COGNITIVE NEUROSCIENCE
https://doi.org/10.1080/17588928.2020.1772214

Routledge
Taylor & Francis Group

Hard criteria for empirical theories of consciousness

Adrien Doerig[a], Aaron Schurger[b,c,d,e] and Michael H. Herzog[a]

[a]Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale De Lausanne (EPFL), Lausanne, Switzerland; [b]Department of Psychology, Crean College of Health and Behavioral Sciences, Chapman University, Orange, CA, USA; [c]Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Irvine, CA, USA; [d]INSERM, Cognitive Neuroimaging Unit, Gif sur Yvette 91191, France; [e]Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin center, Gif sur Yvette 91191, France

ABSTRACT

Consciousness is now a well-established field of empirical research. A large body of experimental results has been accumulated and is steadily growing. In parallel, many Theories of Consciousness (ToCs) have been proposed. These theories are diverse in nature, ranging from computational to neurophysiological and quantum theoretical approaches. This contrasts with other fields of natural science, which host a smaller number of competing theories. We suggest that one reason for this abundance of extremely different theories may be the lack of stringent criteria specifying how empirical data constrains ToCs. First, we argue that consciousness is a well-defined topic from an empirical point of view and motivate a purely empirical stance on the quest for consciousness. Second, we present a checklist of criteria that, we propose, empirical ToCs need to cope with. Third, we review 13 of the most influential ToCs and subject them to the criteria. Our analysis helps to situate these different ToCs in the theoretical landscapeand sheds light on their strengths and weaknesses from a strictly empirical point of view.

ARTICLE HISTORY
Received 9 December 2019
Revised 14 April 2020
Published online 14 July 2020

KEYWORDS
Consciousness; theories; criteria

I The problem of consciousness

Descartes (1637/1996) is usually considered as the founder of the mind-body problem, the precursor of the modern topic of consciousness. He proposed a dualistic framework, in which mind and matter are two separate ontological entities, interacting at the pineal gland. Such dualistic theories are not easy to reconcile with the fundamental laws of physics, namely, the conservation of energy and impulse. For this reason, most modern philosophical and scientific theories of consciousness are compatible with physicalism, in which only matter exists and mental events are identical to or supervene on physical processes (Stoljar, 2017). Many philosophical approaches, however, deny that consciousness can be reduced to physics (e.g. Chalmers, 1996). At the center of the debate is the notion of qualia, the subjective phenomenal qualities of experience: how it is to feel pain, experience a shade of green, or a wonderful piece of music. Levine (1983) introduced the term *explanatory gap* to highlight the difficulty of physicalist theories in explaining phenomenal properties; Chalmers (1996) called it the *hard problem*.

This philosophical debate has been ongoing for decades, hovering around the question of whether we can close the explanatory gap or whether it is impossible for principled reasons. At the highest level, there are two different attitudes. Realists propose that consciousness really exists as a distinct property, and the job of science is to understand how it is generated (e.g. Chalmers, 2004). In contrast, illusionists suggest that we are simply wrong about the nature of consciousness (e.g. Dennett, 2016; Frankish, 2016ba): consciousness, as a non-physical property, is an illusion and the relevant question is: why do people feel that they have such a non-physical property 'inside' them? Another important theoretical dispute concerns the distinction between access vs. phenomenal consciousness (see Block, 1995; Cohen & Dennett, 2011; Naccache, 2018; Phillips, 2018; Ward, 2018). One of the major problems in these discussions is that consciousness seems to evade a rigid definition, making it difficult to pit theories against each other like in other scientific disciplines.

Here, we leave these hard philosophical problems aside and focus on *empirical* approaches to consciousness only, as suggested in the seminal work of Crick and Koch (1990). We wake up every day and change from an unconscious to a conscious state. Obviously, there is something to explain. We can render visible stimuli invisible by backward masking. Obviously, there is something to explain here too. Because of these clear empirical phenomena, there is no need to posit a theoretical definition. Empirical science starts with good observations and aims for definitions.

CONTACT Adrien Doerig adrien.doerig@gmail.com Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale De Lausanne (EPFL), Lausanne, Switzerland

Consciousness science is ripe for such a purely empirical approach. Indeed, a large body of detailed empirical data has been accumulated. Nevertheless, as mentioned, there are dozens of very different ToCs and there is no clear, largely accepted 'winner'. However, not all these ToCs can be correct. We suggest that this conceptual quagmire may result from a lack of stringent criteria guiding how to address consciousness empirically. Here, we will propose a list of such criteria. We will use this checklist to compare current ToCs, aiming to foster constructive discussions about how to address and explain consciousness. The list is neither proposed to be exhaustive nor are the criteria written in stone. Rather, we see this list as a steppingstone to compare ToCs in a structured manner by working out criteria that describe how predictions of a ToC can be falsified empirically and how to pit theories against each other. Currently, there are very few mutual comparisons between ToCs. Instead, authors build on their own ToCs, largely ignoring others (but see Ball, 2019).

Importantly, ToCs often make explicit or implicit metaphysical assumptions, such as subscribing to illusionist or realist ideas or relying on access or phenomenal consciousness. As mentioned, the focus of this contribution is only on how ToCs address *empirical data* about consciousness, such as masking, rivalry, or the difference between sleep and wakefulness. All scientific theories need empirical support regardless of their metaphysical assumptions, and therefore need to address the criteria. Because the criteria are geared toward explaining empirical data, they are strongly linked to behaviour. However, this does *not* mean that ToCs need to be behaviourist or functionalist. Our criteria are neutral regarding metaphysical assumptions and also apply to theories focussing on phenomenal consciousness, which require data if they want to be considered as empirical theories.

Another important note is that this contribution is not aimed to compare *specific* ToCs with each other but to pit common *ideas and principles* of ToCs against each other. This is not a catalog of the latest version of each ToC, but a critical evaluation of basic arguments in consciousness research.

The following points spell out what the aims of this contribution are, and what is not addressed:

(1) Consciousness can be approached from a purely empirical stance.
(2) There is a bewildering number of ToCs, suggesting that stringent criteria are missing.
(3) This contribution is a steppingstone to build such criteria and provide a compass to locate ToCs in a space of these criteria.
(4) Not all current ToCs can be correct. Hence, the ultimate goal is to reduce the number of ToCs, subsuming current ideas into mutually exclusive theories making clear cut, novel, and empirically testable predictions that, if confirmed, will exclude one theory or another (which is rare in consciousness research).
(5) This contribution is not a review of specific ToCs. It is about general arguments in the field.
(6) We think the field is ripe for a coherent research program, which goes beyond proposing consciousness is *identical* with another phenomenon X.
(7) This contribution is *not* about metaphysics, such as realism vs. illusionism. However, ToCs need to meet criteria regardless of metaphysical assumptions if they claim support from empirical data.

II Empirical phenomena of consciousness

In all sciences, it is important to have a good description of the phenomenon of interest, which subsequently needs to be explained by theories. In the following we list several classic distinctions about consciousness.

II.1. Does the theory address the content or the state of consciousness or both?

From an empirical perspective, there are two main aspects of consciousness (Chalmers, 1996; Searle, 2000). First, there are unconscious *states*, such as non-REM sleep and anaesthesia. These are contrasted with conscious *states*, such as being awake (there are also more ambiguous states, such as hypnosis or meditation).

The second main avenue in consciousness research concerns the *content* of consciousness. For example, in visual masking, a target element is followed by a mask, which renders the target not consciously perceived (Figure 1a). The not-conscious target can still influence visual information processing, i.e., its representation is active and influential in the human brain (see Peters & Lau, 2015, for difficulties in studying unconscious perception). For example, words can be semantically processed without being consciously perceived (Gaillard et al., 2006) and expert chess players can analyse chess situations without consciousness (Kiesel et al., 2009). Likewise, in binocular rivalry, different images are presented to the two eyes. When the images are not compatible, only one of the images is perceived and the other one is suppressed (Figure 1b). When the images are of 'equal' strength, they rival, i.e., the suppressed image becomes conscious after a few seconds and the previously visible image is rendered unconscious. During the unconscious

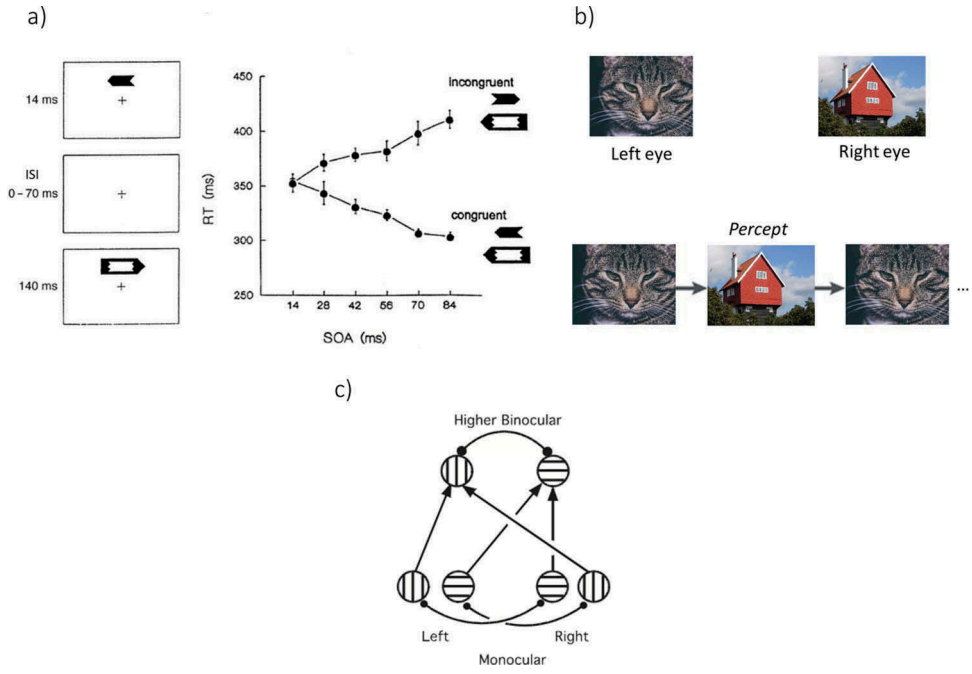

Figure 1. a) Masking. A briefly presented prime (left pointing arrow) can be rendered invisible by a trailing mask (arrow with central gap). Even though observers do not consciously perceive the prime, it can still affect unconscious brain processing. For example, observers are asked whether the masking arrow points either to the left or right. When the prime and mask arrows point in the same direction (congruent trials), reactions are faster than when the arrows point in opposite directions (incongruent trials). Figure reproduced with permission from Vorberg et al. (2003) Copyright (2003) National Academy of Sciences, U.S.A. b-c) Binocular rivalry. b) Two different images are shown to the left and right eye at the same time. When the images are not compatible, only a single image is perceived at a time. After a few seconds, there is a switch and the other image is perceived. Hence, the content of consciousness alternates between the two images. In this example, a cat was presented to one eye and a house to the other. Figure reproduced from Doerig, Schurger, et al. (2019). c) An example of a six-neuron network explaining binocular rivalry. Figure reproduced with permission from Wilson (2003) Copyright (2003) National Academy of Sciences, U.S.A.

periods, the suppressed image can still influence information processing. For example, even if an image is fully suppressed for the entire duration of the experiment, humans can unconsciously learn about its features (Seitz et al., 2009). Other paradigms used to study the content of consciousness include continuous flash suppression (a version of binocular rivalry; Tsuchiya & Koch, 2005), change blindness (Simons & Levin, 1997) and crowding (Atas et al., 2014; Bouma, 1973; see Breitmeyer, 2015; S. Dehaene et al., 2017; Kim & Blake, 2005, for reviews of empirical methods for consciousness).

ToCs may address the content or the state of consciousness, or ideally both. In practice, ToCs are not always clear about which of these aspects they address.

II.2. Is consciousness graded or binary?

We have the intuitive feeling that state consciousness is a gradual phenomenon. There is a continuum from death, to coma, anaesthesia, drowsiness, and fully alert states. Even within these states there seems to be continuity. For example, depth of anaesthesia can be classified by the bi-spectral (BIS) index. However, it is unclear to what extent this intuition is correct. Another open question is whether the *content* of consciousness is gradual.

II.3. Is consciousness unitary?

Although we feel that we have only a single consciousness at one moment in time, Moutoussis and Zeki (1997) argued that the brain may hold many simultaneous consciousnesses because motion and colour are perceived (consciously) at different moments in time, reflected by different neural activities in different brain regions. Most other approaches to consciousness propose that consciousness is unitary.

II.4. Is consciousness temporally continuous or discrete?

Unconscious targets can be re-rendered conscious when a second mask follows the first (Breitmeyer & Öğmen, 2006). Therefore, there must be an extended period of unconscious processing since otherwise there could be no recovery from masking (had the first mask

irretrievably suppressed the target representation, the second mask could not render it conscious again). A similar finding is that a masked stimulus can be re-rendered conscious by a transcranial magnetic stimulation (TMS) pulse (Ro et al., 2003). This unconscious integration period can last for 420 ms as experiments using TMS and feature fusion have shown (Scharnowski et al., 2009; Rüter et al., 2010; see also Pilz et al., 2013). Drissi-Daoudi et al. (2019) have also shown mandatory unconscious integration up to 450 ms. Relatedly, Sergent et al. (2013) showed that a cue presented *after* a target stimulus that would otherwise remain unconscious can retrospectively render the stimulus conscious. These results raise the question of whether consciousness is a continuous stream of percepts or is discrete, i.e., consciousness occurs at only certain time points (Doerig, Scharnowski, et al., 2019; Fekete et al., 2018; Herzog et al., 2016; James, 2013; VanRullen & Koch, 2003; White, 2018).

II.5. What is the fate of unconscious elements?

As mentioned, even when elements are not consciously perceived, e.g., because they are masked, they can still be fully processed as objects and influence conscious processes. Even more surprisingly, features of unconscious elements can be visible at other consciously perceived elements, presented at different times and locations (Nishida et al., 2007; Otto et al., 2006). For example, Otto et al. (2006) used masking to render a target unconscious. They showed that features of this unconscious masked target can be perceived consciously as features of unmasked elements presented later and at different spatial locations. What differentiates processing of unconscious features and elements from conscious ones? If processing of unconscious features is fundamentally different from conscious ones how can they interact with each other?

III Hard criteria for empirical theories of consciousness

While the empirical literature blossomed, a comparable bonanza of ToCs emerged. These theories vary greatly in terms of what they aim to explain and how they explain it – ranging from quantum theories aimed at explaining human understanding, to computational theories aimed at explaining masking and attentional blink experiments. Here, we present a list of criteria that we propose all empirical ToCs need to cope with regardless of their underlying metaphysical positions.

Importantly, we see this list as a starting point and other criteria may be proposed too. However, we propose that this list is constraining enough to provoke interesting discussions about how to study consciousness empirically.

There are obvious criteria that any ToC needs to meet. For example, the scope of the theory must be clear: is the theory about the content of consciousness, the state of consciousness, or both? Moreover, as for any scientific theory, the proposed mechanism must be both *necessary* and *sufficient* to explain data about consciousness. In the following, we outline four further important criteria specific to consciousness research.

III.1. Paradigm cases of consciousness & the unconscious alternative

Paradigm cases of consciousness (Aaronson, 2014) are empirical phenomena focussed specifically on consciousness, and not on other co-occurring aspects. They must have conscious *and* unconscious alternatives, allowing us to pit conscious vs. unconscious processing against each other – otherwise one cannot be certain that consciousness *per se* and not other co-occurring processes are addressed (this is reminiscent of Baars (1986) contrastive approach). In masking, for instance, the target is perceived consciously under certain conditions and it remains unconscious under other conditions. Going from long to short inter-stimulus intervals (ISIs) between target and mask leads to a change from full target visibility to zero visibility. For certain ISIs the target is consciously perceived in 50% of trials. As mentioned, in trials where the target is not consciously perceived, it can still influence processing (Figure 1a). This is the unconscious alternative: the target is processed but not consciously perceived (Figure 2a&b). To explain a paradigm case, ToCs need to show how changes in their mechanism explain changes from consciousness to unconsciousness (Figure 2c).

Importantly, it follows that introspection by itself does not allow one to study paradigm cases of consciousness. Indeed, introspection only addresses the conscious alternative since, by definition, we are unaware of the unconscious alternative. For example, when we are unaware of the masked target in a masking experiment, introspection does not provide any information about unconscious processing. Hence, although introspection may seem to offer a privileged window on consciousness, well-controlled experiments are also needed to address paradigm cases.

Even though problems with this contrastive approach have been highlighted recently (Aru et al., 2012; Balsdon & Clifford, 2018; Kleiner & Hoel, 2020; Lau, 2008; Peters & Lau, 2015), all that is needed is the existence of conscious and unconscious states, and methods to contrast

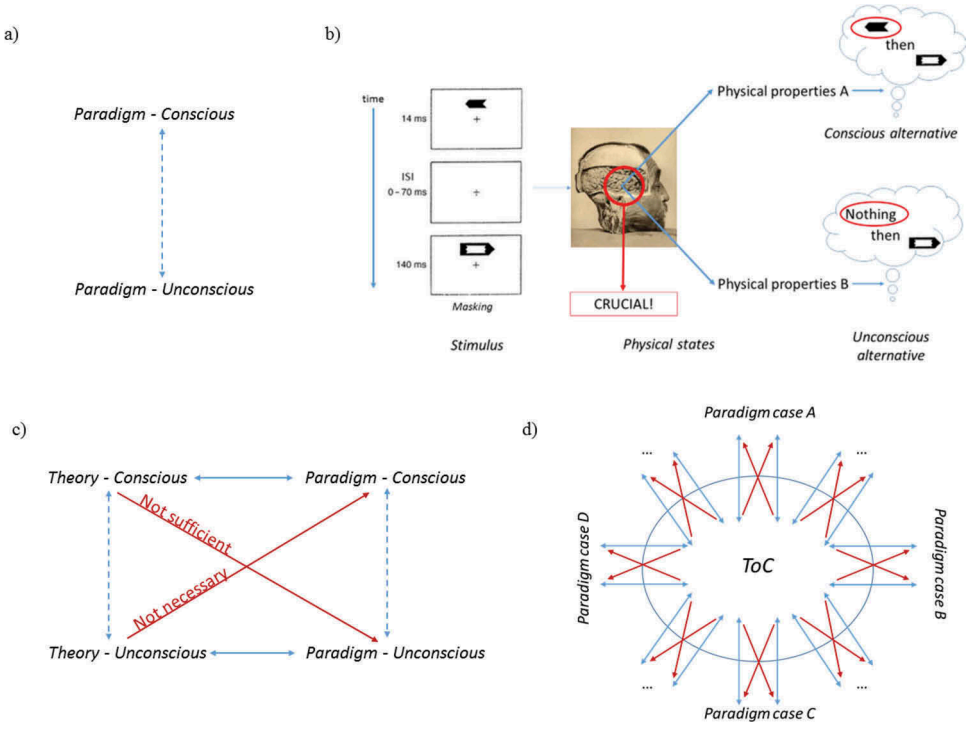

Figure 2. Paradigm cases of consciousness. a) In paradigm cases of consciousness, first, the phenomenology of consciousness is clear cut and, second, there are both a conscious and an unconscious alternative (for example wakefulness vs. death, perceived vs. non-perceived image in rivalry, visible vs. invisible prime in masking, etc). In all these cases, there is obviously something to explain about consciousness. b) In a masking experiment, we can fix one intermediate ISI for which the prime is consciously perceived on certain trials and not on others. The mechanism of a ToC needs to explain what causes this difference. A second avenue is to change the ISI parametrically. For short ISIs, the prime is invisible (the direct measure is zero). For longer ISIs, the prime is clearly visible. The mechanism needs to explain in a parametric manner how consciousness emerges as a function of key components of the mechanism. One caveat with this method is that the mechanism should reflect a change in consciousness and not only in stimulus strength. There are methods to double dissociate the two (Schmidt & Vorberg, 2006). c) Theories must explain paradigm cases by proposing a mechanism explaining in which cases the conscious/unconscious alternatives occur. To this end, changes in the mechanism must reflect the observed changes of the paradigm case. A mechanism is not necessary if consciousness occurs when the mechanism predicts unconsciousness and it is not sufficient if the mechanism predicts consciousness but there is no consciousness. A mechanism that is neither necessary nor sufficient has no explanatory power and must be rejected because there is a double dissociation between the mechanism and consciousness. d) An ideal ToC proposes a necessary and sufficient mechanism for each paradigm case. This can be achieved by showing that changes in the mechanism account for changes from the conscious to the unconscious alternatives for *all* paradigm cases of consciousness. Explaining all paradigm cases is important, because a theory that can only explain binocular rivalry for example is just a theory of rivalry, not a ToC.

them (but see Salti et al., 2019, for an argument that the conscious vs. unconscious dichotomy is misguided). For visual masking and other paradigms, specific procedures have been established to differentiate between conscious and unconscious perception (Morales et al., 2015; Overgaard et al., 2010; Schmidt & Vorberg, 2006).

Paradigm cases with an unconscious alternative ensure that consciousness is the dependent variable in experiments, and contrast with approaches where only conscious states are investigated. We already mentioned introspection. Other approaches study only conscious (and not unconscious) states and their changes, such as how consciousness changes under the influence of LSD, other psychotropic substances or meditation (Carhart-Harris et al., 2014; Lutz et al., 2007), or track changes in body ownership and map them to brain states (Faivre et al., 2015). Other approaches target specific processes thought to be constitutive for consciousness, such as neural binding (Edelman, 2003), learning (Cleeremans, 2007), or insight (Hameroff & Penrose, 2014). Because these approaches lack an unconscious alternative, they need to cope with the problem that they may be addressing other co-occurring processes instead of consciousness *per se* (see section IV.1).

It is not enough to explain just a few paradigm cases. A theory of consciousness must be more general than a theory of binocular rivalry, otherwise it is merely a theory of rivalry. For example, the 6-neuron model of rivalry in Figure 1c is not a model of consciousness. A rich phenomenology of consciousness is needed to

move beyond such models. The underlying commonalities between all paradigm cases need to be understood (Chalmers, 1996; Fingelkurts et al., 2012; Haynes, 2009; Seth, 2016). Hence, ideally, a ToC should explain paradigm cases by a principled mechanism (Figure 2d). It may also be that different aspects of consciousness need different mechanisms (for example, content and state consciousness may involve different mechanisms).

In summary, our first criterion asks whether a ToC addresses paradigm cases of consciousness. If it does not, the ToC needs to provide a principled argument showing that it nevertheless targets consciousness *per se*. In addition, it is important for a ToC to provide principled reasons why it targets consciousness in general, beyond the specific paradigm cases it addresses.[1]

III.2. The unfolding argument

An important aspect is the level on which ToCs describe consciousness. For example, many ToCs situate consciousness at the level of functions performed by a system, while causal structure theories focus on the implementation level. Causal structure theories are ToCs that associate consciousness with the presence of the 'right' kind of causal structure in a system. Consciousness occurs when neurons are connected in just the 'right way'. Information Integration Theory (Tononi, 2004) and Recurrent Processing Theory (Lamme, 2006) are examples of causal structure theories. Both theories imply that recurrent processing is necessary and sufficient for consciousness.

It is a mathematical fact that both recurrent and feedforward networks (along with many other kinds of universal function approximators) can approximate any input-output function to any degree of accuracy (Hornik et al., 1989; Schäfer & Zimmermann, 2006). The inputs may be the stimuli of a rivalry experiment and the outputs are the participant's responses such as button presses. Like all functions, this sensorimotor processing can be implemented equivalently in feedforward and recurrent networks. One can always *unfold* a recurrent network into a functionally equivalent feedforward network and vice versa.

Hence, feedforward and recurrent systems can perform in exactly the same way in a rivalry experiment, for example. The same is true for any possible experiment about consciousness. According to causal structure theories, the recurrent system is conscious whereas the feedforward system is not. As a consequence, the unfolding argument challenges causal structure theories because, if one accepts that two functionally identical systems can have different consciousness, these ToCs cannot be addressed empirically, and are therefore outside the realm of science (see Doerig, Schurger, et al. (2019) for a detailed presentation of the Unfolding Argument; Hanson and Walker (2019) and Kleiner and Hoel (2020) for related arguments, and Kleiner (2019) and Tsuchiya et al. (2020) for replies).

As an example, Recurrent Processing Theory proposes that visual consciousness arises in primary visual cortex when recurrent processing kicks in (Lamme, 2006). However, these regions can in principle be replaced by feedforward equivalents without changing anything about their input-output function. Since there is no functional difference, a patient with such an implant behaves identically and does not report any difference. Hence, since the mechanism suggested by Recurrent Processing Theory can be replaced without any change in data about consciousness, this mechanism cannot be necessary to explain data about consciousness.

Hence, our second criterion asks whether a ToC is subject to the unfolding argument. If so, the ToC needs to explain how experiments can support or falsify it.

III.3. The small & large network arguments

III.3.a. Conscious small networks

Herzog et al. (2007) showed that many ToCs imply that small networks with fewer than ten neurons are conscious. There are two alternative stances to deal with the small network argument.

III.3.b. Fixing theories with additional constraints

First, a ToC may reject that the small networks are in fact conscious. In this case, the theory is not sufficient for consciousness because it cannot explain why the small networks are not conscious. Something is missing from the theory. To avoid this problem, it is often proposed that only small, non-crucial addendums need to be added.

Size. For example, it was claimed that full consciousness occurs only in networks with many neurons (Taylor, 2007). However, a large network with millions of neurons 'behaves' identically to a single neuron if all neurons do the same thing. Thus, size by itself does not suffice.

Complexity. An alternative proposal may be that complexity is crucial. However, for many definitions of complexity, even very small networks with less than 10 neurons exhibit high complexity (Herzog et al., 2007; Oizumi et al., 2014).

Crucial additional ingredients. Even if it were possible to 'fix' these ToCs by adding just the 'right' ingredients, it

[1]These latter two requirements echo the exclusiveness and exhaustiveness requirements of Reingold and Merikle (1988).

may be that these additional ingredients are more important to explain consciousness than the proposed mechanism itself. In general, when a mechanism is proposed to be only necessary, it needs to be clarified how much of the variance the mechanism explains. How crucial is it for consciousness?

III.3.c. Panpsychism

Alternatively, a ToC may propose that small networks are conscious. In this case, the ToC implies a form of panpsychism: all or most systems in the universe are conscious to varying degrees. A criterion needs to be given to specify exactly which systems are conscious. This criterion is not constrained by empirical data because there are no paradigm cases for small networks. Therefore, it must always be chosen arbitrarily. Hence, unless a principled argument is given to show why *this* ToC with *these* criteria is correct, panpsychist ToCs are not sufficient to explain consciousness. See Seth (2018) and Frankish (2016b) for other methodological problems with panpsychism, and Goff (2017) for a defence of panpsychism.

III.3.d. The large network argument

Accepting consciousness in small networks also poses a challenge for *large* systems. The human brain contains more than ten billion neurons. If networks with less than ten neurons can be conscious, the human brain may contain billions of conscious subsystems. One needs to provide an additional non-arbitrary criterion explaining which subsystem is conscious, or to deny the unity of consciousness. Unless a principled argument is given to show why *this* ToC with *these* criteria is correct, the ToC is not sufficient to explain consciousness. This problem is known in philosophy as the combination problem (Chalmers, 2017).

In summary, our third criterion asks whether ToCs are subject to the small (and large) network argument and, if they are, how they deal with it. Do they embrace panpsychism and, if so, how do they cope with the underdetermination issues this raises? It is important for a ToC to be explicit about these points.

III.4. The other systems argument

A ToC should be able to determine which systems, apart from awake humans, are conscious. Hence, ToCs focussing on the normal human brain must either be generalizable to other systems, or provide a strong argument showing why only humans can be conscious.

For example, it was proposed that consciousness is mediated by thalamo-cortical interactions (Llinas et al., 1998). Thus, a question is whether animals without a thalamus can have consciousness? How about the human in Figure 3? One simple reply would be that removing an adult's thalamo-cortical system makes the person unconscious, so it must be that the human in Figure 3 has an equivalent of the thalamo-cortical system. However, then the question is which properties shared by the thalamo-cortical system and its equivalent are important for consciousness?

This question is particularly difficult given the strong multiple realizability of many phenomena.[2] It is well known that any function can be implemented by different physical systems (Bechtel & Mundale, 1999; Fodor, 1974). A word processing system such as Microsoft Word can be run on many operating systems such as Unix, MacOS or Windows. Phenomenologically, there is no difference. However, there are large differences in the software and hardware implementations. Hence, simply pointing to the software and hardware used in the Windows implementation can neither explain what Microsoft Word is nor indicate in which other systems it can be instantiated. More precisely, although the software and hardware used in that implementation are sufficient to instantiate Microsoft

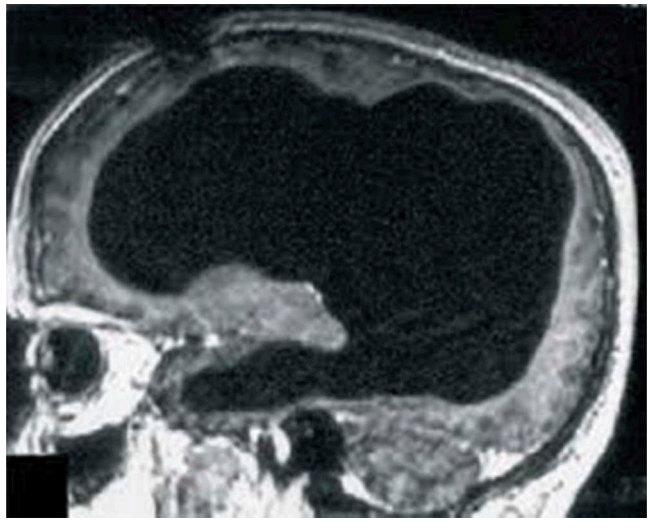

Figure 3. A sagittal scan of a person with strongly reduced brain volume. The patient lives a normal life, has no cognitive problems, and is as conscious as any other human. It is not easy to map explanations of consciousness based on brain anatomy and connectivity to the brain of this patient. Reproduced with permission from (Feuillet et al., 2007).

[2]Multiple realization has been a fruitful topic in philosophy (Fodor, 1974; Kripke, 1972; Putnam, 1967, 1988). Classically, multiple realizability has been used in metaphysical debates such as identity-theories vs. functionalism, or in the context of theory reduction. Here, we use it in a different way: ToCs need to offer a gauge to determine whether or not a given system is conscious.

Word, they are not necessary. Likewise, ToCs that explain consciousness by pointing to certain brain regions or characteristics claimed to be sufficient for consciousness need to explain why they are also *necessary* for consciousness.

Hence, our fourth criterion asks whether ToCs can make clear-cut and specific predictions about which other systems are conscious, apart from humans. If they cannot, they are subject to the other systems argument. If they are subject to the argument, how do they cope with it? Do they claim that only humans are conscious? Otherwise, they are not necessary for consciousness. It is possible for a ToC that lacks the specificity, in its current form, to address the other-systems criterion to later be able to address the criterion after becoming more mechanistically precise.

IV. Facing up to the criteria

Guided by the criteria of the last section, we use the most prototypical ToCs to work out leitmotifs in argumentation, which are dominant in consciousness research. A quick synopsis tries to capture the essentials of these ToCs. More ToCs are described in the appendix. Our analysis is summarized in Table 1. We are aware of the shortcomings of such a minimalistic approach. First, such a review cannot be exhaustive because so much has been said about consciousness (it is potentially quicker to list phenomena to which consciousness has *not* been linked). Still, we think that ToCs not portrayed have much in common in argumentation with the portrayed ToCs. Second, we are very much aware that it is impossible to perfectly portray a ToC in a few lines because these theories are complex, multi-faceted and have seen many updates. For this reason, we are not interested in minute details of specific ToCs. Once more, we are more interested in prototypic argumentations than in an encyclopedia of ToCs. For this reason, we do not wish to imply that when ToCs cannot cope with one of the criteria they should be dismissed. There might be simply fixes. Our aim is to make challenges explicit and

Table 1. Summary of how ToCs cope with our criteria. The classification presented here should be seen as our interpretation, and aims to foster discussion. **Top: Criteria**. Green indicates that a ToC successfully copes with the associated criterion, red that it does not. Orange boxes indicate that the ToC faces either the small network or the other systems argument, depending on how one interprets the ToC. This is due to loosely defined mechanisms: if the mechanism is interpreted as a straightforward computational mechanism (such as a simple implementation of higher-order 'thoughts' in a 2-stage network), the small network argument arises. If instead the mechanism is a complex process found only in the human brain (for example 'thought' understood as a brain process), the other systems kicks in. We propose that theories with orange boxes may be unable to simultaneously avoid the small network and other systems arguments. Details are given in the corresponding parts of section IV and in the appendix. Theories are described in section IV or in the Appendix. **Bottom: The scope of theories**. Light green indicates that the ToC explicitly addresses the associated empirical characteristic of consciousness. Light red means it does not. The empirical characteristics are described in section II. As mentioned, ToCs are not always explicit about the scope of what they address and what exactly they aim to explain. For this reason, the above table partly reflects our own understanding of ToCs. In addition, different versions and interpretations of a ToC can change the above classifications. This table can be seen as an invitation to discuss, take sides and make commitments. **Theory acronyms**: NDT – Neural Darwinisn Theory, Orch OR – Orchestrated Objective Reduction, IIT – Information Integration Theory, RPT – Recurrent Processing Theory, GWT – Global Workspace Theory, HOTT – Higher Order Thought Theory, PPT – Predictive Processing Theory, ART – Adaptive Resonance Theory, TLT – Thalamocortical Loop Theory, NMDA – NMDA Theory, AST – Attention Schema Theory, SMT – Sensorimotor Theory, SCMT – Self Comes to Mind Theory.

	CLASS I: Identify consciousness with other phenomena		CLASS II: Identifiy consciousness with causal structures		Class III: Identify consciousness with computational processes				CLASS IVa: Identify consciousness with biological processes		CLASS IVb: Identify consciousness with cognitive processes		
	NDT	Orch OR	IIT	RPT	GWT	HOTT	PPT	ART	TLT	NMDA	AST	SMT	SCMT
CRITERIA													
III.1: Addresses paradigm cases of consciousness													
III.2: Copes with the unfolding argument													
III.3: Copes with the small network argument													
III.4: Copes with the multiple realization argument													
SCOPE													
II.1: Addresses state, content, both or unclear (S, C, B, Unc)	Unc	B	B	B	B	B	B	B	B	S	B	B	B
II.2: Consciousness is graded, binary or unclear (G, B, Unc)	Unc	G	G	B	B	B	Unc	Unc	Unc	G	B	Unc	Unc
II.3: Consciousness is unitary, non-unitary or unclear (U, N, Unc)	U	U	U	Unc	U	Unc	U	Unc	U	Unc	Unc	Unc	Unc
II.4: Consciousness is continuous, discrete or unclear (C, D, Unc)	Unc	D	Unc	Unc	D	Unc	Unc	Unc	D	Unc	Unc	Unc	Unc
II.5: Fate of unconscious elements is clear or unclear (C, Unc)	Unc	C	C	C	C	C	C	C	C	Unc	C	C	C

thus foster discussion about how best to deal with them. The various subsections can be read independently and, thus, hurried readers can pick their ToCs of interest and skip others.

IV.1. Theories challenged by criterion I: Paradigm cases of consciousness

IV.1.a. Linking consciousness to a cognitive process

Synopsis. Many cognitive processes have been linked to consciousness including learning (Cleeremans, 2007; Cleeremans et al., 2019), body ownership (Faivre et al., 2015), language (Gazzaniga, 1970), integrated and egocentric encoding of the world (Barron & Klein, 2016) and homoeostatic bodily responses (Critchley et al., 2004), to name a few. Some ToCs are not explicit as to whether the proposed cognitive process is necessary for consciousness, sufficient, or both. Other ToCs assert that the proposed cognitive process is necessary but not sufficient for consciousness (e.g. Cleeremans, 2007). Others *identify* consciousness with the proposed cognitive process, which is taken to be necessary and sufficient (Barron & Klein, 2016). Typically, the evidence for linking these cognitive processes to consciousness relies on the fact that they seem to always be associated with consciousness. Importantly, none of these proposals address paradigm cases. Hence for these types of ToCs, it is important to show that consciousness itself and not merely co-occurring processes are addressed. Even though there is a link to consciousness, it needs to be shown that the link is not merely contingent. There may even be double dissociations. For example, language seems neither necessary nor sufficient in order to consciously perceive a patch of colour.

Example: Subjectivity and the Hermann-Hering Grid: As an illustration, consider the illusory ghost spots in the Hermann-Hering Grid (Figure 4). They might be taken to reflect consciousness since they are purely subjective, i.e., only in the eye of the beholder. However, since there is no unconscious alternative, there is no reason to link the ghost spots to consciousness rather than other co-occurring processes. In fact, the illusory spots could be explained by simple facts of retinal processing (Figure 4, Baumgartner, 1978). Recent explanations of the ghost spots rely on basic cortical processing (Blakeslee & McCourt, 2012), but we present here the simpler retinal explanation to illustrate how subjective effects can in theory arise independently of consciousness. More generally, for consciousness, it does not matter if there are real spots in the outer world causing a percept (distal stimulus) or if the retina or basic cortical processing 'creates' these spots instead (proximal stimulus). To illustrate this with a simpler example: imagine a person holding up two fingers and, then, there is a change to five fingers. The content of consciousness has changed but clearly this experiment does not tell us anything about consciousness since it was not the dependent variable: the observer was always conscious. What has changed are states of the outer world and accordingly trivially the content of perception and consciousness. The Hermann-Hering grid is simply an instance of non-veridical perception discussed in the context of consciousness. When a percept does not correspond to the 'objective' reality, consciousness may *seem* to be involved, but this impression is usually mistaken. Illusions and altered states of consciousness (under the effect of psychotropic substances for instance) face similar issues: it does not matter for consciousness how an illusion or the contents of an altered state of consciousness are created by the brain. The question is why these contents – and not others – are consciously perceived? There certainly is a change in content, but consciousness *per se* cannot be shown to change.

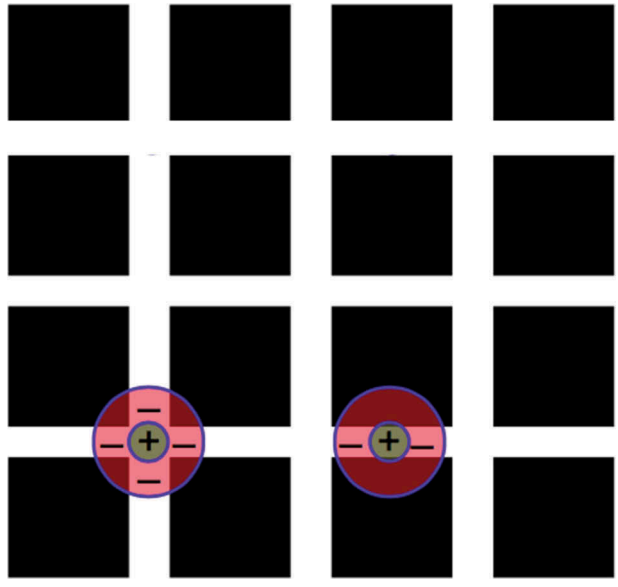

Figure 4. The Hermann-Hering Grid. Black spots are perceived at the intersections of the white strips. Because these ghost spots are subjective, it seems they are directly linked to consciousness. However, the occurrence of the spots may be explained by basic retinal or cortical processing. An on-centre neuron fires more strongly to a white line than to two intersecting white lines because there is less inhibition in the first case, explaining why the spots are perceived only for the latter case. Receptive fields of two on-centre neurons are shown here. The more light falls in the red periphery of these neurons' receptive fields, the less they fire. More recent explanations of the ghost spots rely on basic cortical processing (Blakeslee & McCourt, 2012), but we present here the simpler retinal explanation to illustrate how subjective effects can in theory arise independently of consciousness.

Example: Body Ownership and the rubber hand illusion. In the rubber hand illusion, participants can be tricked into attributing ownership of their real hand to a fake rubber hand (Botvinick & Cohen, 1998). They feel like the fake hand is their hand. There is no unconscious alternative, since the participant is having a conscious ownership experience the whole time. The experience simply changes from owning the real hand to owning the rubber one. The rubber hand illusion has been used to study consciousness (Faivre et al., 2015). However, since there is no unconscious alternative, the situation is similar to the Hermann-Hering grid and the fingers examples mentioned above. There certainly is a change in body representation, but consciousness per se is not shown to change.

In summary, the main challenge for these theories linking consciousness to a cognitive process is that, because they do not address paradigm cases, they may not address consciousness *per se* but instead other co-occurring processes. These ToCs need to make explicit why they are about consciousness *per se*.

IV.1.b. Linking consciousness to a neural or physical process

Synopsis. Many theories identify consciousness with neural or physical processes. For example, neural feature binding has been proposed as the key feature of consciousness, and this binding has been explained by neural synchrony or re-entrant processing (Edelman, 2003; Grossberg, 2017; Llinas et al., 1998; Singer, 2007). One of the most well-known theories linking consciousness to neural binding is Edelman's (2003) Neural Darwinism Theory (NDT). As a quantum physical example, Penrose and Hameroff's Orchestrated Objective Reduction (Orch OR; Hameroff & Penrose, 2014) theory links consciousness with understanding, and then links understanding to quantum mechanical processes. Orch OR faces similar challenges as NDT (see appendix).

Example: Neural Darwinism Theory (NDT). NDT proposes that the crucial property of consciousness is to bind neural information into a unitary percept. This claim is based on the observation that conscious states are highly diverse, yet always unified. Binding is proposed as the mechanism to create these diverse and unified conscious states and is explained in terms of assemblies of neurons firing synchronously in the thalamocortical system. Which neurons fire together is determined by a selectionist mechanism, called Neural Darwinism: circuits that give rise to useful percepts are selected and strengthened through epigenetic and synaptic plasticity. These synchronous neural assemblies are proposed to be in 1–1 correspondence with conscious states.

Similarly to the approaches linking consciousness to a cognitive process, NDT does not have an unconscious alternative. Indeed, binding occurs in both conscious and unconscious processing. For example, unconscious priming can occur at the object level (James et al., 2000), suggesting that parts can be bound into wholes unconsciously. Therefore, binding cannot be used to study differences between conscious and unconscious alternatives in paradigm cases.

In summary, since there is no unconscious alternative, an important challenge for NDT is to show that it really targets consciousness *per se* rather than perception in general (III.1).

IV.2. Theories challenged by criterion II: The unfolding argument

IV.2.a. Causal structure theories

Synopsis. Theories that address paradigm cases aim to predict changes from the unconscious to the conscious alternative in experimental paradigms. In order to do so, several ToCs focus on *how* elements of a system interact. If the system has the 'right' kind of causal structure, i.e., if its elements interact in the 'right' way, it is conscious. Otherwise, it is not. The two most well-known examples are Recurrent Processing Theory (RPT; Lamme, 2006) and Integrated Information Theory (IIT; Tononi, 2004). Causal structure theories need to address the unfolding argument and often the small network argument. In the following, we present IIT as an example. RPT faces similar challenges (see appendix).

Example: Integrated Information Theory (IIT). IIT was first proposed by Tononi (2004). The theory starts with a set of five axioms proposed to capture the phenomenological properties of consciousness. The unity of consciousness (II.3.b) is an axiom, for example. These axioms are translated into mathematical postulates, from which it is derived that consciousness corresponds to integrated information. The amount of information integrated by a system, and therefore its level of consciousness, is quantified by a number: Φ. Importantly, Φ is computed based on the causal structure of the system. In particular, feedforward networks always have $\Phi = 0$ and recurrent networks always have $\Phi > 0$. When several subsystems have $\Phi > 0$, the unitary consciousness of the system as a whole is determined by the subsystem with the maximal Φ.

IIT addresses both state and content paradigm cases of consciousness. Conscious states occur if and only if Φ > 0 (so unconscious states occur when Φ = 0). For conscious content, a feature (e.g., the target in a masking experiment) is consciously perceived if its representation contributes to the network that determines Φ > 0. The unconscious alternative for conscious contents occurs when the feature of interest is not represented in the network that determines Φ > 0. IIT is perhaps the current theory that copes best with the other systems argument, making precise quantitative predictions about which systems are conscious, how much, and even which conscious contents are perceived.

Like all other causal structure theories, IIT is subject to the unfolding argument (III.2) because systems with identical input-output functions can have Φ = 0 or arbitrarily high Φ. The unfolding argument also shows that systems with identical input-output functions can have arbitrary conscious content, according to IIT (Doerig, Schurger, et al., 2019). For example, a system participating in a rivalry experiment may report that it is seeing the cat image from Figure 1 when, according to IIT, it is experiencing the smell of ham. In principle, it can experience *any* content of consciousness while reporting that it sees a cat.

IIT also faces the small network argument (III.3). Proponents of the theory accept that two recurrently connected neurons are indeed conscious and accept panpsychism (Oizumi et al., 2014; Tononi & Koch, 2015). The criterion determining which systems are conscious is Φ > 0. However, this criterion is at least partly arbitrary, since the axioms do not uniquely specify the complexity measure, i.e., other measures than Φ fulfil the axioms as well (Bayne, 2018), but lead to very different empirical results (Mediano et al., 2019). Moreover, IIT also faces the 'large network argument' (III.3.d). For example, gravitational interactions between celestial objects in a galaxy integrate information and therefore have Φ > 0. Consciousness is attributed only to the subsystem with the highest Φ locally. Since a system always has larger or equal Φ than any of its subsystems (either the subsystem is the local maximum of Φ or it is lower than that), depending on how we interpret 'locally', only the Universe has consciousness because it has the highest Φ. IIT needs a criterion for differentiating subsystems.

In summary, the main challenge for IIT is that the unfolding argument suggests that the theory cannot address empirical data about the state nor about the content of consciousness. IIT needs to show how it can deal with the unfolding argument. In this respect, see Tsuchiya et al. (2020) and Kleiner (2019) who defend IIT from the unfolding argument. Moreover, IIT faces the small and large network arguments. It embraces panpsychism and proposes a criterion for determining which small networks are conscious, but this criterion is at least partly arbitrary. IIT also needs a more precise criterion for large systems.

IV.3. Theories challenged by criterion III: The small/large network argument

IV.3.a. Computational theories

Synopsis. Many important ToCs are of a computational nature. These theories have several advantages: they address paradigm cases, fit comfortably within the modern information processing framework of neuroscience and often deal convincingly with the unfolding and the other systems arguments. However, they need to deal with the small network argument. We present two of the leading ToCs, Global Workspace Theory and Higher Order Thought Theory as examples. Two other ToCs, Predictive Processing Theory and Adaptive Resonance Theory are presented in the appendix.

Example: Global Workspace Theory (GWT). GWT was first put forward by Bernard Baars (Baars, 1993; Baars et al., 2013). Dehaene and collegues subsequently proposed the global *neuronal* workspace theory (Dehaene & Naccache, 2001; Mashour et al., 2020). GWT postulates that conscious experiences reflect a flexible binding and broadcasting function in the brain. That is, peripheral coalitions of neurons compete in a winner-take-all fashion and the winner broadcasts information to the whole brain, thus binding features of different modalities into a coherent conscious percept. 'To be consciously accessible, information must be encoded as an organized pattern of neuronal activity in higher cortical regions, and this pattern must, in turn, ignite an inner circle of tightly interconnected areas forming a global workspace' (Dehaene, 2014). For example, different interpretations of a visual scene may compete until a winner is globally 'broadcast', giving rise to a unified conscious experience. In its simplest version, the global workspace model comprises peripheral neurons that project to central neurons, which in turn bind information together and broadcast this information to the entire network.

GWT addresses paradigm cases including the attentional blink (Sergent et al., 2005), masking (Dehaene et al., 2001), and wakefulness vs. sleep (Dehaene et al., 2003). The broadcasting mechanism has conscious and unconscious alternatives. For example, in masking, it was found that activity spreads more in the cortex when the target is consciously perceived than when it is not. The explanation is that the target becomes conscious when it enters the workspace

and remains unconscious when the mask prevents the target from entering the workspace. GWT is not subject to the unfolding argument because broadcasting is a general concept, which can be implemented in many different ways (even though GWT is usually framed in recurrent terms, recurrence is not essential, Doerig, Schurger, et al., 2019; but see Kleiner, 2019; Kleiner & Hoel, 2020). GWT copes with the other systems argument by proposing that systems endowed with a global workspace architecture are conscious.

The small network argument applies because the global workspace architecture and broadcasting can be realized with very few neurons. For example, a network consisting of two peripheral neurons connected to a small recurrent global workspace fulfils the criteria for consciousness proposed by GWT. Proponents of GWT do not usually concede and grant consciousness to these small networks, so additional criteria are needed to explain why they are not conscious. As explained in section III.3.b, simple addendums cannot fix the problem, so something crucial seems to be missing from the theory.

Moreover, billions of subsystems in the brain have a GW architecture so GWT faces the large network argument (III.3.d). Another additional criterion is needed to decide which one of them gives rise to our seemingly unitary conscious experience, or GWT needs to give up the unity of consciousness. Finally, the need for additional criteria is also highlighted by the existence of systems such as the immune and the vegetative nervous systems, which can also be seen as having a GW structure with broadcasting but are not granted consciousness by GWT.

In summary, the main challenge for GWT seems to be that it grants consciousness to too many systems. Therefore, GWT needs to provide criteria to cope with the small (and large) network argument.

Example: Higher Order Thought Theory (HOTT). HOTT is another popular theory subject to the same kind of challenges as GWT. While HOTT originated in philosophical circles (Rosenthal, 1986, 2002) and has mostly been couched in cognitive terms, it has more recently made its way into the arena of neuroscientific theories of consciousness (Lau & Rosenthal, 2011). There are a few variants of higher-order theories of consciousness (Brown et al., 2019; Lau, 2019; Lau & Rosenthal, 2011; Rosenthal, 2004) but the gist of HOTT is this: A mental state *X* is conscious if, and only if, one has a higher-order representation to the effect that one is currently representing *X*. One prominent neuroscientific take on HOTT posits that specific areas of pre-frontal cortex involved in metacognition are directly involved in the formation of higher-order thoughts (Lau & Rosenthal, 2011). HOTT can be contrasted with first-order theories such as IIT, which posit that consciousness of *X* depends only on the first-order neural representation of *X*.

HOTT addresses paradigm cases of consciousness such as visual masking (Lau & Rosenthal, 2011). There are conscious and unconscious alternatives, depending on whether or not the relevant higher-order thoughts occur. In the example of masking, there is usually a higher-order thought to the effect that one perceives the mask (subjects usually report being conscious of the mask in masking experiments), but no higher-order thought that one perceives the stimulus that is masked.

HOTT faces different challenges depending on how it is interpreted, because different authors have different takes on what exactly constitutes a higher order thought (compare for instance Lau (2019) and Rosenthal (2002)). Depending on the interpretation, either the small network or the other systems arguments may apply. At one extreme, if we interpret higher order thoughts as a simple 2-stage computation, HOTT is subject to the small network argument because any 2-stage computer program is conscious. In this case, HOTT needs to propose criteria to distinguish which systems are conscious and to explain which subsystems of the brain contribute to our seemingly unitary conscious experience. At the other extreme, higher order thoughts may simply refer to the everyday concept of human thinking. In this case, HOTT is not a computational theory and the other systems argument applies: HOTT needs to explain what is crucial to be a 'thought' and which systems can have equivalents to 'thoughts' and be conscious, apart from humans. These extreme interpretations of HOTT illustrate why intermediate interpretations may face a mix of the small network and other systems arguments.

Since there is currently no clear-cut mechanism, it is difficult to say to what extent these arguments apply but, eventually, HOTT will need to include a specific mechanism that can cope with both the other systems and the small network argument.

IV.4. Theories challenged by criterion IV: The other systems argument

IV.4.a. Biological theories

Synopsis. Many studies have investigated the neural correlates of consciousness (Koch et al., 2016; Rees et al., 2002). For example, brain areas correlating with the conscious percept in binocular rivalry were found (Tong et al.,

1998). However, in general, correlates cannot tell about the underlying causes and mechanisms. Quite a few researchers went further and identified certain biological properties of the brain as essential for consciousness. Examples include NMDA synapses (Flohr, 1992), processing in layer 5 of V1 (Crick & Koch, 2003), the claustrum (Crick & Koch, 2005), gamma-band oscillations (Buzsáki, 2004), thalamocortical loops (Llinas et al., 1998), or the microgenesis of consciousness in the thalamus (Bachmann, 2000). In general, these theories need to address the other systems argument because they focus only on consciousness in humans and cannot tell which other systems may be conscious. We portray Llinas' Thalamocortical Loops Theory (TLT) as an example.

Example: Thalamocortical Loops Theory (TLT). Llinas et al. (1998) link consciousness to neural binding. The proposed mechanism involves gamma-band oscillations in thalamocortical loops, which bind information from different modalities by synchronizing the firing of neurons. Different cortical regions represent different contents, which are bound into a unitary conscious percept by gamma band oscillations. The intralaminar nucleus in the thalamus generates the oscillation. Accordingly, it has been observed that damage to the intralaminar nucleus leads to coma. Consciousness is temporally discrete because the gamma oscillations create conscious snapshots (Joliot et al., 1994).

TLT links consciousness to neural binding, addresses paradigm cases and proposes an unconscious alternative. For example, the difference between wakefulness and dreamless sleep (a paradigm case) is explained by a difference in the amplitude of the thalamocortical oscillations. High amplitude corresponds to the conscious and low amplitude to the unconscious alternative.

The other systems argument applies: what is really necessary for consciousness in the thalamocortical system? Even if it is true that removing the thalamocortical system from adult humans renders them unconscious, a sentient being without a thalamocortical system may still possess consciousness because his brain may implement the crucial functions fulfilled by the thalamocortical system in a different way. Moreover, the unfolding argument opens up the possibility that other implementations of the thalamocortical system are possible.

In summary, TLT faces the other systems argument and needs to explain what is special about the thalamocortical system.

IV.4.b. Cognitive theories

Synopsis. Certain theories propose a cognitive mechanism for consciousness. These theories are usually not subject to the unfolding argument and small network argument because their mechanisms are very complex, involving, for example, language or attention. The price paid is that the mechanisms are usually vaguely described, provoking the other systems argument. These theories are different from theories in section IV.1, because, rather than merely *linking* consciousness with a cognitive process, they provide an *explanation* of how the proposed cognitive mechanism differentiates between conscious and unconscious alternatives. Here, we portray the Attention Schema Theory (AST) as an example. O'Regan and Noë's (2001) Action Percetion Loop Theory and Damasio's (2010) Self Comes to Mind Theory are presented in the appendix.

Example: Attention Schema Theory (AST). According to AST, the brain is equipped to model the attention of others via cues, such as gaze direction and body language. This modelling mechanism is used to infer the mental state of others and to predict their behavior. Here 'attention' is defined in general terms as the highlighting of a subset of all currently available neural information for more in-depth processing. The focus of attention can, in any given instance, be directed exogenously at specific sensory information or endogenously at a memory or fantasy that one is currently entertaining. Not having direct access to the physical processes that underlie the directing of attention in the brains of others, the brain represents the attention controller as something ethereal that we come to know as 'consciousness' residing in others. When this same model is turned inwards and used to describe our own attention-directing mechanism, we come to attribute consciousness to ourselves. With notable exceptions, every bit of neural information to which we direct our attention becomes tagged with a special attribute, an ineffable 'something extra', that we have come to know as qualia. Thus, just as the brain maintains a body schema, which continuously models the relative position of one's body parts in space, the brain also maintains an 'attention schema' which continuously models one's current state of attention. According to the theory, consciousness and qualia are perceptual attributions, not fundamentally different from other perceptual attributions such as color or texture. According to Graziano, the temporo-parietal junction is a central hub in the implementation of the attention schema in the human brain.

AST is special because it is the only explicitly illusionist theory reviewed here. Illusionists usually argue against metaphysical ideas about the nature of consciousness, mainly against the existence of qualia as distinct non-physical entities (see Section I). If AST or other illusionist

ToCs argue only at this metaphysical level, our criteria do not apply. However, if AST seeks empirical support from data about consciousness (i.e., if it purports to explain masking, sleep, etc.), our criteria apply.

The proposed mechanism, i.e., the modeling of attention, can address paradigm cases of consciousness in principle: There are both conscious and unconscious alternatives, depending on whether information is modeled by the attention schema. Content paradigm cases are addressed by looking at which contents are the subject of attention modelled by the attention schema. Different states of consciousness could be explained by different states of the modeling process. Since AST is not a causal structure theory, it does not face the unfolding argument. AST requires that the system generates a model of what attention *is* (i.e., it must describe attention as something ethereal that we come to know as 'consciousness') not just what attention is pointed at. This plausibly requires a reasonable level of complexity, so the small network argument seems not to apply.

AST currently does not provide a clear-cut description of its mechanism, which leads to the other systems argument. The main question is: what defines attention modelling? If there is no computational explanation of what really counts as attention modeling, the other systems problem arises: what is crucial for consciousness in the human implementation of attention modelling? Which other systems have it and are thus predicted to be conscious? One issue for AST in coping with the other system argument is that, if attention modelling is defined computationally (for example as maps representing where enhanced information processing occurs), the small network argument may pose a problem. Eventually, AST will need a specific mechanism that can satisfy the other-systems criterion without conjuring up the small network argument.

V. Discussion

Modern neuroscience has developed many tools and paradigms to investigate consciousness empirically (masking, rivalry, crowding, continuous flash suppression, anaesthesia, …), and many more may come. Through these phenomena, consciousness can, at least partly, be addressed as any other scientific topic, leaving metaphysical questions aside. In parallel, a puzzling plethora of ToCs were proposed. We suggest that this exuberance of different theories points to a lack of stringent criteria in consciousness science.[3] Consequently, the major goal of this contribution is to put forward a list of criteria that empirical ToCs need to address.

As mentioned, we do not claim to have fully described the many facets of the theories we reviewed. In addition, we do not propose that ToCs challenged by our criteria should be discarded. To the contrary, our analysis is meant to explicitly formulate these challenges and, in this way, to provide a common framework to discuss and compare ToCs systematically. We do not propose that the criteria are exhaustive and unique. To the contrary, we see this contribution as a steppingstone to develop agreed-upon criteria that foster theory evaluation and selection. Further criteria may be added and proposed criteria may be modified or dropped. In the following we classify ToCs in four classes, depending on how they cope with our four major criteria.

V.1 Class I: ToCs that do not address paradigm cases of consciousness

A first, large class of ToCs addresses consciousness only indirectly via phenomena proposed to be identical or closely associated with consciousness (see NDT and Orch OR, Table 1). As we have shown, most of these ToCs first motivate why the proposed phenomenon is crucial for or identical with consciousness and then work out the details of the phenomenon rather than the details of consciousness. These ToCs usually do not address paradigm cases of consciousness. For this reason, it is not always clear whether the proposed phenomenon is in fact linked with consciousness at all. Sometimes the link can be described as a mystery meeting another mystery. To quote Pinker (2007), '[Q]uantum mechanics sure is weird, and consciousness sure is weird, so maybe quantum mechanics can explain consciousness'. Class I ToCs need to specify whether the proposed phenomenon is necessary, sufficient, or only linked to consciousness, and clarify why and how. It is important to show that consciousness is really addressed, and not only a co-occurring process.

V.2. Class II: ToCs that are subject to the unfolding argument

ToCs of class II identify consciousness with causal structures (see IIT and RPT, Table 1). In some way, these ToCs locate consciousness on the level of hardware rather than software. These theories address paradigm cases and certain characteristics of consciousness, such as unity.

[3]On top of a potential lack of criteria, Lau and Michel (2019) propose a socio-historical take on this question.

They provide clear cut mechanisms proposed to be necessary and sufficient, and make quantitative predictions about the empirical characteristics of consciousness listed in section II. However, these ToCs are challenged by the unfolding argument because they imply that empirically identical systems have different consciousness. Any function can be implemented by many different systems with different causal structures. Hence, it seems that there can be no consistent link between causal structures and experimental results. Proposing simple, clear cut causal structures as *sufficient* for consciousness seems to open the door too wide. This is also the reason why the small and large network arguments apply. Hence, class II ToCs need to address and clarify how they can cope with the unfolding and small network arguments or why they refute these criteria.

V.3. Class III: ToCs that are subject to the small network argument

Class III ToCs propose that computational aspects are identical with consciousness (see GWT, HOTT, PPT and ART, Table 1). These ToCs address paradigm cases, address the empirical characteristics of consciousness listed in section II and usually successfully deal with the unfolding and other systems arguments because they are independent of the specific implementation and apply to any type of creature. However, they seem to fall prey to the small (and large) network argument because they are too unconstrained, and hence apply to too many systems. Hence, the proposed mechanisms do not seem sufficient and therefore additional criteria are required (section III.3.b). The extent to which the proposed mechanism is crucial for consciousness needs to be demonstrated.

V.4. Class IV: ToCs that are subject to the other systems argument

ToCs of Class IVa identify consciousness with biological processes. Some ToCs propose that certain biological systems or processes are crucial for consciousness (see TLT and NMDA, Table 1). ToCs of Class IVb identify consciousness with cognitive processes (see AST, SMT and SCMT, Table 1).

Class IV ToCs address paradigm cases by specifying how the candidate process plays a crucial role in the transition from unconsciousness to consciousness. Most of these ToCs avoid the unfolding and the small (and large) network arguments because they attribute consciousness mainly to humans instead of proposing a clear-cut mechanism applicable to any system. However, for this reason, the other systems argument kicks in, and it is difficult for these ToCs to address the empirical characteristics of consciousness presented in section II. One way out of this problem is to provide a precise computational formulation of the proposed biological or computational process. However, in this case, the ToC becomes a class III theory and the small network argument may apply.

VI. Conclusions

The many ToCs in classes II–IV directly address consciousness through paradigm cases. In general, it is important for each ToC to unearth the common characteristics of paradigm cases (Chalmers, 1996; Fingelkurts et al., 2012; Haynes, 2009; Seth, 2016). Whether we have sufficient data, experimental paradigms, etc., at the moment remains an open question. Maybe the plethora of ToCs simply reflects the fact that we have too few experimental constraints. It is possible that with more data and a more detailed view of the subprocesses of consciousness, the mystery will evaporate, similarly to what happened with the discussion about the 'nature' of life. Nowadays biologists understand what life is, but there is no 'theory of life' (Machery, 2012). It is the entirety of subprocesses such as homeostasis, reproduction, etc., that differentiates life from non-life.

Current ToCs do not take this approach. Instead, a characteristic of all ToCs is that they seemingly identify consciousness with something else. To illustrate this with a metaphor: Banksy is a street artist whose private identity is unknown. It may be your neighbour. You know there is a street artist and you know your neighbour, but you do not know that they are identical (Kripke, 1972). Similarly, all ToCs suggest that consciousness is identical to something we know already and propose to elucidate the link between the two. Consciousness is not something 'new'. ToCs differ in *what* they identify with consciousness. For example, GWT identifies broadcasting with consciousness, IIT identifies systems with $\Phi > 0$ with consciousness, and AST identifies the modelling of attention with consciousness.[4] One difficulty is that identifying consciousness with clear cut mechanisms such as recurrence or a mathematical definition of integrated information easily leads to the small network or the unfolding argument. Relatedly, a theory explaining rivalry, masking and sleep is a theory of the three but not necessarily a theory of consciousness. As soon as a concrete model is proposed, the small network argument becomes

[4]Even though AST is an illusionist theory, it still identifies the illusion of consciousness with modelling attention.

threatening and the question arises whether a small system implementing this model is conscious – which is nothing other than the hard problem of consciousness. Vague mechanisms camouflage this issue. Perhaps for this reason, ToCs often identify consciousness with a rather vague, metaphorical or little understood aspect, such as models of attention, complexity, neural binding, or quantum states. In this line, it is not surprising that few ToCs make detailed predictions and are therefore difficult to compare. In short, identifying consciousness with something precise leads to a slippery slope with the small network and unfolding arguments at the bottom, and identifying consciousness with a vaguer property makes it difficult to make detailed predictions and to cope with the other systems argument. Although it may turn out that these hurdles can be overcome so that identifying consciousness with a known phenomenon will ultimately succeed, we propose that these challenges are serious and must be confronted.

Another option is that consciousness is something 'new'. The current situation in consciousness research may be similar to that of magnetism in ancient times. The ancient people of Greece, India and China knew the empirical phenomena of magnetism. For example, Thales knew that certain stones could move certain other objects and attributed this power to souls residing in the magnetic stones. For two millennia, there was no widely accepted theoretical explanation or definition, and the discussion might have resembled what we encounter nowadays in consciousness research: either magnetism was deemed fundamentally mysterious or it was identified with known entities (e.g., linking magnets and souls). Likewise, consciousness is most often explained by entities and theories of today, such as neural, cognitive, or computational processes. However, just like consciousness, magnetism was well 'defined' empirically, e.g., by the attraction or repulsion between magnetic stones, compasses, etc. After centuries of research, and despite the lack of rigid definitions to start with, magnetism has lost its mysteries through the Maxwell equations and subsequent theories of electro-magnetism, which provided a clear scientific explanation (see Dennett, 1991, p. 44). To explain magnetism, it was necessary to understand other phenomena, such as electricity, beforehand. Maybe consciousness is a 'solution', a by-product, or a core component of a computational challenge that information processing systems need to solve – and that we have not discovered yet.

Whatever the final answer to these questions is, theoretical frameworks clarifying the link between empirical data and ToCs are crucial in order to compare theories and make progress in consciousness science. The criteria we propose are intended as a first set of guidelines to foster discussions about consciousness as an empirical phenomenon.

Acknowledgments

This work was funded by the SNF grant "Basics of visual processing: from elements to figures" (176153). We thank Michael Graziano, Stephen Grossberg, Victor Lamme, Hakwan Lau, Kevin O'Regan, David Rudrauf, David Rosenthal, and Nao Tsuchiya for helpful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the SNF grant 'Basics of visual processing: from elements to figures' (176153).

ORCID

Adrien Doerig http://orcid.org/0000-0001-5120-9750
Aaron Schurger http://orcid.org/0000-0003-2985-3253
Michael H. Herzog http://orcid.org/0000-0001-5433-1030

References

Aaronson, S. (2014). *Giulio Tononi and me: A phi-nal exchange*. http://www.scottaaronson.com/blog/?p=1823

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience and Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Atas, A., Faivre, N., Timmermans, B., Cleeremans, A., & Kouider, S. (2014). Nonconscious learning from crowded sequences. *Psychological Science*, *25*(1), 113–119. https://doi.org/10.1177/0956797613499591

Baars, B. J. (1986). What is a theory of consciousness a theory of?—The search for criterial constraints on theory. *Imagination, Cognition and Personality*, *6*(1), 3–23. https://doi.org/10.2190/WJER-XABV-QM4W-KD6V

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. J., Franklin, S., & Ramsoy, T. Z. (2013). Global workspace dynamics: Cortical "binding and propagation" enables conscious contents. *Frontiers in Psychology*, 4, 200. https://doi.org/10.3389/fpsyg.2013.00200

Bachmann, T. (2000). *Microgenetic approach to the conscious mind* (Vol. 25). John Benjamins Publishing.

Ball, P. (2019). *Neuroscience readies for a showdown over consciousness ideas*. Quanta Magazine. https://www.quantamagazine.org/neuroscience-readies-for-a-showdown-over-consciousness-ideas-20190306/

Balsdon, T., & Clifford, C. W. (2018). Visual processing: Conscious until proven otherwise. *Royal Society Open Science*, *5*(1), 171783. https://doi.org/10.1098/rsos.171783

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, *113*(18), 4900–4908. https://doi.org/10.1073/pnas.1520084113

Baumgartner, G. (1978). Physiologie des zentralen Sehsystems. Gauer OH, Kramer K, Jung R (Hrsg) Sehen: Sinnesphysiologie DI. Urban & Schwarzenberg, München Wien Baltimore, 263-348.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1), niy007. https://doi.org/10.1093/nc/niy007

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, *66* (2), 175–207. https://doi.org/10.1086/392683

Blakeslee, B., & McCourt, M. E. (2012). When is spatial filtering enough? Investigation of brightness and lightness perception in stimuli containing a visible illumination component. *Vision Research*, *60*(May 2012), 40–50. https://doi.org/10.1016/j.visres.2012.03.006

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel'touch that eyes see. *Nature*, *391*(6669), 756. https://doi.org/10.1038/35784

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, *13*(4), 767–782. https://doi.org/10.1016/0042-6989(73)90041-2

Breitmeyer, B., & Öğmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. Oxford University Press.

Breitmeyer, B. G. (2015). Psychophysical "blinding" methods reveal a functional hierarchy of unconscious visual processing. *Consciousness and Cognition*, *35*(September 2015), 234–250. https://doi.org/10.1016/j.concog.2015.01.012

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Burdick, R. K., Villabona-Monsalve, J. P., Mashour, G. A., & Goodson, T. (2019). Modern anesthetic ethers demonstrate quantum interactions with entangled photons. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-47651-1

Buzsáki, G. (2004). Neuronal oscillations in cortical networks. *Science*, *304*(5679), 1926–1929. https://doi.org/10.1126/science.1099745

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

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chalmers, D. J. (2004). *How can we construct a science of consciousness?*. MIT Press.

Chalmers, D. J. (2017). The combination problem for panpsychism. *Panpsychism: Contemporary Perspectives*, 179–215.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cleeremans, A. (2007). Consciousness: The radical plasticity thesis. *Progress in Brain Research*, *168*(2007), 19–33. https://doi.org/10.1016/S0079-6123(07)68003-0

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2019). Learning to be conscious. *Trends in Cognitive Sciences*, *24*(2), 112–123. https://doi.org/10.1016/j.tics.2019.11.011

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364. https://doi.org/10.1016/j.tics.2011.06.008

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*(2), 119–126. https://doi.org/10.1038/nn0203-119

Crick, F. C., & Koch, C. (2005). What is the function of the claustrum? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1458), 1271–1279. https://doi.org/10.1098/rstb.2005.1661

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189. https://doi.org/10.1038/nn1176

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. Vintage.

Dehaene, S. (2014). *Consciousness and the brain: deciphering how the brain codes our thoughts*. Penguin.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758. https://doi.org/10.1038/89551

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, *100*(14), 8520–8525. https://doi.org/10.1073/pnas.1332574100

Dennett, D. C. (1991). *Consciousness explained*. Little Brown & Co.

Dennett, D. C. (2016). Illusionism as the obvious default theory of consciousness. *Journal of Consciousness Studies*, *23(11-12)*, 65–72.

Descartes, R. (1996). *Discourse on the method: And, meditations on first philosophy*. Yale University Press.

Doerig, A., Scharnowski, F., & Herzog, M. H. (2019). Building perception block by block: A response to Fekete et al. *Neuroscience of Consciousness*, *2019*(1), niy012. https://doi.org/10.1093/nc/niy012

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and*

Cognition, *72*(July 2019), 49–59. https://doi.org/10.1016/j.concog.2019.04.002

Drissi-Daoudi, L., Doerig, A., & Herzog, M. H. (2019). Feature integration within discrete time windows. *Nature Communications*, *10*(1), 4901. https://doi.org/10.1038/s41467-019-12919-7

Edelman, G. M. (2003). Naturalizing consciousness: A theoretical framework. *Proceedings of the National Academy of Sciences*, *100*(9), 5520–5524. https://doi.org/10.1073/pnas.0931349100

Faivre, N., Salomon, R., & Blanke, O. (2015). Visual consciousness and bodily self-consciousness. *Current Opinion in Neurology*, *28*(1), 23–28. https://doi.org/10.1097/WCO.0000000000000160

Fekete, T., Van de Cruys, S., Ekroll, V., & van Leeuwen, C. (2018). In the interest of saving time: A critique of discrete perception. *Neuroscience of Consciousness*, *2018*(1), niy003. https://doi.org/10.1093/nc/niy003

Feuillet, L., Dufour, H., & Pelletier, J. (2007). Brain of a white-collar worker. *Lancet (London, England)*, *370*(9583), 262. https://doi.org/10.1016/S0140-6736(07)61127-1

Fingelkurts, A. A., Fingelkurts, A. A., & Neves, C. F. (2012). "Machine" consciousness and "artificial" thought: An operational architectonics model guided approach. *Brain Research*, *1428*(January 2012), 80–92. https://doi.org/10.1016/j.brainres.2010.11.079

Flohr, H. (1992). Qualia and brain processes. In A. Beckermann, H. Flohr, and J. Kim, eds., *Emergence or Reduction*, 220–238.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115. https://doi.org/10.1007/BF00485230

Frankish, K. (2016b, September 20). *Why panpsychism is probably wrong*. The Atlantic. https://www.theatlantic.com/science/archive/2016/09/panpsychism-is-wrong/500774/

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127. https://doi.org/10.1038/nrn2787

Friston, K. (2013). Consciousness and hierarchical inference. *Neuropsychoanalysis*, *15*(1), 38–42. https://doi.org/10.1080/15294145.2013.10773716

Gaillard, R., Del Cul, A., Naccache, L., Vinckier, F., Cohen, L., & Dehaene, S. (2006). Nonconscious semantic processing of emotional words modulates conscious access. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(19), 7524–7529. https://doi.org/10.1073/pnas.0600584103

Gazzaniga, M. S. (1970). *The bisected brain* (Vol. 2). Appleton-Century-Crofts New York.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press.

Grossberg, S. (2017). Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Networks*, *87*(March 2017), 38–95. https://doi.org/10.1016/j.neunet.2016.11.003

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR'theory. *Physics of Life Reviews*, *11*(1), 39–78. https://doi.org/10.1016/j.plrev.2013.08.002

Hanson, J. R., & Walker, S. I. (2019). Integrated information theory and isomorphic feed-forward philosophical zombies. *Entropy*, *21*(11), 1073. https://doi.org/10.3390/e21111073

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, *13*(5), 194–202. https://doi.org/10.1016/j.tics.2009.02.004

Herzog, M. H., Esfeld, M., & Gerstner, W. (2007). Consciousness & the small network argument. *Neural Networks*, *20*(9), 1054–1056. https://doi.org/10.1016/j.neunet.2007.09.001

Herzog, M. H., Kammer, T., & Scharnowski, F. (2016). Time slices: What is the duration of a percept? *PLoS Biology*, *14*(4), e1002433. https://doi.org/10.1371/journal.pbio.1002433

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

James, T. W., Humphrey, G. K., Gati, J. S., Menon, R. S., & Goodale, M. A. (2000). The effects of visual object priming on brain activation before and after recognition. *Current Biology*, *10*(17), 1017–1024. https://doi.org/10.1016/S0960-9822(00)00655-2

James, W. (2013). *The principles of psychology*. Read Books Ltd.

Joliot, M., Ribary, U., & Llinas, R. (1994). Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences*, *91*(24), 11748–11751. https://doi.org/10.1073/pnas.91.24.11748

Kiesel, A., Kunde, W., Pohl, C., Berner, M. P., & Hoffmann, J. (2009). Playing chess unconsciously. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(1), 292. https://doi.org/10.1037/a0014499

Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible.'. *Trends in Cognitive Sciences*, *9*(8), 381–388. https://doi.org/10.1016/j.tics.2005.06.012

Kleiner, J. (2019). Brain states matter. A reply to the unfolding argument. *PsyArXiv*. https://doi.org/10.31234/osf.io/jdcfh

Kleiner, J., & Hoel, E. (2020). Falsification and consciousness. *ArXiv:2004.03541 [q-Bio]*. http://arxiv.org/abs/2004.03541

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Kripke, S. A. (1972). Naming and necessity. In: Davidson D., Harman G. (eds) *Semantics of Natural Language. Synthese Library (Monographs on Epistemology, Logic, Methodology, Philosophy of Science, Sociology of Science and of Knowledge, and on the Mathematical Methods of Social and Behavioral Sciences)*, vol 40. Dordrecht: Springer.

Lamme, V. (2015). The Crack of Dawn: Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision. In T. K. Metzinger & J. M. Windt (Eds.), *Open MIND*: 22(T). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570092

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, *8*(4), 529–535. https://doi.org/10.1016/S0959-4388(98)80042-1

Lau, H. (2008). Are we studying consciousness yet. In L. Weiskrantz and M. Davies, eds., *Frontiers of Consciousness:*

Chichele Lectures, (pp. 245–258). Oxford: Oxford University Press.

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv*. https://doi.org/10.31234/osf.io/ckbyf

Lau, H., & Michel, M. (2019). A socio-historical take on the meta-problem of consciousness. *Journal of Consciousness Studies*, *26*(9–10), 136–147. doi: 10.31234/osf.io/ut8zq

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Llinas, R., Ribary, U., Contreras, D., & Pedroarena, C. (1998). The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*(1377), 1841–1849. https://doi.org/10.1098/rstb.1998.0336

Lutz, A., Dunne, J. D., & Davidson, R. J. (2007). Meditation and the neuroscience of consciousness. *Cambridge Handbook of Consciousness*, 499–555.

Machery, E. (2012). Why I stopped worrying about the definition of life … And why you should as well. *Synthese*, *185*(1), 145–164. https://doi.org/10.1007/s11229-011-9880-1

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mediano, P. A., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy, 21*(1), 17. https://doi.org/10.3390/e21010017

Morales, J., Chiang, J., & Lau, H. (2015). Controlling for performance capacity confounds in neuroimaging studies of conscious awareness. *Neuroscience of Consciousness*, *2015*(1). https://doi.org/10.1093/nc/niv008

Moutoussis, K., & Zeki, S. (1997). A direct demonstration of perceptual asynchrony in vision. *Proceedings of the Royal Society of London B: Biological Sciences*, *264*(1380), 393–399. https://doi.org/10.1098/rspb.1997.0056

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170357. https://doi.org/10.1098/rstb.2017.0357

Nishida, S., Watanabe, J., Kuriki, I., & Tokimoto, T. (2007). Human visual system integrates color signals along a motion trajectory. *Current Biology*, *17*(4), 366–372. https://doi.org/10.1016/j.cub.2006.12.041

Noë, A. (2004). *Action in perception*. MIT press.

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Otto, T. U., Ögmen, H., & Herzog, M. H. (2006). The flight path of the phoenix—The visible trace of invisible elements in human vision. *Journal of Vision*, *6*(10), 7. https://doi.org/10.1167/6.10.7

Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and Cognition*, *19*(2), 682–684. https://doi.org/10.1016/j.concog.2009.12.018

Park, H.-D., & Tallon-Baudry, C. (2014). The neural subjective frame: From bodily signals to perceptual consciousness. *Soc.*, *369*(1641), 20130208. https://doi.org/10.1098/rstb.2013.0208

Penrose, R. (1999). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford Paperbacks.

Peters, M A, & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife*, *4*, e09651. doi:

Phillips, I. (2018). The methodological puzzle of phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170347. https://doi.org/10.1098/rstb.2017.0347

Pilz, K. S., Zimmermann, C., Scholz, J., & Herzog, M. H. (2013). Long-lasting visual integration of form, motion, and color as revealed by visual masking. *Journal of Vision*, *13*(10), 12. https://doi.org/10.1167/13.10.12

Pinker, S. (2007). *The mystery of consciousness*.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (eds.), *Art, Mind, and Religion*, (pp. 37–48). University of Pittsburgh Press.

Putnam, H. (1988). *Representation and reality*. MIT press.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79. https://doi.org/10.1038/4580

Rees, G., Kreiman, G., & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience*, *3*(4), 261–270. https://doi.org/10.1038/nrn783

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575. https://doi.org/10.3758/BF03207490

Ro, T., Breitmeyer, B., Burton, P., Singhal, N. S., & Lane, D. (2003). Feedback contributions to visual awareness in human occipital cortex. *Current Biology*, *13*(12), 1038–1041. https://doi.org/10.1016/S0960-9822(03)00337-3

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*(3), 329–359. https://doi.org/10.1007/BF00355521

Rosenthal, D. M. (2002). Explaining consciousness. *Philosophy of Mind: Classical and Contemporary Readings*, 406–421.

Rosenthal, D. M. (2004). Varieties of higher-order theory. *Advances in Consciousness Research*, *56*, 17–44.

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, *428* (September 2017), 106–131. https://doi.org/10.1016/j.jtbi.2017.05.032

Rüter, J., Kammer, T., & Herzog, M. H. (2010). When transcranial magnetic stimulation (TMS) modulates feature integration.

European Journal of Neuroscience, *32*(11), 1951–1958. https://doi.org/10.1111/j.1460-9568.2010.07456.x

Salti, M., Harel, A., & Marti, S. (2019). conscious perception: Time for an update? *Journal of Cognitive Neuroscience*, *31*(1), 1–7. https://doi.org/10.1162/jocn_a_01343

Schäfer, A. M., & Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. In International Conference on Artificial Neural Networks (pp. 632–640). Berlin, Heidelberg: Springer.

Scharnowski, F., Rüter, J., Jolij, J., Hermens, F., Kammer, T., & Herzog, M. H. (2009). Long-lasting modulation of feature integration by transcranial magnetic stimulation. *Journal of Vision*, *9*(6), 1. https://doi.org/10.1167/9.6.1

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Attention, Perception & Psychophysics*, *68*(3), 489–504. https://doi.org/10.3758/BF03193692

Searle, J. R. (2000). Consciousness, free action and the brain. *Journal of Consciousness Studies*, *7*(10), 3–22.

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron*, *61*(5), 700–707. https://doi.org/10.1016/j.neuron.2009.01.016

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400. https://doi.org/10.1038/nn1549

Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., & Tallon-Baudry, C. (2013). Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Current Biology*, *23*(2), 150–155. https://doi.org/10.1016/j.cub.2012.11.047

Seth, A. K. (2016). *The hard problem of consciousness is a distraction from the real one – Anil K Seth | Aeon Essays*. Aeon. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one

Seth, A. K. (2018, February 1). Conscious spoons, really? Pushing back against panpsychism. *NeuroBanter*. https://neurobanter.com/2018/02/01/conscious-spoons-really-pushing-back-against-panpsychism/

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, *22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

Singer, W. (2007). Binding by synchrony. *Scholarpedia*, *2*(12), 1657. https://doi.org/10.4249/scholarpedia.1657

Stoljar, D. (2017). Physicalism. *The Stanford Encyclopedia of Philosophy, Winter 2017 Edition*. https://plato.stanford.edu/archives/win2017/entries/physicalism/

Taylor, J. G. (2007). Special issue: commentary on the'small network'argument. *Neural Networks*, *20*(9), 1059–1060.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, *21*(4), 753–759. https://doi.org/10.1016/S0896-6273(00)80592-9

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Soc.*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*(March 2020), 102877. https://doi.org/10.1016/j.concog.2020.102877

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), 1096. https://doi.org/10.1038/nn1500

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, *7*(5), 207–213. https://doi.org/10.1016/S1364-6613(03)00095-0

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, *100*(10), 6275–6280. https://doi.org/10.1073/pnas.0931489100

Ward, E. J. (2018). Downgraded phenomenology: How conscious overflow lost its richness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170355. https://doi.org/10.1098/rstb.2017.0355

White, P. A. (2018). Is conscious perception a series of discrete temporal frames? *Consciousness and Cognition*, *60*(April 2018), 98–126. https://doi.org/10.1016/j.concog.2018.02.012

Wilson, H. R. (2003). Computational evidence for a rivalry hierarchy in vision. *Proceedings of the National Academy of Sciences*, *100*(24), 14499–14503. https://doi.org/10.1073/pnas.2333622100

Appendix

Linking consciousness to neural or physical processes

Orchestrated Objective Reduction (Orch OR). Based on Gödel's incompleteness theorems, the Orchestrated Objective Reduction (Orch OR) theory by Penrose and Hameroff proposes that mental aspects such as understanding, free will or insight, are not Turing machine computable (Penrose, 1999). Hence, a non-computable mechanism must be at work. Penrose proposed the objective reduction (OR) of quantum wave functions, which has an inherently non-computable stochastic component. OR events are proto-conscious moments, without meaning or information. Proto-conscious OR events are proposed to be a basic property of the entire universe. A minimal moment of consciousness occurs every time an OR event occurs. However, these OR events are usually not organized, like individual instruments in an orchestra being tuned: noise rather than music. In the brain, OR events do not occur at random. Instead, microtubules 'orchestrate' (Orch) OR events in such a way that meaningful conscious moments emerge, resulting in Orch OR moments of full, rich conscious experience (music rather than noise). In this way, human consciousness corresponds to Orchestrated Objective Reduction (Orch OR) events. Orch OR events are not functionally inert: they are proposed to select particular classical states of microtubules which then govern neuronal function, e.g., regulating axonal firing to exert causal action on behavior.

Orch OR links putatively non-computable phenomena such as free will, creativity, insight and understanding to consciousness. A quantum mechanism explains these cognitive phenomena. There is evidence that terahertz quantum vibrations in microtubules are dampened during anaesthesia (Burdick et al., 2019). However, this does not support the link between consciousness and OR. Indeed, Orch OR probably also occurs in unconscious processing. Hence, there is no unconscious alternative, so Orch OR cannot address paradigm cases. Furthermore, quantum effects in anaesthesia may simply alter neural processing without any link between OR and consciousness. As a side comment, it is also far from consensual that Orch OR events can occur at all in the brain or that they can explain cognitive effects (but see Hameroff & Penrose, 2014).

In summary, because it does not address paradigm cases, Orch OR needs to explain why it targets consciousness *per se*.

Causal structure theories

Recurrent Processing Theory (RPT). Lamme (2006) argued that a neural description of consciousness is needed because reportability might not always be adequate for measuring consciousness. For this reason, he proposed that consciousness can simply be measured by whether or not recurrent processing is present. In vision for example, there is first a forward sweep of information processing that occurs unconsciously. Consciousness emerges when subsequent recurrent processing allows different specialized visual regions to communicate. The idea motivating this identification of consciousness with recurrent processing is the following: it was shown that recurrent processing is required for perceptual organization and grouping (Lamme et al., 1998), which are functions that, it is argued, are explicitly linked to conscious experience (Lamme, 2015). For this reason, RPT proposes that recurrent processing is necessary and sufficient for consciousness, making it a causal structure theory.

RPT addresses paradigm cases (masking for example in Lamme, 2006), and recurrent processing is a mechanism with both conscious (there is recurrent processing) and unconscious (there is no recurrent processing) alternatives. The other systems argument does not apply: other systems are conscious when they process information recurrently.

The unfolding argument (III.2) suggests that there may be a double dissociation between RPT and paradigm cases of consciousness because any computational task performed by a recurrent network can be carried out identically by feedforward networks. The small network argument (III.3) arises since a two-neuron network can implement recurrent processing and should therefore be conscious, leading to a sort of panpsychism. For the same reason, RPT also needs to address the large network argument (III.3.d).

In summary, RPT needs to explain how it copes with the unfolding and small network arguments.

Computational Theories

Predictive Processing Theory (PPT). The brain is often seen as a predictive processing machine (Clark, 2013; Friston, 2010; Rao & Ballard, 1999). The idea is that the brain uses a generative model to explain its input stream with a complex web of top-down predictions. For example, if a dog excites the retina, the generative model predicts what the retinal activation looks like, and how it will change. Failures to predict the sensory input result in prediction errors, and top-down predictions try to cancel these error signals. We experience a structured world with dogs, cats, houses, and even abstract entities such as parliaments or mental states because they are contents of the generative model. Predictive processing has often been associated with consciousness (Friston, 2013; Rudrauf et al., 2017; Seth & Tsakiris, 2018). For example, Rudrauf et al. (2017) identify core aspects of consciousness with a set specific computational mechanisms, of which predictive processing is paramount (3D spatial phenomenology of subjective experiences also plays an important role in their model, but we focus here on the predictive processing component of the model).

PPT addresses paradigm cases such as binocular rivalry for example: only one image at a time is inferred as the cause of the input stream. The other image is not, and therefore it is not perceived (unconscious alternative). The unfolding argument does not apply, because predictive processing can be unfolded (Doerig, Schurger, et al. (2019); but see Kleiner (2019) and Kleiner and Hoel (2020)).

The main question is what defines predictive processing? If it is restricted to the human brain, there is no computational understanding of the crucial characteristics and the other systems problem arises. If instead it is defined by simple mechanisms (such as classic predictive coding networks proposed by Rao & Ballard, 1999), the small network argument kicks in.

In summary, PPT needs to show that it can cope simultaneously with the small network and other systems arguments.

Adaptive Resonance Theory (ART). Grossberg (2017) proposed that consciousness occurs when neurons are in an Adaptive Resonant (AR) state. AR states occur when top-down expectations are combined with bottom-up sensory information. Top-down expectations take the form of a memory template that is compared with the actual features of an object as detected by the senses in a recurrent loop. All conscious states are proposed to be AR states, but the converse is not true. In this framework, consciousness is one of several emergent properties of self-organizing neural systems, which work together to enable brains to autonomously learn to attend, recognize, and predict objects and events in an ever-changing world. AR states provide a mechanism to combine sensory input, expectations, attention, memory, etc. into a coherent subjective percept. It is proposed that hierarchical AR computations are needed to deal with uncertainty. In computational models, AR states match subjective effects such as illusory contours, which are only in the eye of the beholder (see IV.1 and the Hermann-Hering grid). These effects are not 'in' the stimulus, but they are present in conscious experience.

ART offers both conscious and unconscious alternatives (there may or may not arise an AR state after stimulus presentation for example), and addresses paradigm cases of consciousness such as masking. There is no other systems problem: systems that implement AR states are conscious (although the fact that not all AR states are conscious states may pose a problem in this respect). However, AR states can be implemented with very few neurons (even a hierarchy of AR computations can be implemented with <100 neurons), so the small (and the large) network argument applies (III.3). Moreover, depending on the interpretation of ART, the unfolding argument may apply (III.2).

In summary, ART needs to specify how it copes with the small network argument and how exactly conscious resonant states differ from unconscious resonant ones.

Cognitive theories

Sensorimotor Theory (SMT). The sensorimotor theory of phenomenal consciousness (as proposed initially in O'Regan and Noë (2001) and developed more extensively in Noë (2004) and O'Regan (2011)) proposes a view about what sensory experiences or 'feels' really consist of. Instead of assuming that feels are things that are generated by the brain and happen to you, the theory suggests that feels should be understood as 'things that you do'. Understood this way, the quality of a feel lies in the law that describes the sensorimotor interaction involved when you experience the feel. For example, the softness of a sponge is an abstract law that describes the fact that when you press the sponge it squishes under the pressure of your fingers. Feeling the softness of the sponge involves mentally probing whether at this moment your interaction with the world obeys the sensorimotor laws of softness. More generally, having a feel with a particular quality means being currently mentally poised to confirm that the sensorimotor laws corresponding to that quality are valid. For example, vision feels different from audition because sensorimotor input from seeing depends differently on your movements than from hearing.

SMT can address paradigm cases of consciousness by proposing different sensorimotor interactions in conscious vs. unconscious cases. The unfolding argument does not apply, since sensorimotor loops could be unfolded. The sensorimotor loops proposed by SMT seem very complex so that the SNA does not apply. The main challenges for SMT stem from its vague mechanism. The question is what defines sensorimotor interactions? If sensorimotor interactions are restricted to humans, there is no generalizable understanding of the crucial characteristics and the other systems problem arises: what makes human brains special? Which other systems have this special kind of sensorimotor interactions and are conscious? One issue for SMT in trying to propose a precise mechanism to cope with the other systems argument is that the small network argument may kick in. If, for example, sensorimotor interactions are defined as information processing loops or other simple computational features, a thermostat would be conscious.

Self Comes To Mind (SCM). For Damasio (2010), the self is the key to consciousness. The self is a collection of neural processes centred on representing and monitoring the state of the body in order to maintain homeostasis. The crucial step in the emergence of consciousness is not about perception, i.e., the creation of the content of consciousness, but making the percepts *our* percepts. The self is vital because it acts as a witness to the mind, and this is the only way we can know about mental events. Thus, we become conscious of events when the corresponding representations interact with the self.

The brain maps the world around it and it maps its own properties. Those maps are experienced as images in our minds. The special kind of mental images of the body produced in body-mapping structures, constitutes the *protoself*. Interacting with an object leads to that object's representation in maps, and changes the state of the body, thus altering to protoself. This brings the object into consciousness. It becomes salient. This intrinsically present-moment form is termed *core consciousness*. In brains endowed with abundant memory, language, and reasoning, narratives with this same simple origin and contour are enriched, thus producing a well-defined protagonist, an *autobiographical* self. Thus, the entire fabric of a conscious mind is created from the same cloth – images generated by the brain's map-making abilities. Critchley et al. (2004), Park and Tallon-Baudry (2014) and Faivre et al. (2015) have also argued that self representations in the brain are crucial for consciousness.

The proposed mechanism is the interaction of the self and other representations. SCMT is not geared towards paradigm consciousness but is rather about subjectivity and cognition in general, such as how the inner and outer world are mapped and how these maps may interact. Nevertheless, state paradigm cases such as coma vs. wakefulness are addressed (they correspond to different states of the self representations) as well as content paradigm cases such as masking (because the mask precludes the target from interacting with the self representations). There is an unconscious alternative when the object representation and the self do not interact. The other systems argument is a challenge, asking what is special about the human implementation of the self. The mechanism seems to vaguely defined to tell which other systems are conscious.

In summary, SCM needs to explain what is special about the human self, to cope with the other systems argument.