Beyond IIT and GWT

Alternative Empirically Testable Theories of Conscious

BRIE LINKENHOKER, WORLDVIEW STUDIO

At a meeting sponsored by the Templeton World Charity Foundation (TWCF) in early 2018, thirteen scholars of consciousness met to generate ideas for experiments to advance our collective understanding of consciousness. They generated a set of experiments that would empirically test some of the core claims of two of the leading theories, integrated information theory (IIT) and global workspace theory (GWT). IIT, developed by Tononi (2004) and elaborated by Koch and others (2016), holds that the degree of consciousness in a system correlates with the ability of the system to integrate information, as measured by Φ. Consciousness *is* integrated information, and the degree of consciousness can vary across species. In IIT, conscious experiences are defined (Tononi, 2008) as subjective, structured (or relational), specific, unified, and definitive (or bounded). GWT, developed by Baars (1989) and later elaborated by Dehaene (2014) and others, holds that conscious cognitive content can be "broadcast" to engage cognitive processes including attention, evaluation, memory, and verbal report. The experience of consciousness is created by what is in the global workspace at any given moment, engaging many other circuits in the brain. The workspace has a bottleneck, so only one conscious percept can be present at once, and as one slips away, a new one takes its place, thus accounting for the temporal stream of conscious experience.

The experiments to compare these theories are moving forward now.

TWCF's aim with the larger "Accelerating Research on Consciousness" project is to use adversarial collaboration and other open science practices to advance our collective understanding of consciousness, in part by generating new experiments that would rigorously test competing claims across two or more theories of consciousness. The hope would be to reduce the number of plausible empirically testable theories either through collapsing previously differentiated theories into a unified theory, or through the experimental rejection of one or more existing theories.

These are laudable goals. To further this process, Worldview Studio set out to identify plausible alternative theories to IIT and GWT. TWCF's overarching aim is to empirically test theories of consciousness that not only describe awakeness or awareness, but also the phenomenal experience of being conscious, so we limited ourselves first to theories that at least aim to describe the "what is it likeness" of consciousness. Second, we focused on theories of consciousness that are rooted in the cognitive, computational, and/or biological architecture of the brain as we know it. This focus left out a large body of important conceptual work in philosophy. Some of that philosophical work has influenced the more brain-based theories of consciousness below, but other philosophical work has been left out of scope due to its lack of near-term testability. This includes significant work in the "consciousness as illusion" tradition. We have also left out theories that focus on quantum level explanations of consciousness as these

appear to fail the testability criterion, at least with our current techniques.

Our hope at the outset of this project was that we would be able to identify and succinctly summarize three to five alternative theories to IIT and GWT that share those two theories' testability and rootedness in the cognitive, computational, and/or biological architecture of the brain. We also hoped that these theories would be testable *against each other*, meaning that they would share enough common definitions and assumptions about the *nature* of consciousness to be of the same type, but have enough differences in proposed mechanisms that good experimental design could – theoretically – distinguish between them.

What we found were theories that did indeed share some of the same explananda, specifically a desire to account for both the so-called "easy" and "hard" problems of consciousness (Chalmers, 1995). The easy problems include wakefulness, the ability to discriminate between and react to stimuli, the deliberate control of behavior and the focus of attention, and the ability of a conscious system to access and report on its own internal states. The hard problem, on the other hand, is the problem of subjective experience. All of the theories below are at least *aiming* at the hard problem, even if it is still a long way off.

The challenge is that the theories we found differed in ways that are so fundamental that imagining experiments that would compare and test them head-to-head is almost impossible right now. Some of these differences concerned the definitional aspects of consciousness, and others the purpose of consciousness or the nature of mental representation, but the biggest differences were in their fundamental framing and starting points. A joke among neuroscientists is that everyone has a theory of consciousness that starts with their own lab's research, and we found some truth to this! The good news, however, is that these different lenses on consciousness, while not making for easy head-to-head comparison, do have incredible potential to enrich and elaborate each other with the right forms of collaboration. If TWCF could make *that* happen, the Foundation might still achieve its goal of reducing the number of plausible theories of consciousness *and* make the remaining ones that much stronger.

One important author's note before the review of theories: My goal has been to helpfully summarize the following theories of consciousness in line with their presentation by their authors. I have therefore quoted liberally and directly from their work, and I in no way claim intellectual credit for the ideas they have put forth. Any errors in summarizing their theories or in asking questions that might have been answered in other publications not reviewed here are my own. The summaries below in no way try to *infer* answers to common questions across theories or fill in explanatory gaps left by the authors. That is a much longer project, and one I return to in the report's conclusion.

REVISITING HIGHER ORDER THEORIES

Higher order theories (HOTs) of consciousness may already be on TWCF's radar, but recent advances in these theories are worth taking another look. Higher order theories of consciousness claim that phenomenal consciousness requires higher order representations (HORs) of lower level (first order) perceptions or other more basic representations in order to achieve consciousness of the first order state. The HORs are about first order representations, and while you need HORs for consciousness, you are not aware of HORs (though you could be aware of HORORs, higher order representations of representations; more on that below). Higher order theories differ from first order theories (such as those put forth by Ned Block) that claim that a higher order representation is not required for consciousness; in first order theories, the brain needs only to access the first order representation in order to become conscious of it.

Higher order theorists, like David Rosenthal, Richard Brown, Joseph LeDoux, and Hakwan Lau, argue that "access" to stimulus processing is not enough, and that "access" to first order perceptions is difficult to define, verify or falsify in the brain. According to higher order theorists, general networks of cognition (GNC), which are cortical circuits involving regions in lateral and medial prefrontal cortex, posterior parietal cortex, and insular cortex, make conscious the information that may be processed in, for example, primary and secondary sensory cortex. HOT supporters don't claim that these regions are the "seat" of consciousness, only that they are involved in complex, cross-circuit interactions known to be essential to attention, working memory, and meta-cognition, and therefore, also (according to HOTs) critical to the introspective properties of phenomenal consciousness.

Different higher order theories appeal to different kinds of cognitive activities that create HORs, including attention and working memory, and in some cases, potentially even processing by a global workspace. Because HOTs are a *class* of theories, all the HOTs sometimes face criticisms that ought to be leveled at only a subset of theories. They can also be put forth as encompassing more finely tuned theories, including all or part of GWT (LeDoux and Brown, 2017), or as being a variation on the theme of another theory. In their recent paper, "The Misunderstood Higher-Order Approach to Consciousness," Brown, Lau and LeDoux (2019) outline and respond to seven misconceptions about HOTs. These are included at length below in part because it is just this kind of clear response to specific criticisms, confusions and questions that will help advance the field.

Misconception 1: HOT is a single theory. The authors note several "currently active" HOTs, including

» The phenomenal self theory (Metzinger, 2003): A conscious brain models the world and the self. The experienced self is a representational model of self that is created continuously by local processes in the brain. The continuity of self is adaptively important as our bodies and environments

REVISITING HIGHER ORDER THEORIES

change. Part of what makes this theory a HOT is that the conscious self-models cannot be recognized as models. (Note that this theory shares some themes in common with the Projected Consciousness Model (PCM) below.)

Higher order theories of consciousness assume "minimal cognitive functions for consciousness. In this sense, [they occupy] an intermediate position between GWT and early sensory views, and plausibly account for shortcomings of both."

(BROWN, LAU & LEDOUX, 2019)

» The radical plasticity hypothesis (Cleeremans, 2011): Consciousness is something that the brain learns to do rather than an intrinsic property of specific neural circuits. "Consciousness arises as a result of the brain's continuous attempts at predicting not only the consequences of its actions on the world and on other agents, but also the consequences of activity in one cerebral region on activity in other regions. By this account, the brain continuously and unconsciously learns to redescribe its own activity to itself, so developing systems of meta-representations that characterize and qualify the target first-order representations." (Note similarities between this theory and *unlimited associative learning* as the marker of the *evolutionary transition to minimal consciousness* briefly described below. Cleeremans theory could easily have been included here as one of the primary alternative theories to GWT and IIT.)

» Several versions of HOTT (Rosenthal, 2008), higher order thought theory, which postulate that the higher-order state is "thought-like." In HOTT, a mental state is conscious if and only if one is conscious of oneself, in some suitable way, as being in that mental state. Consciousness of cognitive and desiderative states is not likely to directly "enhance processes of rational thought and planning, intentional action, executive function, and the correction of complex reasoning," but rather "come to be conscious as a result of other useful psychological developments, including language."

Misconception 2: *HOT suggests that sophisticated thoughts are necessary for conscious experiences.* This is a criticism primarily aimed at HOTT, and captures a concern that requirements for "thought," "introspection," and "self" might rule out conscious experience for other mammals, and even in humans be more complex than required for conscious experience. The authors counter that the problem lies in the colloquial understanding of these words, and that HOTs are actually quite "lean" in their notions of what it means to be aware or to introspect.

REVISITING HIGHER ORDER THEORIES

Misconception 3: *HOT says consciousness is the same as metacognition or confidence.* This confusion likely results from the fact that many experimental paradigms used to study HOT employ metacognitive tasks, which require subjects to give confidence ratings to perceptual decisions or memories. While these tasks are useful probes of the mechanisms of consciousness, their use does not imply that metacognition is *necessary* for consciousness.

Misconception 4: *HOT is a variant of GWT.* Both HOT and GWT require downstream processes in addition to first-order representations to account for consciousness. In GWT, these processes *serve the purpose* of global broadcast to different modules/modalities, which then strengthen and stabilize conscious signals and give them power in behavioral and cognitive control. HOTs do *not* propose a higher purpose for higher-order representation beyond giving rise to conscious experience itself. The authors offer blindsight as an example of how these differences play out. In GWT, the lack of conscious experience "must mean that the relevant signal is not globally available to all major modules of the brain; somehow, some local pathway must have made possible the successful guessing and stimulus identification." In HOT, blindsight can be explained by the idea that a "percept can reach a relatively stable global state and remain nonconscious, which accounts for potentially powerful forms of unconscious perception."

Conscious experiences, regardless of their content, arise from one system in the brain. In this view, what differs in emotional and nonemotional states are the kinds of inputs that are processed by a general cortical network of cognition, a network essential for conscious experiences.
(LEDOUX & BROWN, 2017)

Misconception 5: *HOT is about reports and access, not experiences per se.* Because some HOT experiments involve self-report, some may confuse the experimental paradigm with what the theory holds to be critical for consciousness. HOT proponents argue that subjective experience is *not* just about access to information and self-report, but that HOT is much more focused on experience.

Misconception 6: *HOT is implausible because conscious experiences are more tightly associated with first-order (sensory) activity.* This criticism essentially asks what HOTs contribute beyond first order theories. In the visual system, if V4 is processing information about color and MT about motion, then any visual processing downstream of these areas may just allow access and reporting. But since unconscious or subliminal stimulation also engages these areas, their activation is probably not *sufficient* for conscious experience (though there are cogent arguments against this assertion). The current data suggest that it is at least plausible that, at least in

the visual system, extrastriate visual activity has to be re-represented by HORs to become conscious.

Misconception 7: *HOT suggests that consciousness is* in *the prefrontal cortex*. While the PfC has been implicated experimentally in aspects of higher cognition that may be essential to phenomenal consciousness, different HOTs take different perspectives on its role. In HOTT, which of the aspects of visual system activity enter consciousness are jointly determined by concurrent activity in PfC areas and areas that maintain lower-order representations. In HOROR theory, the higher order state in PfC is *itself* phenomenally conscious. Despite these differences, virtually no HOT proponents assume that one is "more conscious" when PfC is more active.

The authors conclude by saying that HOT is empirically testable, and particularly valuable in explaining "everyday experiences" of most people as well as those who suffer from mental disorders, and they call for more attention to those phenomenal, self-reported experiences - not just behaviors and measurable forms of cognition - in the treatment of mental disorders.

While this paper was helpful in its detailed responses to criticisms, more work needs to be done on differentiating HOTs from other theories of consciousness - especially GWT - in ways that go beyond the understood *purpose* of higher order brain activity related to lower level percepts. HOTs (and especially HOTTs or HORORs) may be the best situated for head-to-head experimental comparison to GWT *provided that they differ significantly beyond the purpose of higher order representation,* which will be difficult to probe experimentally.

On the topic of purpose in HOT, LeDoux and Brown (2017) push HOT into the realm of emotion, which differs from the significant focus on sensory (and particularly visual) perception in much of the consciousness literature (with the notable exception of Antonio Damasio's work; see below). Emotions, they argue, "can never be unconscious. Responses controlled by subcortical survival circuits that operate nonconsciously sometimes occur in conjunction with emotional feelings but are not emotions." Their view is that emotional experience depends not just on HORs, but on HORORs. They trace these representations of representations through the example of fear:

» A threat, like a snake, is perceived by the visual system

» That first order representation enters into a HOR, along with long-term memories and subcortical survival circuit activation associated with snakes. All of this is still nonconscious.

» A HOROR then allows for a conscious noetic experience of the perceived snake as dangerous (where conscious noetic experience is defined as not involving the self, and as being connected to semantic or factual knowledge)

REVISITING HIGHER ORDER THEORIES

- To get to an experience of *fear*, the HOROR must integrate autonoetic conscious experience (where autonoetic experience is connected to autobiographical episodic memories that involve the self)

- Since the cognitive functions (working memory, attention, metacognition, etc.) and neural circuits involved in the general networks of cognition (especially in PfC) support HORs, these same circuits should support the "self-HOROR" required for the experience of fear of the snake.

- The self-HOROR integrates visual perception, subcortical survival circuit activation, long-term memories (both semantic and autobiographical), brain and body arousal, and self- and emotion-schema.

This schematic is quite complicated, so much that the resemblance to other HOTs is somewhat tenuous. But the self-HOROR doesn't require any different fundamental circuitry than other HORs, and the HORORs have the advantage of being able to explain how one could feel fear of something imagined that isn't really there. This, the authors claim, allows them to potentially account for all forms of fear: those accompanied by brain arousal and bodily responses and those that are purely cognitive and even existential.

One can argue to what extent the more complicated emotional self-HOROR described by LeDoux and Brown is really in the same class as the simpler HOTs rooted in visual perception. The authors appear to see more similarity than difference. The advantages of the self-HOROR work in emotion lay in a) the way that emotional experience challenges HOT and requires its adaptation, b) the fact that the self-HOROR requires integration of the self through autonoetic experience/autobiographical memory, and/or a "schema" of self, and c) the construction of a framework in which HORs can represent things that are imagined, remembered, or contemplated. These advances may move higher order theories as a class further along towards capturing more of our most critical "What is it like to..." subjective experiences.

If you're going to read just two papers:

Brown R, Lau H, and LeDoux JE (2019). The Misunderstood Higher-Order Approach to Consciousness. https://doi.org/10.31234/osf.io/xpy8h

LeDoux JE and Brown R (2017). A higher-order theory of emotional consciousness. *PNAS* 114 (10) E2016-E2025. https://doi.org/10.1073/pnas.1619316114

ATTENTION SCHEMA THEORY (AST)

Subjective awareness, claims AST, is the brain's internal model of attention - and nothing more. AST's primary proponent, Michael Graziano, prefers the term "subjective awareness" to consciousness simply because it is a less loaded term. Graziano writes that AST is in line with Daniel Dennett's view that "the problem of subjective experience consists only in explaining why and how the brain concludes that it contains an apparently non-physical property," but that AST goes further than other "illusionist" models by naming the functional utility of the construct of awareness: that it serves as a model of attention.

Just as a battlefield with model armies can provide a general with a simplified, imperfect, but *useful* model for strategic planning, so can a constantly updated model of attention help the brain direct its attentional resources towards internal states, external stimuli, or a combination of both. Thus, AST does appear able to answer the question, "What is consciousness (or subjective awareness) for?" The role of consciousness is to direct attention and process information about the subjects of attention. There is, Graziano concludes, "adaptive value for a brain to build a construct of awareness."

Graziano's position is that the internal model of attention has no need for – and thus ignores – the neurobiological mechanisms of attention (lateral inhibition, signal competition, etc.). The internal model is instead focused on "the most salient aspects of attention – the ability to take mental possession of an object, focus one's resources on it, and, ultimately, act on it." The "quick and dirty" or "fast and frugal" version of awareness-as-simplified-model-of-attention is described as similar to our perception of white light. We see white light as brightness with the absence of color, rather than seeing it for what it really is – a combination of all colors of light. Another example is the brain's model of the body, its movements, and its position in space. This model is influenced by things like muscle tension, limb orientation, and balance perception, but the *experience* of tracking the body in space doesn't include this granular information. Similarly, consciousness gives us a unified "good enough" picture of what we're attending to inside our own bodies and thoughts, or in the external world. This attention schema (model) can then guide top-down decisions about what we should attend to in the next moment in time.

The content of this model, says Graziano, leads the brain to *conclud*e that it has a subjective experience. This isn't necessarily a true or false conclusion; it simply is. Graziano claims that AST can be consistent with both GWT and IIT, in that the process of attaching the construct of awareness to an item (a thought, an apple, etc.) requires the integration of multiple streams of information about self, the item, and their relationship into

ATTENTION SCHEMA THEORY (AST)

a coherent, unified experience, which could require a brain-wide global workspace. The unique addition of AST is that awareness does not happen only because the brain integrates information or broadcasts it across multiple networks; awareness must be *constructed* in the form of an attentional model that allows not only a report that "the apple is red," but also that "I am aware of that apple."

We propose that the top–down control of attention is improved when the brain has access to a simplified model of attention itself. The brain therefore constructs a schematic model of the process of attention, the 'attention schema,' in much the same way that it constructs a schematic model of the body, the 'body schema.' The content of this internal model leads a brain to conclude that it has a subjective experience.
(WEBB & GRAZIANO, 2015)

In AST, "awareness is part of the control machinery for attention." This means that awareness and attention should be closely correlated, but separable in some situations. A classic example showing the tight correlation is the well known video of basketball players dressed in white or black that viewers are asked to watch so they can count the number of passes between players wearing white. During the video, a person dressed as a gorilla walks into the scene and out again. Most viewers fail to see the gorilla at all since they are attending only to the players in white. Thus, attention and awareness (or lack thereof) are tightly linked. However, blindsight experiments and experiments in which a stimulus is masked or presented very briefly show that bottom-up attention can improve task performance with *no awareness* of the stimulus.

Another implication of the AST model is that if awareness is the model of current attention that guides the deployment of future attentional resources, attention *without* awareness should be more poorly controlled. In social psychological experiments on race bias cited by Graziano, subjects who scored low on explicit measures of racism showed *more* stereotyping biases when presented with subconscious stimuli that primed those biases than when they were shown explicit racially stereotyped priming stimuli. In other words, awareness of the priming stereotyping made it easier for them to inhibit their own stereotyped responses, whereas lack of awareness made it harder to direct their attention towards stereotype suppression. Similar results have been shown in visual discrimination tasks, in which subjects performed better on a central task when they were aware of distracting stimuli on the edge of the screen than when they were not awarel

ATTENTION SCHEMA THEORY (AST)

AST is a parsimonious theory with roots in both cognitive neuroscience and neurophilosophy, but it leaves several questions unanswered, including: Where in the brain – and how – is the model of attention created? How is it updated? No one would dispute the critical role of attention in human cognition, but why is the model of attention, as opposed to some other internal model (e.g., the self in space), the seat of awareness? How does a *representation* of current attention become causal, and by what means does it direct attention in the next moment? Can the attention schema power other forms of intentionality that go beyond directing attention in the near future?

Graziano's writing is very convincing when it comes to considering a major role for attention in consciousness, and perhaps even in the proposal that a model of attention *could* be (or constitute a large part of) phenomenal awareness (consciousness). But there are still many holes in the evidence that phenomenal awareness *is* the internal model of attention, and gaps in the theory in terms of where and how the model is instantiated and acts within the brain. All that said, it isn't hard to imagine some compelling additional brain imaging experiments that might start to untangle a model of attention from attention itself, thus making it a potentially experimentally tractable model.

If you're going to read just one paper:

Webb TW and Graziano MSA (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2015.00500

For a good overview written for public audiences:

Graziano MSA (2013). How consciousness works and why we believe in ghosts. *Aeon*. https://aeon.co/essays/how-consciousness-works-and-why-we-believe-in-ghosts

THE PROJECTIVE CONSCIOUSNESS MODEL

The team proposing the Projective Consciousness Model (PCM) includes a philosopher, a mathematician, and two neuroscientists. As might be expected from such a team, the PCM has roots in each of these fields, making it unique among theories of consciousness. The function of consciousness, says the team, is "to address a general 'cybernetic' problem or problem of control. Consciousness enables a situated organism with multiple sensory channels to navigate its environment and satisfy its biological (and derived) needs efficiently." Seeing consciousness as a solution to a problem of control is similar to the framing used by Graziano, but the two theories differ in many other ways.

The PCM starts from the idea that all our sensory experiences are unified into one perceptual experience of the world around us *with oneself at the center* in a first person point-of-view. The first personness of consciousness, says the team, is non-trivial and at the core of conscious experience. Starting from this observation, the team proposes a set of postulates that define consciousness, in which #s 2-5 are implicated in #1:

» Relational phenomenal intentionality: All consciousness involves the appearance of a world (of objects, properties, etc.), in various qualitative or representational ways, to an organism. Consciousness is at root relational in its structure in a manner that is at least proto-spatial and is always framed around a point of view.

» Spatiality: The space of the presented world (objects, etc.) is 3-dimensional and perspectival, unfolding in an oriented manner between a point of view and a horizon at infinity (where all parallel lines converge). The origin of the point of view is elusive, though it normally seems to be located in the head. The space is normally organized around the lived body.

» Multimodal synchronic integration: Consciousness involves the synthesis or integration into a unified whole of a multiplicity of qualitative and representational components (from sensory modalities, memory, and cognition).

» Temporal integration: Consciousness involves at least the retention of immediately past experiences and the protention of immediately future experiences, as integral elements of its "specious present" and a foundation for more expansive forms of temporal integration (distant memories, long-term plans, etc.).

» Subjective character: Consciousness involves a pre-reflective, non-conceptual awareness of itself and its individuality.

The focus on consciousness as relational from a first-person perspective, which entails location and situatedness, leads the team to incorporate an FoC (field of consciousness) into their model. They posit that the FoC cannot be Euclidean in its geometrical structure despite the fact that it allows us to navigate

THE PROJECTIVE CONSCIOUSNESS MODEL

what is an essential Euclidean environment. The team goes to great lengths to describe the features of the first-person, projective three-dimensional space, and claims that the perspectival projective transformations of the world around us can account for some of the more peculiar phenomena of consciousness (e.g., out-of-body experiences), and well known visual phenomena like the multistable perception of Necker Cubes.

The projective consciousness model "would give us a theory of consciousness that intelligibly links the phenomenology of first-person conscious experience with its realization in the brain (or other substrate) via a mathematical and computational model and simultaneously accounts for its biofunctional properties."
(WILLIFORD ET AL., 2018)

Another key component of the PCM is its focus on "free energy (FE) minimization, "the concept that some biological systems seek to minimize the difference between their prior model of the world and new sensory information and perceptual experience. One of the PCM authors, Karl Friston, has written previously about FE minimization, noting that the brain can minimize free energy as defined above by 1) continuous updating of the prior model of the world based on "surprise" or mismatches between the prior model and real experience, and 2) actively changing the world into a state expected by the internal model. In the PCM, this FE minimization, or optimization, happens in reference to the first-person FoC. The conscious brain builds internal models of the world from a first-person perspective, and then uses cycles of perception, action, imagination, and reflection to minimize FE. Conscious brains also use FE minimization to maximize preference satisfaction through choice and action.

While the first-person perspective is at the heart of the PCM, the team proposes that the brain also stores a "world model," which is generally subconscious, but accessible in memory and stores prior experiences, beliefs, and knowledge about the world. Critically, this model of the world largely conforms to Euclidean space, and like the first-person model, can be continuously updated. The FoC accesses prior information in the world model, and updates the world model with information gained from its first-person perspective.

The PCM, says the team, accounts for "the phenomenologically available, generic structure of consciousness and shows how consciousness allows organisms to integrate multimodal sensory information, memory, and emotion in order to control behavior, enhance resilience, optimize preference satisfaction, and minimize predictive error in an efficient manner." The goal of the PCM

THE PROJECTIVE CONSCIOUSNESS MODEL

is to generate specific empirical predictions that can be both mathematically modeled and tested through the collection of physiological, self-report, and behavioral data. The team describes the mathematics of these predictions, and the modeling of simulated conscious agents, as "demanding," with work underway and as yet unpublished. But the *goal* of generating a model of consciousness that is mathematically modellable, explains observed (and often tricky) phenomena of consciousness, makes falsifiable predictions, and lends itself to empirical testing seems like a major advance in the field.

The PCM is a computational model of consciousness, and as such, will face some predictable philosophical pushback, to which the team responds extensively. They write: "As long as the computational model captures the phenomenological invariants, meets the functional constraints, and survives the tribunal of prediction and experiment, we should happily accept that some apparently contingent and only empirically knowable constraints pertaining to the physical realization of consciousness are actually essential to it." They also address the subjective character of phenomenal selfhood, and claim that the PCM can account for critical elements of the experience of consciousness, including "(i) the first-person point of view, (ii) pre-reflective self-consciousness, (iii) global self-consciousness, (iv) social self-awareness, (v) ipseity or "mineness" (foundational to the sense of body ownership and agency), (vi) the "transcendental ego", and (vii) the autobiographical self or "empirical ego."

In their summary of the PCM, the authors anticipate and respond to an impressively broad set of potential questions and objections from computational, philosophical, clinical and biological perspectives. These are too extensive to summarize here. They do note that the idea of global availability that is a core feature of the GWT is built into their FoC, "a virtual projective space in which various contents become accessible for further multimodal processing." The PCM also includes the concept of information integration (referencing IIT), but notes that the treatment of information integration through the FE minimization and FoC constructs goes beyond what has been formally described by the authors of IIT.

The team set out to develop a model that "would give us a theory of consciousness that intelligibly links the phenomenology of first-person conscious experience with its realization in the brain (or other substrate) via a mathematical and computational model and simultaneously accounts for its biofunctional properties." This new theory of consciousness seems to have made major advances on this goal, and appears very promising for future development.

If you're going to read just one paper:

Williford K, Bennequin D, Friston K, and Rudrauf D (2018). The Projective Consciousness Model and Phenomenal Selfhood. *Frontiers in Psychology* 9:2571. https://doi.org/10.3389/fpsyg.2018.02571

THE DECISION TO ENGAGE MODEL

Shadlen and Kiani (2011) don't propose a fully developed theory of consciousness, but rather note potential close ties between the neural mechanisms of decision making and the neural mechanisms that could give rise to conscious states. Simple forms of sensory-motor decision making help an animal "engage the environment" by connecting sensations and perceptions of the external world to internal states (e.g., memories, appetitive states, near-term goals), and then computing informed courses of action. Similarly, the ultimate function of consciousness may be to help the individual engage the environment, but through increasingly sophisticated forms of decision making, including deciding to engage, "deciding to consider," or "deciding to decide to..."

Shadlen and Kiani define two types of consciousness that differ (perhaps helpfully) from other typologies. N-consciousness (for neurology) refers to consciousness as "a state of wakefulness with organized interaction with the environment. N-consciousness is absent in sleep, coma, general anesthesia, and generalized seizures, and present but impaired in states of delirium, some psychiatric illnesses, and some infectious diseases. P-consciousness (for philosophy) refers to the subjective features of consciousness, including "perceptual awareness, self-awareness, volition with awareness (i.e., authorship), a sense of free will, a sense of what it is like to be, a capacity to report narrative, introspection," etc. Their central thesis is that "similar neural mechanisms, computations and structures underlie many if not all of these forms of consciousness. The common feature is a decision to engage. Waking from sleep is a decision to engage the environment, and acting with awareness of purpose (authorship or will) involves a decision to engage a form of narrative associated with potential reportability."

Readers new to the decision making literature may find this use of "decision making" confounding, as we are accustomed to using the term to refer to fully conscious, volitional decisions. Shadlen and Kiani are referring to *decision making computations* performed by different neural circuits, as in the way sleeping brains make a "go" or "no-go" decision about sounds in the environment that should alert us or be ignored: birdsong at dawn might get a "no go," while a crying child in the night gets a "go."

Shadlen and Kiani use the neural mechanisms of visual motion perception to illustrate four key principles of neural decision making that could scale to support broader consciousness:

» Neural activity can evolve gradually over time to represent accumulation of evidence from multiple sources in time (thus integrating and unifying multiple inputs)

» There is a termination rule, or criterion for finishing a decision, which could be based on the amount of evidence, a temporal deadline, or other computation (cost, value, etc.)

THE DECISION TO ENGAGE MODEL

» Computations resemble probabilistic inference, meaning that neural activity can represent intensities that correspond to "degree of belief in..., " or "expected loss if"

» The neurons involved defy classification as sensory or motor. They lie at the nexus of sensory processing and motor planning.

It is easy to follow the authors supposition that there may be similar groups of neurons or circuits in the brain with similar properties that could make more complex decisions, moving, for example, from "performing an action to achieve a goal, in imitation of another's actions, i construing from another's action the goal that led to the other's action, in mirroring this goal – steps toward a neurobiology of "theory of mind."

Shadlen and Kiani contrast the *intentionality* of their framework with an "overenthusiasm for the representational framework" that leads to bizarre solutions that "elevate agnostic representations of information to the status of perceptions and awareness by oscillating it, synchronizing it, or enhancing its power spectrum in some frequency band." They find these latter solutions frustrating in that they don't answer the (to them, obvious) question about which brain structure *decides "to wiggle some part of the representation and thus render it available for conscious awareness?"*

The decision to engage is fundamentally an "evaluation of evidence leading to turning on another circuit and configuring the flow of information." This conceptualization shares some features in common with the broadcast function in GWT. In Shadlen and Kiani's model, N-conscious processes allow the brain to make more non-conscious decisions about what to engage, and P-conscious processes are decisions to engage primarily for communicating and reporting. While the processes by which P-conscious processes might actually emerge through a cascade of decisions are not fully elucidated, the authors do identify several brain structures that could play important roles, including the midbrain reticular formation and intralaminar nuclei of the thalamus, both of which play key roles in arousal, and the latter of which are part of

The concept of a decision to engage links the neurobiology of consciousness to the field of decision making. It has the virtue of tying together characterizations of consciousness employed in clinical neurology with the phenomenology that we associate with the mind's most precious pursuits. It may guide future experiments and, if correct, it would render broad areas of systems, cellular and molecular neuroscience relevant to the study of consciousness.
(SHADLEN & KIANI, 2011)

THE DECISION TO ENGAGE MODEL

the thalamic matrix, which may play a role in cortical circuit selection. Shadlen and Kiani also point to the cortical areas that are part of the default network, which appears to play a role in monitoring the world precisely when we are *not* engaged. Language areas may play a role in decisions to engage in narrative ways.

The decision to engage model has several strengths. It points to a set of mechanisms that are fairly well understood at the neurophysiological level in simpler forms of decision making, and it makes a compelling case for *where else* in the brain we might look for similar mechanisms at work. These assertions make this model more experimentally tractable than most, which is probably its most compelling strength. By focusing on decision making mechanisms that activate different bits of cortical circuitry, the model also allows for different forms of conscious experience among animals that have different cortical circuits and cognitive abilities (e.g., language areas) – another important strength. The notion that similar mechanisms could underlie different forms of consciousness (N- and P-) adds a nice explanatory efficiency, and provides a plausible evolutionary structure for the development of more complex forms of consciousness across animals with different brains.

The main weaknesses of the model are that 1) its parallel intentional architecture fails to capture the unity of conscious experience, and 2) it does not explain the subjective "what it is like to be" aspects of consciousness (qualia). The parallel intentional architecture could turn out to be a strength, they write, in comparison to global workspace theory, which seems to require a brain region or circuit with more computational and representational power and features than any they can identify in reality. While the authors admittedly don't know how a unified experience can arise from a parallel architecture, they see this outcome as being more plausible than a powerful central structure capable of organizing a global workspace that has yet to be identified. Regarding qualia, Shadlen and Kiani surmise that different *ways* of engaging the world via different brain structures could eventually provide an explanation for subjective experience, and note that "the decision to engage is the first building block of a subjective conscious experience, not the entirety of it."

If you're going to read just one paper:

Shadlen MN and Kiani R (2011). Consciousness as a decision to engage. In S Dehaene and Y Christen (eds.) *Characterizing Consciousness: From Cognition to the Clinic,* Research and Perspectives in Neurosciences. DOI 10.1007/978-3-642-18015-6_2.

Antonio Damasio's theories of consciousness: Damasio has written several books and given numerous public lectures about consciousness. While his conceptualization of consciousness has evolved, feelings have always been at its core. In his more recent writing, he moves beyond the brain to focus on the role of emotion in the body (and of the body in emotion). All animals with bodies and the nervous systems required to control them have some form of awareness, and the more complex the nervous system, the more elaborate and nuanced forms of subjective experience will be possible.

Across multiple books, Damasio makes the compelling case that if the experimental study of consciousness remains stuck in the world of sensory-motor perceptions and actions, we are unlikely to get to the part of consciousness that is so important in our lived experience: our feelings - about what it's like to be ourselves in any given moment. He writes in *The Strange Order of Things* (2018),

> "The aspect of mind that dominates our existence, or so it seems, concerns the world around us, actual or recalled from memory, with its objects and events, human and not, as represented by myriad images of every sensory stripe, often translated in verbal languages and structured in narratives. And yet, a remarkable yet, there is a parallel mental world that accompanies all those images, often so subtle that it does not demand any attention for itself but occasionally so significant that it alters the course of the dominant part of the mind, sometimes arrestingly so. That is the parallel world of affect, a world in which we find feelings traveling alongside the usually more salient images of our minds. The immediate causes of feelings include (a) the background flow of life processes in our organisms, which are experienced as spontaneous or homeostatic feelings; (b) the emotive responses triggered by processing myriad sensory stimuli such as tastes, smells, tactile, auditory, and visual stimuli, the experience of which is one of the sources of qualia; and (c) the emotive responses resulting from engaging drives (such as hunger or thirst) or motivations (such as lust and play) or emotions, in the more conventional sense of the term, which are action programs activated by confrontation with numerous and sometimes complex situations; examples of emotions

Consciousness is a particular state of mind, enriched by a sense of the particular organism in which a mind is operating; and the state of mind includes knowledge to the effect that the said existence is situated, that there are objects and events surrounding it. Consciousness is a state of mind with a self process added to it... Conscious states are felt.

(DAMASIO, 2010)

include joy, sadness, fear, anger, envy, jealousy, contempt, compassion, and admiration. The emotive responses described under (b) and (c) generate provoked feelings rather than the spontaneous variety that arises from the "unaffected" homeostatic flow."

Damasio's focus on *embodiment, especially the embodiment of feelings,* as a core concept in consciousness is a welcome change from much of the neuroscientifically-oriented work on consciousness. While it is difficult to outline his current, comprehensive model of consciousness without interviewing Damasio directly, his perspectives on the role of feelings, the body, and in his latest book, homeostasis and the evolution of consciousness, deserve to be part of a continuing conversation. He writes: "We can think of feelings as mental deputies of homeostasis... Feelings are *for* life regulation, providers of information concerning basic homeostasis or the social conditions of our lives (Damasio, 2018)."

If you're going to read just one book:

Damasio, A (2018). *The Strange Order of Things: Life, Feeling and the Making of Cultures.* New York: Pantheon.

Unlimited associative learning (UAL) as the marker of the evolutionary transition to minimal consciousness:

Echoing both Damasio on the evolution of consciousness and Axel Cleeremans on consciousness as something that the brain learns to do, Zohar Bronfman, Simona Ginsburg, and Eva Jablonka recently published a new theory of the evolution of consciousness that depends on learning, and varies across the animal kingdom. (Ginsburg and Jablonka have also just published a longer book, *The Evolution of the Sensitive Soul* (2019) on this theory.) At its core, they propose "unlimited associative learning (UAL) as the marker of the evolutionary transition to minimal consciousness (or sentience)." The criteria they outline as fundamental to sentience are influenced by a wide range of experts on consciousness, and include:

» "Flexible value systems and goals that reflect or give rise to the motivational values of the organism's ever-changing internal states and actions"

» "Unity and diversity through sensory binding leading to the formation of a compound stimulus; the multiple underlying features of the compound are coherently and conjointly perceived, rather than each feature being perceived independently"

» "Global availability of information, involving multidirectional feedback and reentrant interactions that generate a state in which information is available to different specialized cognitive processes that are otherwise 'computationally isolated'"

» "Temporal thickness – the temporal persistence of mental states"

» "Selection – involvement of processes of exploration and selective stabilization at different levels (neural, behavioral), including processes of action selection and selective attention"

» "Intentionality (aboutness). There are processes of representation/referral; inputs from the body and the world are "mapped" onto dynamic perception and action models that are necessary for the constitution of phenomenal consciousness"

» "Self and embodiment – no account of consciousness is possible without addressing the obvious fact that there is an agent that is sentient: it is the animal rather than its nervous system that is minimally conscious... We focus here on the animal's ability to form a representation of its body as distinct from the external world, yet embedded in it."

Bronfman, Ginsburg and Jablonka take a reverse-engineering approach to understanding sentience as thus defined, and look for an evolutionary marker that might have enabled the transition to sentience in the animal kingdom. The marker they propose is *unlimited associative learning* (UAL), which builds on traditional notions of associative learning that depend on an association being formed between a conditioned stimulus (e.g., the sound of a bell, or the push of a lever), and an unconditioned, naturally rewarding stimulus (e.g., drops of juice or a ripe berry). Since associative learning is predictive, only unanticipated events lead to learning, and the greater the difference between prediction and outcome, the stronger the learning signal. (This is similar to a core tenet of IIT.) The authors understand associative learning to include both "self-learning (learning only about the consequences of one's own actions – about *how* things are learned) and world learning (learning about *what* there is in the world through the reinforcing effects of relations between stimuli in the world that are independent of one's own action)."

Unlimited associative learning has more advanced criteria, which include:

» The conditional stimulus or the reinforced actions are *compound*; the animal learns to respond to a coherent set of stimulus features (e.g., color, shape, sound, temporal presentation, context) that don't activate the conditioned response alone. Similarly, the animal could learn that complex combinations of actions will result in a reward when individual components of those actions will not.

» The conditional stimulus and/or the reinforced actions are *novel*; they don't elicit reflexes or build upon past reinforcements.

» The learned CS or reinforced actions can support second order conditional learning, in which, for example, a conditioned stimulus becomes an unconditioned stimulus that is rewarding in itself.

The authors go to lengths to show how the capacity for UAL also underlies the emergence of the seven criteria for sentience. Their perspective is that if an animal has the capacity for UAL, it is endowed with sentience. However, animals without the immediate capacity for UAL, like infant humans, may still have sentience because they have the fundamental architecture for UAL in place, even if it is not yet fully functional.

Understanding UAL as forming the critical architecture for sentience can align well with other existing theories of consciousness, say the authors. For example, GWT requires perceptual, motor, memory, value and attention systems to unite to construct mental states, and these same systems are required for UAL. IIT, the authors say, "posits that consciousness is based on composite, integrated (irreducible and non-interdependent subsets), intrinsic cause-effect processes,...and that this integrated information (Φ) can be measured." Their model also required composition, integration and exclusion of information, but has additional requirements for sentience including the role of reinforcement and memory for compound patterns. The authors also see similarities to Damasio's work.

Animals with an integrated body, actions, and world models, that distinguish between self and world and act according to reinforcement signals based on these integrations in multiple domains of sensation and action, can be said to have a self.

UAL as an architecture for minimal consciousness represents a fundamentally new approach to understanding consciousness that is both rooted in evolution (the authors suggest that minimal consciousness emerged during, and may have helped drive, the Cambrian explosion), and offered as a positive marker for sentience (i.e., if an organism shows capacity for UAL, it is sentient). This framing takes us away from discussions about the specific neural architecture needed for consciousness, and focuses on the *capabilities* needed for consciousness - a move that could permit a much wider array of organisms and neural architectures to achieve the status of having minimal consciousness. While there is still work to be done in connecting UAL to the subjective experience of phenomenal consciousness, Bronfman, Ginsburg, and Jablonka offer a unique perspective on the evolution and functional requirements for minimal consciousness, and have done impressive work surveying a wide range of consciousness literature to try to distill the most common and critical elements of a definition of consciousness, which represents an important step forward for the entire field.

If you're going to read just one paper:

Bronfman ZZ, Ginsburg S and Jablonka E (2016). The transition to minimal consciousness through the evolution of associative learning. *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2016.01954

The theories described above were chosen primarily for their potential for empirical testability (and previous relevant experimentation), and for their contribution of new perspectives that counter or extend beyond the theories that TWCF and its network are currently focusing on (IIT and GWT). At the outset of this project, we had hoped to identify theories that were ready for direct experimental comparison to each other and/or to IIT and GWT. Unfortunately, each theory starts from a fundamentally different position, a lens through which the many problems of consciousness are framed, prioritized, and addressed or ignored. This isn't to say that they can't be experimentally probed and compared, only that they may need to be more fully aligned in terms of their positions, explananda and explanans, philosophical references, predictions, and hypotheses before head-to-head experiments can be envisioned and realized.

We think that the next step is to try to develop a taxonomic framework for empirically testable theories of consciousness. The first step in *that* process is to outline the key questions that empirically testable theories – when fully mature – ought to be able to answer. This would include questions about experiments that would advance the theory in some way, but may not extend yet to experiments that would compare theories head to head. Questions about each theory could include:

» What is the purpose of phenomenal consciousness, if any?

» Does consciousness require one or more representations of lower level experience? If so, what is the definition of representation? How and where are representations instantiated in the brain?

» How does the theory of consciousness define and explain the self? What role does the self play in organizing neural activity in ways that support or direct consciousness?

» Can non-biological systems be phenomenally conscious? What would be the requirements of such a system?

» Are there identified lower level neural mechanisms that could be plausibly scaled up to give rise to phenomenal conscious experience?

» What brain circuits are likely to be critical to supporting consciousness in humans? What about in other animals?

» To what extent does the theory of consciousness account for disorders of consciousness?

» How does the theory of consciousness account for emotional experience, especially of thoughts or ideas evoked in thought as opposed to experienced through the physical presence of a stimulus that provokes an emotional response?

» In this theory of consciousness, do animals need to *learn* to be conscious? If so, why? What kind of learning, and architecture for learning, are required?

» What role does attention play in the theory of consciousness?

» What is the evolutionary or homeostatic advantage of consciousness? How has it shaped the mechanisms of consciousness?

» What is being computed by the brain during phenomenally conscious states?

» What role does information play in this theory? What information is critical, and how and where is it processed?

» What philosophical lines of inquiry have influenced this theory?

» Are there critical aspects of the theory that are not testable? Are there ways to accumulate evidence that would support or not support those aspects, even if they can't be fully proven or ruled out?

Some of these questions are already addressed by the authors of the theories above (as well as IIT and GWT), but not *all* questions are consistently addressed, nor can they be at the outset. Shadlen and Kiani, for example, start from an observed neural mechanism that they think could scale and *enable consciousness,* whereas Graziano begins by asking how the illusionist philosophical tradition might relate to our collective knowledge about how attention works and to the logic of model-based control. Shadlen and Kiani are unlikely to have an advanced answer to questions about rootedness in philosophical theories, and Graziano may not have elaborate answers about neurobiological mechanisms.

That said, we think that one of the best ways that TWCF could advance the field of consciousness would be to develop a kind of taxonomic questionnaire for developing theories of consciousness, and ask the representatives of each theory to contribute not only their answers but also additional questions for others. The next step would then be to convene the authors of these theories. Rather than simply presenting and advocating for his/her own theory, each participant would be asked to question and help improve *another person or group's theory,* and also to answer questions from the team at work on advancing his/her theory. Another option would be to have representatives from multiple theories work together on a piece of a larger "alignment" project, answering questions together like, "Neurological states (e.g. dreaming) or conditions (e.g. blindsight) a theory of consciousness ought to be able to account for include..."

We believe that this kind of socializing and aligning of theories of consciousness is a necessary prerequisite for designing experiments that might be able to enhance or diminish the evidence for one theory versus another. But perhaps more importantly, by asking experts who have mostly focused on their own theory to dive deep into theories advanced by others, we think that you could see some theories collapse or combine in fruitful ways. Some theories might also be elaborated with useful ideas from others, or developed more comprehensively with help from one of the other lenses described above.

EXPERTS TO CONSIDER INVITING TO A FUTURE FORUM THAT INCLUDE THE THEORIES ABOVE:

In addition to the neuroscientists and philosophers you have already brought together on consciousness to date, we recommend including the following people at future events or in interviews, all of whom contributed to one of the theories described above:

Daniel Bennequin
Université Pierre et Marie Curie-Université Paris Diderot (Projective Consciousness Model)

Zohar Bronfman (graduate student)
Tel Aviv University (Evolution of Consciousness)

Richard Brown
City University of New York (Higher Order Theories/Higher Order Representations of Representations Theory)

Axel Cleeremans
Université Libre de Bruxelles (Radical Plasticity Theory)

Antonio Damasio
University of Southern California

Karl Friston
University College London, (Projective Consciousness Model)

Simona Ginsburg
Open University of Israel (Evolution of Consciousness)

Michael Graziano
Princeton University (Attention Schema Theory)

Eva Jablonka
Tel Aviv University (Evolution of Consciousness)

Roozbeh Kiani
New York University (Decision to Engage Model)

Hakwan Lau
University of California, Los Angeles (Higher Order Theories)

Joseph LeDoux
New York University (Higher Order Theories)

Thomas Metzinger
Johannes Gutenberg-Universität Mainz (HOT; Phenomenal Self Model)

David Rosenthal
City University of New York (Higher Order Theories)

David Rudrauf
University of Geneva (Projective Consciousness Model)

Michael Shadlen
Columbia University (Decision to Engage Model)

Taylor Webb (postdoc)
Princeton University (Attention Schema Theory)

Kenneth Williford
University of Texas, Arlington (Projective Consciousness Model)

The following people are also doing interesting work and/or producing insightful commentary about consciousness and could help elaborate and question some of the theories above:

Susan Blackmore
University of Plymouth

Ned Block
New York University

Sam Parnia
New York University

Adina Roskies
Dartmouth College

Anil Seth
University of Sussex

We would also suggest adding at least these two philosophers, both of whom have deep backgrounds in neuroscience. Neither works specifically on consciousness (and thus are less likely to be dedicated to a specific position on consciousness), but both are very thoughtful about explanations in the brain and connecting the theories of philosophy to the experiments of neuroscience and the biological "stuff" of the brain.

Rosa Cao
Stanford University

Carl Craver
Washington University in St. Louis

ADDITIONAL REFERENCES (NOT INCLUDING THOSE LISTED ABOVE)

Baars BJ (1989). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.

Block N, Carmel D, Fleming SM, Kentridge RW, Koch C, Lamme VA, Lau H and Rosenthal D (2014). Consciousness science: real progress and lingering misconceptions. *Trends in Cognitive Sciences*. 18, 556–557. DOI: 10.1016/j.tics.2014.09.004

Chalmers DJ (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2: 200-19. DOI: 10.1.1.103.8362.pdf

Cleeremans A (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology* 2: 86. DOI: 10.3389/fpsyg.2011.00086.

Damasio A (2010). *Self Comes to Mind: Constructing the Conscious Brain*. New York, USA: Pantheon/Random House.

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

Metzinger, T (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, USA: MIT Press.

Michel M, Fleming SM, Lau H, Lee ALF, Martinez-Conde S, Passingham RE, Peters MAK, Rahnev D, Sergent C, and Liu K (2018). An informal internet survey on the current state of consciousness science. *Frontiers in Psychology* 9:2134. DOI: 10.3389/fpsyg.2018.02134

Oizumi M, Albantakis L, and Tononi G (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Computational Biology*. DOI: 10.1371/journal.pcbi.1003588

Rosenthal DM (2008). Consciousness and its function. *Neuropsychologica* 46: 829-40. DOI: 10.1016/j.neuropsychologia.2007.11.012.

The Stanford Encyclopedia of Philosophy entries on "Consciousness" (6/18/04; revised 1/14/14); "Higher-Order Theories of Consciousness" (4/3/01; revised 8/29/16); "The Neuroscience of Consciousness" (10/9/18); and "Representational Theories of Consciousness" (5/22/00; revised 4/17/15).

Tononi, G. 2004 An information integration theory of consciousness. *BMC Neuroscience 5*: 42. DOI: 10.1186/1471-2202-5-42.

Tononi, G. 2008 Consciousness as integrated information: A provisional manifesto. *Biological Bulletin* 215, 216-42. DOI: 10.2307/25470707.

Tononi G, Boly M, Massimini M and Koch C (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* 17:450–461. DOI: 10.1038/nrn.2016.44