*, 517–519.

Carruthers, P. (2011). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/consciousness-higher/>.

Chabris, C., & Simons, D. (2010). *The invisible gorilla*. New York: Crown.

Chalmers, D. J. (1996). *The conscious mind*. Oxford: Oxford University Press.

Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology, 2*. http://dx.doi.org/10.3389/fpsyg.2011.00086. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3110382/>.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. London: Simon and Schuster.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*, 119–126.

Damasio, A. (2012). *Self comes to mind: Constructing the conscious brain*. New York: Vintage.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience, 14*, 143–152.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York: Viking.

Dehaene, S. (2009). Neuronal global workspace. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 466–470). Oxford: Oxford University Press.

Dehaene, S., & Changeux, J. (2011). Experimental and theoretical approaches to consciousness processing. *Neuron, 70*, 200–227.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford: Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al (2012). A large-scale model of the functioning brain. *Science, 338*, 1202–1205.

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science, 25*, 245–286.

Engel, A. K., Fries, P., König, P., Brecht, M., & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition, 8*, 128–151.

Findlay, S. D., & Thagard, P. (2012). How parts make up wholes. *Frontiers in Physiology, 3*. <http://www.frontiersin.org/Journal/Abstract.aspx?s=1086&name=systems_biology&ART_DOI=10.3389/fphys.2012.00455> http://dx.doi.org/10.3389/fphys.2012.00455.

Garey, M., & Johnson, D. (1979). *Computers and intractability*. New York: Freeman.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416–1419.

Goldfarb, L., & Treisman, A. (2013). Counting multidimensional objects: implications for the neural-synchrony theory. *Psychological Science, 24*, 266–271.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences, 13*, 194–202.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554–2558.

Isokawa, M. (1997). Membrane time constant as a tool to assess cell degeneration. *Brain Research Protocols, 1*, 114–116.

Kant, I. (1929 (Originally published 1787)). *Critique of pure reason* (N. K. Smith, Trans.). New York: St. Martin's Press.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.

Koch, C. (2012). *Consciousness: Confessions of a romantic reductionist*. Cambridge, MA: MIT Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*, 494–501.

LeDoux, J. (1996). *The emotional brain*. New York: Simon and Schuster.

Maia, T. V., & Cleeremans, A. (2005). Consciousness: Converging insights from connectionism modeling and neuroscience. *Trends in Cognitive Neuroscience, 9*, 397–404.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

Plate, T. (2003). *Holographic reduced representations*. Stanford: CSLI.

Porkka-Heiskanen, T., & Kalinchuk, A. V. (2011). Adenosine as a sleep factor. *Sleep and Biological Rhythms, 9*, 18–23.

Prinz, J. (2012). The conscious bra. In *How attention engenders experience*. Oxford: Oxford University Press.

Prior, H., Schwarz, A., & Güntürkün, O. (2008). Mirror-induced representation in magpies: Evidence of self-recognition. *PLoS Biology, 6*(8), e202.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge MA: MIT Press/ Bradford Books.

Schröder, T., & Thagard, P. (in press). Priming: Constraint satisfaction, competition, and creativity. *Social Cognition*.

Schröder, T., Stewart, T. C., & Thagard, P. (2014). Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive Science, 38*, 851–880.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review, 120*, 255–280.

Selfridge, O. G. (1958). Pandemonium: A paradigm of learning. In *Mechanisation of thought processes: Proceedings of a symposium held at the National Physical Laboratory* (pp. 513–526). London: Her Majesty's Stationery Office.

Slagter, H. A., Johnstone, T., Beets, I. A. M., & Davidson, R. J. (2010). Neural competition for conscious representation across time: An fMRI study. *PLoS ONE, 5*, e10556.

Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Upper Saddle River, NJ: Pearson Prentice Hall.

Steriade, M., & McCarley, R. W. (2005). *Brain control of wakefulness and sleep*. New York: Springer.

Stewart, T. C., & Eliasmith, C. (2012). Compositionality and biologically plausible models. In W. Hinzen, E. Machery, & M. Werning (Eds.), *Oxford handbook of compositionality* (pp. 596–615). Oxford: Oxford University Press.

Thagard, P. (2010). *The brain and the meaning of life*. Princeton, NJ: Princeton University Press.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P. (2014a). Creative intuition: How EUREKA results from three neural mechanisms. In L. M. Osbeck & B. S. Held (Eds.), *Rational intuition: Philosophical roots, scientific investigations*. Cambridge: Cambridge University Press, pp. 287–306.

Thagard, P. (2014b). Explanatory identities and conceptual change. *Science & Education, 23*, 1531–1548.

Thagard, P. (2014c). The self as a system of multilevel interacting mechanisms. *Philosophical Psychology, 27*, 145–163.

Thagard, P. (2014d). Thought experiments considered harmful. *Perspectives on Science, 22*, 288–305.

Thagard, P., & Schröder, T. (in press). Emotions as semantic pointers: Constructive neural mechanisms. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotions*. New York: Guilford.

Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and Cognition, 17*, 811–834.

Thagard, P., & Stewart, T. C. (2011). The Aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1–33.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*. <http://www.biomedcentral.com/1471-2202/5/42> http://dx.doi.org/10.1186/1471-2202-5-42.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin, 214*, 216–242.

Tononi, G. (2010). Information integration: Its relevance to brain function and consciousness. *Archives italiennes de biologie, 148*, 299–322.

Tononi, G. (2012). *PHI: A voyage from the brain to the soul*. New York: Pantheon.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences, 1124*, 239–261.

Tripoli, M., & Torg, J. (2011). Pathophysiology of concussion: A review of the literature. *Temple University Journal of Orthopaedic Surgery & Sports Medicine*, <http://www.templeconcussion.com/files/Pathophysiology_of_Concussion_2011-spring.pdf>.

Villars, P. S., Kanusky, J. T., & Dougherty, T. B. (2004). Stunning the neural nexus: Mechanisms of general anesthesia. *AANA Journal, 72*, 197–205.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge, MA: Harvard University Press.