The objects of consciousness: a non-computational model of cell assemblies

Abstract

The premise of this paper is that an adequate model of consciousness will be able to account for the fundamental duality in experience typified by thought *and* feeling, objectivity *and* subjectivity, science *and* art, and that it will do so without any of these terms assimilating its counterpart. The paper argues that such an account is possible using existing models of the cell assembly, but only if consciousness is conceived in structural rather than information-processing terms. To this end, the paper contests the viability of information-processing models that identify consciousness with a substrate-independent flow of information, and instead identifies consciousness with the physical structure of the cell assembly itself. This allows a fuller and more parsimonious account of consciousness than existing information-processing models, as well as the integration of a range of key related matters from the fields of neuroanatomy, psychology, philosophy, and the physical sciences.

Overview of cell assemblies

The networks of neurons that populate the brain go by a range of names including 'neuronal ensembles', 'neural networks', and 'cell assemblies'. These terms are often synonymous, but the present paper focusses on the cell assembly (CA), best known from the work of Donald Hebb (1949). Although 'there is no general agreement on its definition', the CA (cell assembly) is generally understood as an 'organized collection of neurons' that functions as an 'internal cognitive process' and thereby 'represents a particular concept' such as an object.[1] CAs are populations of neurons, and the term 'population coding' denotes a fundamental principle of cell assemblies whereby a given concept is 'coded by a set of neurons that fire at an elevated rate when the concept is perceived' or recalled (H&P, p. 266). Population coding is contiguous with 'overlapping coding', another tenet of Hebb's theory according to which 'some, perhaps most, neurons participate in multiple CAs', and 'new concepts composed of

[1] Huyck & Passmore, 2013, pp. 263-4; Harris, 2005, p. 403. Subsequent references to Huyck & Passmore (2013) will be abbreviated to 'H&P'. All italic emphasis in quotations is my own.

pieces of old concepts will contain some new neurons, though they will also contain some of the neurons from the base concepts' (H&P, p. 275).

The CA comes into being as 'an anatomically dispersed set of neurons among which excitatory connections have been potentiated' due to repeated co-activation of those neurons. The 'theory is based on the premise, now known as Hebb's rule, that synaptic connections become strengthened by synchronous activity of presynaptic and postsynaptic neurons' (Harris, 2005, p. 400). In other words, CAs are learned (i.e. established by experience), and once established, a CA is able to 'act as a categoriser of sensory stimuli' such that 'the presentation of an object (for example, an orange) to an individual may cause a particular CA to become active, allowing the individual to identify the object as an orange.' Moreover, the same CA can also 'be activated without direct sensory stimuli, so a person's CA for [an] orange will ignite when they think about an orange'. 'The CA hypothesis (and indeed Hebb's related learning rule) has been increasingly supported by biological, theoretical, and simulation data since it was made' (H&P, p. 263).

Overview of information-processing models of the brain

In recent decades, CA modelling has been influenced by the widespread hypothesis that the brain is a structure or hierarchy of information-processing units. In this paper, the term 'information processing' refers to any model that presents the brain (either in part or totality) as an input-operation-output system in which the throughput is encoded and processed as information. This information-processing paradigm has become the textbook account of functional brain anatomy, whereby both neurons and the larger brain systems of which they are component parts are described as information processors.[2] At the micro level, this information processing is typified by the signalling between neurons, across synapses, by neurotransmitters; at the macro level the brain itself constitutes the processor, whereby 'the *primary input* to the brain comes from our senses', and the '*primary output* of the brain is the control of our muscles and movements, which are the basis of behaviour and language' (Hill, 2014, p. 121). The functional relations between these inputs and outputs can broadly be

[2] See, for example, Crossman and Neary, 2000, p. 33.

described as 'cognition', which is also conceived in information-processing terms as the 'mental activities involved in acquiring and *processing information*' (Colman, 2015).

For brevity, I will refer to information-processing based models as 'IP models', and take them to be most fully exemplified by computational models of the brain. Computational models assert that even though brains are radically different from existing computers, they share a fundamental identity insofar as both are systematic information processors (Dehaene, 2014, p. 106). The computational model is typified by the claim that brains 'surely encode *information*' by (for example) 'transducing inputs into patterns of chemical and electrical *information*'; that they operate 'over that encoded *information*'; and that brains therefore '*are* computers, in the sense of being systems that operate over inputs and manipulate information systematically' (Marcus, 2014, p. 209).

Critique of IP models of consciousness

Recent work on cell assemblies has likewise proceeded in terms of information processing, based on the premise that in the CA itself, 'information is passed between neurons' in the form of neuronal spike trains (H&P, p. 266). The present paper will contest this position by arguing (i) that IP modelling of the brain is only possible if the singular nature of consciousness is overlooked, and (ii) that developing an adequate brain-based model of consciousness—i.e. one that *does* attend sufficiently to its unique character—will require relinquishing the IP hypothesis entirely. In making this argument, the term 'consciousness' is used to refer to awareness, that is to say the experiential character of brain activity—'the state or fact of being mentally conscious or aware of something'.[3]

Consciousness presents unique problems to scientific inquiry due to its subjective character, and in consequence consciousness has not been a mainstream subject of investigation in neuroscience.[4] For example, Huyck and Passmore's 2013 review of over 200 academic publications on cell assemblies does not mention 'consciousness' once. Such reticence is well founded because when the character of consciousness *is* sufficiently attended to, IP-based models exhibit two serious problems, viz (i) the explanatory gap, and (ii) what will be referred

[3] *OED (Oxford English Dictionary) Online*. Oxford University Press, September 2016. Hereafter *OED*
[4] There are notable exceptions such as the work of Stanislas Dehaene

to as 'the locale problem'. In brief, the explanatory gap is the failure of IP models to provide a persuasive account of why consciousness *feels* like it does; while the locale problem is the failure of IP models to provide a persuasive account of the spatiotemporal locale of consciousness. A fuller account of both problems, together with an explanation of why they are endemic to IP models are given in the appendix. The immediate aim of this paper, however, is to offer a brain-based, *non*-IP model of consciousness, and to show how this non-IP model not only avoids the problems inherent in IP models, but has altogether greater explanatory power with regards to consciousness itself.

As mentioned above, CAs instantiate 'concepts'. These concepts include representations of physical objects, and this type of CA forms the focus of the present paper as it most clearly relates Hebb's model to the broader questions of perception, memory, and representation that this paper takes up, as well as to the entailed matters of subjectivity and objectivity. In order to show the distinct way in which CAs instantiate representations of physical objects, the model begins with a brief preliminary account of the general structure of physical objects, in contrast to that of informational representations of those same objects.

The structure of physical objects

All physical objects have two aspects: every object is both (i) a unitary entity that can be meaningfully distinguished from other objects, and (ii) a spatiotemporally-distributed structure of that object's constituent elements.[5] At any scale—from atomic to cosmic—the character of a given object derives not simply from the totality of its constituent parts, but from the spatiotemporal relations of those parts. For example, a block of ice, a pool of water, and a cloud of steam may have the same constituent elements as one another (H_2O molecules), but the spatiotemporal relations between those parts differs in each case, and those differences give these objects their distinct characteristics. The 'relations' under discussion here are not a figure of speech, they are space and time themselves, the dimensions through which any given object has its physical being, its place in the cosmos. Physical objects could not exist if their constituent elements had no spatiotemporal extension (as in the state of the universe before the Big Bang), yet that spatiotemporal distribution is

[5] The term 'object' is synonymous with 'system' in this context

not a constituent element, as such, of the objects themselves; space and time (and by extension, causality) are only manifest in and through objects.

The structure of information

In these terms, the key difference between a physical object and an informational representation of that same object is that while any given physical object is a unique, material instantiation of a spatiotemporal structure, the informational representation of that object is not. For example, if either the constituent elements *or* the spatiotemporal structure of a physical object is changed, then the object itself is changed (as in the ice-water-steam example above). This is not, however, true of the informational representations of that same object: a digitized image, set of measurements, descriptive account, and so on, are *not* bound to a specific physical form, and the same digitized information could be instantiated on paper, in silicon, in pulses of light, and so forth.

This distinction matters for the question of consciousness because even though the brain itself is a physical object (i.e. a spatiotemporal structure of constituent elements), IP models hypothesize that states of consciousness (as information processes or outputs) have no intrinsic relationship to the underlying structural states of the brain.[6] As Matteo Carandini puts it, neural computations 'resemble a set of instructions in a computer language, *which does not map uniquely onto a specific set of transistors*, or serve solely the needs of a specific software application' (Carandini, 2014, p. 181). In other words, in IP models, the flow of information that is assumed to constitute consciousness is not bound to a specific physical form or 'hardware': it is decoupled from the brain's neural substrate and could, in principle, be instantiated in an entirely different substrate such as silicon. This is the basis of the claim that 'research in neural computation needs not rest on an understating of the underlying biophysics' (Carandini, 2014, p. 180). And it is this decoupling of consciousness from brain (through the medium of information processing) that the present paper contests.

Contesting the information-processing hypothesis is not a radical enterprise given that neural encoding is itself not proven. As Gary Marcus—a strong advocate of computational models—writes, 'we know (or think we know) roughly what neurons do, and that they communicate

[6] See appendix for more detail

with one another, *but not what they are communicating*' (pp. 214). And although neural spike trains in CAs *may* be evidence for temporal coding, other explanations are available: for example, they 'might instead reflect an underlying organization of cell assemblies' (Harris, 2005, p. 399). Despite this, the idea that consciousness may be grounded in something other than information processing has been deemed 'pre-scientific' (Dehaene, 2014, p. 262), so it should be emphasized at this point that the present model invokes no exotic concepts such as panpsychism. More specifically, this paper does *not* contest the existence of signalling within CAs at the biophysical level, but it *does* contest the interpretation of this neural activity as the encoding, transmission, and processing of information. By analogy, the pulse is an intrinsic part of bodily function, and a raised or lowered, strong or weak pulse can be read as indications of that body state, but this does not mean that the pulse pattern is a stream of encoded information that has been processed by the heart.

In short, CAs are not necessarily information processors, but they are by definition neural structures (H&P, p. 264), and the present paper argues that their structural characteristics offer a stronger basis for a model of consciousness than any putative IP function. Hence instead of interpreting the synaptic activity of neurons in a CA as the encoding and transmission of information prior to its later manifestation as consciousness, the proposed non-IP model conceives that activity as the interrelated elements of a single neural-conscious object (the CA) interoperating as a unified whole and thereby manifesting a specific conscious experience. This changes the key question in this area from one of how consciousness might be an outcome of information processing, to one of how the structure of conscious states correlates to the structure of brain states.

Cell assemblies as 'neural-conscious' objects

Information processing is predicated on input-operation-output systems which inherently push consciousness (as *out*put) downstream of *in*put (e.g. cortical stimulation).[7] As there is no persuasive model of how this conscious output is instantiated (see appendix), the present model proposes that even without an explanatory mechanism for consciousness, brain

[7] Even parallel and re-entrant IP systems are inherently linear because of the movement from input, through processing, to output.

modelling is both better served and more parsimonious if consciousness is located *in situ* at the point of neural activity. The non-IP model therefore locates the conscious experience at the site of the CA itself, meaning no intermediary encoding or processing of information is required between stimulation of the CA and the corresponding conscious experience. Hence in the non-IP model the CA constitutes both (i) a neural object, and (ii) the corresponding conscious experience, and CAs can therefore be described as 'neural-conscious' objects.

As neural-conscious objects, CAs allow consciousness to be modelled as a physically-structured experience rather than as a physically-decoupled flow of information. This is because CAs, like other physical objects, are themselves organized structures of constituent elements. This structure derives from the formation of the CA: for example, repeated sensing of a given physical object stimulates a specific combination of neurons in the sensory cortices, and connections (e.g. strengthened synapses) are thereby established between those neurons, resulting in the creation of the relevant cell assembly. The object of perception is thereby reflected in two different aspects of the CA: (i) the neuronal 'nodes' (i.e. the specific neurons in the sensory cortices stimulated by the physical object), and (ii) the neural structure (i.e. the connections between those neurons which reflect the spatiotemporal structure of the object of perception itself). And because the CA is both a physical (neural) structure *and* the corresponding conscious experience, the dual-aspect character of the former is reflected in the dual-aspect character of the latter. In the discussion that follows, these dual experiential aspects of the CA will be referred to as 'phenomenal' and 'noumenal' consciousness.

Phenomenal consciousness: structure and character

In this paper, the term 'phenomenal' refers to that which is known to the senses, and is used with particular reference to the awareness of physical objects. And because the non-IP model locates consciousness *in situ* at the point of neural activity, any CA that corresponds to the phenomenal apprehension of a physical object will necessarily be at least partly located in the relevant sensory cortices (cf. H&P, p. 272 ff.). For example, Stanislas Dehaene provides the following account of a CA corresponding to a given physical object (in this case, a fire extinguisher):

> All the neurons in a cell assembly support one another by sending excitatory pulses. As a result, they form a delimited 'hill' of activity in neural space. And because many such local assemblies can activate independently at different places in the brain, the outcome is a combinatorial code capable of representing billions of states. For instance, any visual object can be represented by a combination of color, size, and fragments of shapes. Recordings from the visual cortex support this idea: a fire extinguisher, for instance, seems to be encoded by a combination of active 'patches' of neurons, each comprising a few hundred active neurons and each representing a particular part (handle, body, hose, etc.)
> (2014, p. 175)

As Dehaene shows here, the CA is made up of constituent elements (in this case the 'color, size, and fragments of shapes') located in the sensory cortices. These rudimentary elements of sense experience are sometimes termed 'protophenomena', and it is 'generally assumed that the firing of neurons linked by synapses within microscopic circuits gives rise to the basic phenomena of mind-making, conveniently called the "protophenomena" of cognition' (Damasio, 2012, p. 252).

The current paper takes 'protophenomena' in its literal sense ('proto': early stage; 'phenomena': 'known to the senses') to refer specifically to the constituent elements of sensory (i.e. phenomenal) experience. These protophenomena, located in the sensory cortices, are not individual neurons, but small neuronal populations. They are functionally 'atomic' because they correspond to sensory experiences (such as seeing a uniform patch of colour) that do not resolve into more fundamental perceptual elements. Through overlapping coding, these protophenomena are foundational to many different, larger CAs, meaning that from a comparatively sparse vocabulary of protophenomenal resources it is possible to have consciousness of, as Dehaene puts it, 'billions of states'.

The structure and character of perceptual objects

The senses are able to detect some aspects of physical objects but not others (for example, the naked eye can see a pool of water, but not the individual water molecules). Moreover, some of the more familiar characteristics of perceptual objects, such as colour, do not actually inhere in the objects themselves. What is perceived depends on the nature and functioning of both the sense organs themselves, and on the corresponding protophenomena. In consequence, there is not a simple correlation between objects in themselves and objects as perceived. To mark this distinction, the term 'physical objects' will be used to refer to the things in themselves, while the term 'perceptual objects' will be used to refer to objects as they are known through the senses (i.e. as they are established as CAs). This is an important distinction, but it does not mean that the relationship between physical and perceptual objects is arbitrary. On the contrary, there is a coherent, structured relationship between the two, sufficient that we can recognize the same objects across space and time, and that we live in a world of shared perceptual objects with our fellow beings. This correlation is provided for by the nature of the CA, and will be described in more detail below.

Noumenal consciousness: structure and character

To recap, the non-IP model proposes that (i) consciousness is present at the level of the CA itself, instantiated as protophenomenal populations located in the sensory cortices, and (ii) these protophenomena can combine, through co-activation, in the CAs that constitute perceptual objects. Yet this is not the whole story, because the co-activation of the constituent parts of a CA does not in itself instantiate consciousness of the perceptual object as a unified entity, but merely the consciousness of its constituent protophenomenal parts. Recognizing a given physical object requires not simply awareness of those constituent parts, but the ability to grasp that particular combination and interrelation of parts as constituting a unitary object. As will be seen, this is not a rhetorical distinction: the difference between seeing only the parts, and seeing those parts as a unified whole, is powerfully exemplified by visual agnosia, a condition in which this ability is impaired such that 'the individual has intact

elementary vision but cannot identify the nature of objects' (Crossman & Neary, 2000, p. 147).

The matter can be brought into focus by this question: what is it that differentiates (i) consciousness of a number of unrelated protophenomena (such as 'colors, sizes, and fragments of shape'), from (ii) consciousness of the same protophenomena as constituting a unitary perceptual object (such as a fire extinguisher)? It is not the introduction of any additional protophenomenal elements, as they are the same in each case. In the non-IP model, the answer is provided by the nature of the CA itself. As has been said, the CA exists due to repeated co-activation of its constituent neurons which establishes a 'high mutual synaptic strength' between them: that is to say, a CA is not only an assemblage of co-activated protophenomenal elements, but also a physical structure of associations between those elements (H&P, p. 264). And because in the non-IP model, consciousness is located at the level of the CA itself, those neural associations are not simply useful bits of 'wiring', but are—as an intrinsic part of the CA—a constituent part of the conscious experience that the CA in question instantiates.

In the present model, these neural associations correlate to what will be referred to as 'noumenal' consciousness, and it is this noumenal consciousness that enables the apprehension of the unity of objects. For example, recognizing a fire extinguisher requires not only sensing the constituent protophenomena, but *knowing* that they are part of the same object, and recognizing their relational unity. This noumenal *knowing* is not sensory, because it is not protophenomenal, rather, it corresponds to the neural association of the protophenomenal elements, and thereby constitutes a qualitatively distinct type of consciousness. In short, CAs have a dual-aspect structure, viz (i) the neuronal (e.g. protophenomenal) elements, and (ii) the neural associations of those elements; and the dual aspects of this structure correspond to the dual aspects of consciousness, viz (i) the *phenomenal*—'that which can be known by the senses', and (ii) the *noumenal*—that which is 'knowable only by the mind or intellect, not by the senses' (*OED*).[8] Please note that in this context the terms 'phenomenal' and 'noumenal' refer only to consciousness at the level of

[8] This usage does not imply any wider (e.g. Kantian) meaning

the constituent parts of a CA (e.g. unorganized protophenomena), not to the apprehension of the CA as a unified object.

At the cortical level, it might be inferred that this distinction between the phenomenal and the noumenal would map directly onto the widely-used distinction between 'primary' and 'associative' cortex. However, the latter terms are troubled (Zeki, 1993; Kaas, 1999), so the present model proceeds without invoking general cortical divisions into primary and associative areas. Instead, the model assumes that the nature of the physical associations that structure the CA (e.g. strengthened synaptic connections between co-activated neurons) will differ in scale, locale, and intricacy depending on the type and complexity of the CA in question (H&P, p. 272). After all, CAs can vary greatly in size, from an estimated 10^3 to 10^7 neurons per CA, and the neural connections in large-scale CAs are more complex than the synaptic connections at the local level of protophenomena, given that the cortex may 'support CAs that cross brain areas and thus integrate features over a range of complexities and modalities' (H&P, pp. 264, 266).

Normal object perception and recognition require both the phenomenal and noumenal aspects of consciousness to be intact, yet both are susceptible to damage. In the case of the former, damage to protophenomena can mean loss of the phenomenal elements of sensation such as the ability to see (and imagine) a given colour.[9] In the case of the latter, damage to the neural associations of those protophenomena can have an equally dramatic effect on sensory perception. This is evident from lesion studies in which the 'primary' (i.e. protophenomenal) sensory cortex is intact, but the cortical associations between those protophenomenal elements are damaged. The brief accounts of agnosia and amnesia that follow describe the impact of this sort of damage on the perception and recall of objects, and also on the recall of the relationships between objects.

Agnosia as damage to cell assembly structure

Visual agnosia 'is caused by lesions in the visual association cortex, sparing primary visual cortex' (Álvarez, 2016, p. 85). The condition takes different forms because the association of

[9] 'focal brain damage often causes simultaneous deficits in perception and imagery' (Damasio, 2012, pp. 149-150)

protophenomena can be interrupted both at a fundamental level of the protophenomena themselves ('visual form agnosia'), and at the level of the integration of those protophenomena into perceptual objects ('associative visual agnosia'). The two forms are outlined below.

Visual form ('apperceptive') agnosia is the more rudimentary condition, in which individuals 'have an absolute inability to recognize the simplest forms, which impedes them from, for example, differentiating a straight line from a curve or determining the size of objects' (Álvarez, 2016, p. 86). In this case the protophenomena themselves are active but lack even the most basic structural organization: as Martha Farah puts it, there is 'preserved stuff vision in the absence of thing vision', 'a kind of rich but formless visual goo', which is to say 'visual form agnosics lack the ability to *group* local visual elements into contours, surfaces, and objects' (Farah, 2004, p. 18-19). In short, at this rudimentary level the neuronal elements of perception are present, but not the structure; there is phenomenal consciousness but without the organizing form of noumenal consciousness; the stimuli are registered but do not make sense. Visual form agnosia is associated with damage to subcortical white matter in the occipital lobes and surrounding regions (Farah, 2004, p. 25).

Associative visual agnosia, by contrast, appears to be a condition in which the protophenomena *are* functioning correctly, but there is damage at a neural level that interrupts their binding into CAs that correspond to more complex unitary objects (i.e. perceptual objects). Hence an individual suffering from associative visual agnosia may be able to perceive the constituent parts of an object, but will have 'disproportionate problems with the global structure of complex objects relative to simpler shapes or simpler parts of shapes' (Farah, 2004, p. 77). Here, noumenal consciousness is functional at the level of the organization of protophenomena into basic shapes, but not at the level of organization of those basic shapes into more complex objects. As might be anticipated, associative visual agnosia has a different neuropathology from visual form agnosia, and is usually associated with bilateral occipitotemporal lesions (Farah, 2004, p. 88).

The agnosias indicate how, within a specific modality such as vision, damage to the structure of a CA impacts on the integrity of the perceptual object in question. This matter of structural integrity may, however, extend beyond perceptual objects, as certain forms of

amnesia—particularly those affecting episodic memory—present a corresponding case at a higher level of complexity again, whereby the perceptual objects that make up a remembered scene are intact, but the noumenal relations between them are interrupted such that the scene cannot be bound together, and the memory itself is lost.

Amnesia as damage to cell assembly structure

Memory is sometimes thought of as being like a photograph or film footage that is encoded as information, stored in the brain, and later retrieved. This is the case in prevailing IP models of memory[10] which depend on 'distinct perceptual and mnemonic tokens' such that 'visual recognition requires that memory be searched for a representation that resembles the current stimulus input' (Farah, 2004, pp. 76, 72). In the non-IP model however, memory is not a stored record of information, but is—as in the case of perception itself—a neural structure (i.e. CA) that can be reactivated. This is why in the non-IP model a CA correlates to *both* the perceptual recognition of a given object and to the memory of that object: both the perception and the memory of a given object involve activation of the same CA, and the CA is, to borrow Martha Farah's phrase, 'a single perceptual-mnemonic representation'. Consequently, in the non-IP model, amnesia is not the loss of self-contained, photograph-like memories, rather it is the inability to establish or retain the associations between the relevant perceptual objects that would enable them to constitute a remembered scene, event, or narrative.[11] So, whereas in visual agnosia there is an inability to recognize the spatial integration of parts into unitary objects, in some forms of amnesia there may be a corresponding inability to visualize the spatiotemporal integration of perceptual objects into unified scenes or narrative structures. This non-IP model of memory is consistent with a range of recent work on memory, particularly on episodic memory and the hippocampus.

Episodic memory is 'long-term memory for personal experiences and events' (Colman, 2015) and is dependent on the hippocampus: if the hippocampus is destroyed, episodic memory is lost. Nonetheless, even in the face of such loss, the constituent elements of episodic memories—e.g. the CAs that correspond to perceptual objects—may remain intact: patients

[10] See, for example, Burgess & Hitch (2005)
[11] See Rosenbaum, et al., 2005, and Maguire, et al., 2016

with hippocampal damage and a corresponding loss of the ability to recall scenes can nevertheless still bring to mind single objects (Maguire, et al., 2016, p. 432). For example, the amnesic individual may not be able to recall the scene of a day at the beach, but may nonetheless be able to bring to mind the individual objects that might make up such a scene (the sea, the sun, a beach towel, and so on). Hence in this kind of amnesia the loss seems to be not of the constituent perceptual objects themselves, but of the ability to bind those objects into spatially-unified or temporally-ordered scenes in the contexts of recall and visualization. In response to such findings, Eleanor Maguire and colleagues have proposed a 'Scene Construction Theory', which 'posits that a primary function of the hippocampus is to facilitate the construction of scenes by allowing details to be martialled, bound, and played out in a coherent spatial context' (Maguire, et al., 2015, p. 433).

The cell assembly hypothesis is more usually associated with the mental instantiation of objects than of memories, but the latter is also present in Hebb's work. And although some of the recent works on memory use Richard Semon's term 'memory engram' rather than Hebb's 'cell assembly', they nonetheless draw on Hebb's principles as the two models are cognate (Tonegawa, et al., 2015). These recent studies provide empirical evidence that memories themselves are indeed cell assemblies ('engram cell pathways') distributed across multiple connected brain regions. 'For instance, the hippocampus and the associated cortex are known to play a crucial role in episodic memories by associating the emotionally neutral components of the episode: information like what, where, and when' (Tonegawa, et al., 2015, p. 926). In like manner, in the non-IP model, an episodic memory is a complex CA in which perceptual objects are co-activated in networks involving areas outside the sensory cortices including, critically, the hippocampus.

Unity of perceptual objects

In the cases above, amnesia impedes the recall of unitary scenes or narratives, while agnosia disrupts the unitary perception of objects. Under normal perception, however, we experience perceptual objects in a unitary way. For example, although we only ever see objects partially (due to perspective, occlusion, and so on), we do not live in a confusing world of fragments, but experience those partial glimpses as unitary entities. This unitary experience of objects is not an optional supplement to perceptual awareness, as we cannot

prevent ourselves from seeing perceptual objects as unitary forms (rather than as aggregates of disparate elements). In other words, we cannot induce agnosia. Why is this? In the non-IP model, perceptual objects are unitary experiences because of the way in which CAs are activated or 'triggered'. Take the example of vision: despite its composite, saccadic character, vision is not experienced as a flicker of successive images, nor of separate impressions of colour, shape, and movement. This is because a given CA can 'ignite' (i.e. be activated in full) 'when only a small subset of its neurons initially fire' (H&P, p. 265). For instance, 'an object partially hidden behind another one can frequently be identified' because 'full ignition of a cell assembly' is able to occur when only some of its neurons are stimulated – such as when an object is glimpsed (Pulvermüller, 1999, p. 256). And because a given CA is both the neural and the conscious counterpart to a given perceptual object, when that CA is triggered (i.e. partially stimulated), it can ignite in full, meaning that the CA (i.e. the perceptual object) becomes conscious in its totality almost immediately. In this way, the smoothness of perceptual experience is accounted for by the relative stability of established CAs: CAs are the stable neural-conscious unities that underlie the flux of sensory stimulation. As stated earlier, consciousness is not a stream of processed sense data, but the activation of the CAs—and therefore the corresponding conscious states—to which the initial stimuli correlate.

Words and concepts

In the examples considered so far, the unitary character of the perceptual object (i.e. CA) derives from the unitary character of the corresponding physical object, and once that perceptual object has been established, any sensory stimulation provided by the corresponding physical object (such as a glimpse of it) will ignite the CA, instigating a perceptual recognition of the whole. This ignition can, however, also be triggered by stimuli that are *not* a part of—and do *not* derive from—the physical object in question. This can occur through the association of words with objects, whereby the association of a given word with a given object allows them to function as a single CA. The word 'dog', for example, can ignite the relevant perceptual object (i.e. mental image) even though the word 'dog' itself does not derive from, and has no intrinsic physical or sensory relation to the object (animal)

in question. For the same reason, different languages can have different words for the same object.

This independence of words from physical objects also allows words to act as triggers for perceptual formations that are organized by concepts rather than by the unitary character of objects themselves. This is evident in our ability to name and recognize parts of objects even when those parts do not have an existence as independent physical units. For example, a human body is a discrete, unitary, physical object, but a forehead, temple, or elbow is not. The latter are conceptually—but not physically—discrete parts of the body, the knowledge of which comes *after* the establishment of the perceptual object (CA) of which they are part (we recognize what a face is before we learn what a forehead or temple is). Groups of discrete physical objects can likewise be grasped as unitary wholes (e.g. seeing stars as constellations), and again, this sort of conceptual grouping comes *after* we establish the ability to perceive individual stars and the night sky at large. These conceptual formations are not 'natural' (insofar as they are not determined by the unitary integrity of the physical objects themselves), but are cultural: they are public, conceptually-determined naming conventions that require learning.

Conceptual objects

Physical objects cannot be known directly (they are outside the body), but as has been shown, CAs are a means by which those objects may be instantiated as conscious experiences. Moreover, CAs allow physical objects to be apprehended both as *perceptual* and as *conceptual* objects. A perceptual object is the awareness of the phenomenal aspect of the CA, manifest as the unitary sensory apprehension of the physical object in question. A conceptual object is the awareness of the noumenal aspect of the CA, manifest as the intellectual apprehension of the same physical object as a unitary structure of spatiotemporal relations.

All objects—physical, perceptual, conceptual—are by definition unitary entities, but the unitary character of perceptual objects is different from that of conceptual objects. To grasp a physical object as a *perceptual* object, is to grasp it as a sensory whole, a phenomenal unity, a unified totality of protophenomena. By contrast, to grasp a physical object as a *conceptual*

object, is to grasp it not in phenomenal terms, but as the totality of spatiotemporal relations between its constituent parts; it is to know the object intellectually. The difference is exemplified by the contrast between, say, the sight of a triangular patch of light, and the thought of 'a plane rectilinear figure having three angles and three sides' (*OED*).

It is important to note at this point that perceptual objects are not simply 'phenomenal objects' and conceptual objects are not simply 'noumenal objects'. As the discussion of agnosia indicated, perceptual objects depend on both phenomenal *and* noumenal consciousness (i.e. on both the protophenomena themselves, and on their interrelations in the cell assembly). This also holds true for conceptual objects, which despite instantiating intellectual (rather than sensory) apprehensions of objects, likewise depend on some level of phenomenal consciousness in order that those noumenal relations may be known at all (just as space and time are only made manifest in and through physical objects).

Although perceptual objects require noumenal consciousness, it remains muted, whereas in conceptual objects this noumenal aspect is to the fore. Conceptual objects involve the apprehension of the structure of physical objects, and thereby depend on the subdivision of those objects into constituent conceptual parts (as described earlier). Once these parts are established and differentiated, the relations between them can be articulated, and it becomes possible to say of an object that this part is next to that, that is before this, after this will come that, and so on. The ability to observe and articulate meaningful spatiotemporal relations (such as cause and effect) between parts of an object is what characterizes rational thought itself, and in the present model, conceptual objects are the basis of rational thought.

The parts of a conceptual object are therefore related to one another in a similar fashion to the discrete but interrelated objects that constitute episodic memories (as described earlier). Indeed, the present model proposes that just as the same protophenomena may be common to a wide range of perceptual objects (via overlapping coding), so the same abstract spatiotemporal relations may be common to a wide range of conceptual objects (again via overlapping coding). In consequence, these relations do not need to be established anew for every conceptual object, but rather the 'over,' 'after,' 'above,' and so on that obtain between the parts of one conceptual object may be shared not only by other conceptual objects, but

also by, for example, the interrelated objects that constitute remembered or imagined scenes.

In terms of the structure of the CA itself, the proposal is that while the protophenomenal elements of perceptual objects are located in the sensory cortices along with the corresponding 'natural' noumenal associations (i.e. those which are instantiated by the unitary form of the physical object itself, and which are susceptible to agnosia), the development of a conceptual understanding of those same objects may extend the structure of the CA in question outside the sensory cortex and into, for example, the hippocampus.

Representational objects

This final section of the paper offers greater detail on the distinction between conceptual and perceptual objects through their representational counterparts in scientific models and works of art. The significance of science and art in the present context is that they not only differ in their respective foci on objectivity and subjectivity, but that this difference is instantiated in their distinct representational forms. Scientific models constitute abstract, analytical instantiations of the spatiotemporal relations between the elements of a given object, while works of art synthesize disparate materials into new phenomenal unities that embody subjective experience in a communicable form. These two types of representation are incommensurable, which is why an equation is so different from a painting, an opera cannot be meaningfully performed in hexadecimal code, a scientific hypothesis cannot be usefully expressed as a watercolour, and so on.

Scientific models

Science is objective: it is concerned with the conceptual representation and understanding of physical objects. The function of the scientific model, as a representational object, is to manifest the spatiotemporal relations that constitute the physical object under consideration. Science is concerned with the spatiotemporal relationships between and within things (of *a* to *b*), not the things themselves (*a*, *b*). As Niels Bohr concisely expressed it, the purpose of the scientific description of nature 'is not to disclose the *real essence of*

phenomena but only to track down as far as possible *relations between* the multifold aspects of our experience.'[12]

This focus on the structure of physical objects is not concerned with invoking the subjective character of perception. Scientific representation therefore instantiates these relations without reference to the perceptual character of the object under consideration. The matters of concern here are abstract and relational: space, time, causation and so on have no independent phenomenal character. The scientific model therefore bears no intrinsic relation to the phenomenal character of the object being modelled, it only requires component parts that manifest internal relations which correlate to those of the physical object in question. For example, certain aspects of the spatiotemporal relationship of the sun to its satellites may be modelled by an orange and some walnuts, but a cup and some coins might do the job just as well. This is because the phenomenal form of the representation (walnuts, coins) is immaterial to the spatiotemporal relations under consideration, all that matters is that the materials have the capacity to model the object's spatiotemporal relations with the desired degree of accuracy: the focus is on the spatiotemporal structure of the object, not its phenomenal character. Because the latter is incidental to science, it can be pared down to a minimum, and the complexities of the cosmos can be modelled with the phenomenally-minimal language of mathematics and theoretical physics (variables, symbols, equations, formulae).

To put that in other terms, scientific representation is abstract because it is not concerned with instantiating the sensory character of its subject: there is no colour, sound, or aroma to an equation. Scientific hypotheses are not improved (as scientific hypotheses) by being set to music, or redrafted in iambic pentameter, and an oil painting of a molecule offers no advantage (and quite possibly disadvantages) over a blackboard sketch or a computer-generated model of the same. In this context, subjective responses would be a distraction, and scientific language is therefore objective, disinterested, unfeeling, and indifferent to the expressive qualities of language such as metaphor. Scientific models do not speak the

[12] Quoted in Velmans and Schneider, 2007, p. 302

language of the body, senses, pain and pleasure, and therefore do not instantiate compassion, empathy, or subjectivity.

Works of art

Art is subjective: it provides an embodied experience. To be a subject means to have a body, senses, and other corporeal experiences such as the emotions. Art is only concerned with physical objects insofar as they are experienced through the body, i.e. as perceptual objects. The work of art instantiates the character of subjectivity as a representational object that can be shared with others. It provides a sensory, emotional experience, a structure of feeling, *not* a piece of information.

Whereas scientific models are *conceptually* unified, works of art are *perceptually* unified. Thought may contribute to the creation of a work of art, but the final work itself is not conceptual, abstract, or disembodied. On the contrary, the composition of a work of art involves coordinating the relationship of its component elements (i.e. its materials), whatever they may be, into a harmonious and resonant relationship that itself constitutes a phenomenally-unified representational object. If the work of art does not attain this unitary character—if it is perceptually fragmented—the subjective experience will be broken up into elements that need to be conceptually related.[13] This phenomenal dis-integration is an aesthetic shortfall equivalent to the failure to rationally integrate the elements of a scientific model.

Art is not objective. It is concerned with manifesting the character of subjectivity, not with the spatiotemporally exacting replication of physical objects (photographic reproduction is not the acme of artistic achievement). To this end, art—unlike science—is not concerned with modelling accurate spatiotemporal relations between its elements, but rather with integrating those elements as a perceptual unity that discloses the character of subjectivity. Hence it is no error if a landscape painter moves buildings or magnifies mountains or renders vegetation as geometrical forms, nor if a dramatist conflates historical figures or dates. This same latitude of interpretation is present in performance art, whether the reading of a poem,

[13] It is, of course, possible to do this volitionally, to objectivize works of art by considering them in conceptual terms, and this is the function of art criticism etc.; in much the same way it is possible to subjective science, making it 'fun', 'exciting', 'elegant', and so on in the manner of popular science. Whatever their merits, popular science and art criticism are second-order activities distinct from science and art themselves.

staging of a play, or playing of a musical score. The re-forging of the relations of the constituent elements into a new unity can bring a new emotional dynamic to a work, one which would not be present were the piece to be rendered with complete accuracy by a computer. A MIDI rendition of a piano sonata is not better than a virtuoso performance of the same, even though the former may have greater tonal, dynamic, and temporal consistency, and can be a more objective reproduction of the written score.

To attain this perceptual unity, art is attentive in the highest degree to its materials and to the integration and synthesis of these materials into a sensual unity through the possibilities offered by metaphor, harmony, timbre, tone, proportion, colour, and so on. These processes may interfere with objectivity, but allow works of art to manifest and feelingly communicate empathy, compassion, subjectivity, and so on.

In conclusion

To summarize: conscious experiences (exemplified in this paper by perceptual and conceptual objects) are not information-processing outputs, but the activation, however fleeting, of cell assemblies (i.e. as physical structures rather than information conduits). A perceptual CA is established through encounters with a given physical object leading to repeated stimulation of the same group of protophenomena. The integrated structure of associations intrinsic to that CA (e.g. strengthened synaptic connections) reflect the spatiotemporal integrity (unitary form) of the physical object itself. The dual-aspect nature of physical objects is thereby reflected in the dual-aspect structure of the CA, and this is the ground of the dual aspect (i.e. phenomenal and noumenal) character of consciousness. This dual-aspect structure unites—and is common to—physical objects, to CAs (as perceptual and conceptual objects), and to the representations of those objects as exemplified by scientific models and works of art.

This non-IP model provides a structure and spatiotemporal locale for consciousness, a parsimonious account of perception and memory (and corresponding anomalies), a closure of the explanatory gap, a distinct account of the neural basis and dual-aspect character of consciousness, and a means to bind the matter of consciousness to that of space and time. It does so by attending to the range and qualitative character of conscious experience—

subjectivity as well as objectivity; feeling as well as thought; the arts as well as the sciences—and without assimilating any of these terms to its counterpart.

Appendix

The locale problem

IP models of consciousness typically envisage the brain as being composed of a diverse range of localized, unconscious information processors, variously referred to as 'agents', 'units', 'demons', 'homunculi', and so on (Dennett, 1993, pp. 261-2; Dehaene, 2014, pp. 175-8). These processors operate in parallel, across functionally-differentiated areas of the brain, and loosely correspond to CAs. In these computational models, consciousness itself is typically envisaged as 'global information sharing' of the outputs of these local unconscious processors, perhaps functioning as a 'lingua franca' across the different units (Dehaene, pp. 163, 91).[14] Such models are not, however, clear about the locale of consciousness, and whether it is intrinsic to that information-processing function, or an output of it.

For it to be *intrinsic*, consciousness would need to be endemic to the processing function whereby 'global information sharing' takes place. But this operation is itself a mechanism—a 'router' in Dehaene's terms—a kind of Turingesque serial processing amid the parallel processing of the brain (Dennett, 1993; Dehaene, 2015), and the mechanical transparency of a router or a Universal Turing Machine offers no solution here because it reinstates the problem of Leibniz's mill: all physical parts of the system can be inspected, but no consciousness found.

Could, then, consciousness be an *output* of such a system? Might the 'input—operation—output' that characterizes processors be envisaged in terms of, say, 'senses—cognition—muscles' (Hill, 2014)? Here the physical system of inputs and outputs would be clear, but in the case of consciousness this model stalls because there is no physical counterpart to the muscles where consciousness, as an output of that process, might exist. That is to say there is no post-processing locale where the various outputs of distributed unconscious processing might coalesce as conscious experience. So although the distributed processing of computational models has the merit of avoiding problems such as the homunculus, the

[14] for variations of the model, see, for example, Carandini, p. 177 ff., and Dennett, p. 276 ff.

'central meaner,' and the 'Cartesian theatre' (Dennett, 1993), it does not overcome them. The alternative accounts that computational models offer—such as 'global broadcasting' to a 'global workspace' (Dehaene, 2015)—do not solve the locale problem, but rather reinstate it in different terms, replacing consciousness-as-local-integration with consciousness-as-global-distribution.

Daniel Dennett, as an advocate of computational models of consciousness, is aware of this locale problem (1993, pp. 254-6), yet it persists in his own work. Take a stick and touch things around you, he suggests, and the 'transactions between stick and touch receptors under the skin […] provide the information [that] your brain integrates into a conscious recognition of the texture of paper, cardboard, wool, or glass, but these complicated processes of integration are all but transparent to consciousness' (1993, p. 47). Here Dennett invokes an unconscious information-processing stage (i.e. the 'information' undergoing 'all but transparent' 'processes of integration') from which the 'conscious recognition' emanates, but the model provides no spatiotemporal locale for the output of the processing (the 'conscious recognition') to occur.

In the face of this locale problem, computers offer a beguiling model of the brain not only because they are paradigmatic information processors, but because they implicitly circumvent the locale problem by outsourcing consciousness to users via interfaces such as screens, keyboards, and speakers. This outsourcing can be deceptive, so to be clear: despite the claims of IP models, it is the computer operator—not the computer itself—that provides the locale, the 'Cartesian Theatre' in which the processed information 'integrates into a conscious recognition'. The attribution of consciousness to a computer in such an arrangement gets the model the wrong way round: the computer itself is not a conscious entity served by a passive user, rather the unconscious computer functions as an extended sense for the conscious user in a manner similar to the 'extended mind' of Andy Clark and David Chalmers (1998). It is, moreover, noteworthy that the information-processing ability of computers is actually incidental to this role, which is why the extended mind model works equally well with a mobile phone, a paper notebook, a computer, or indeed a stick.

The explanatory gap

A widely-discussed difficulty faced by IP accounts of consciousness it that they *feel* incomplete. This shortfall is referred to as 'the explanatory gap' which—although it is not always cast in these terms—is centrally concerned with information processing, as Joseph Levine (who coined the term) makes clear:

> The basic problem [constituting the explanatory gap] is this. If we consider the sequence of events that begin, say, with light reflected off a ripe tomato and ending with a visual experience of a red shape, the role of the intervening physical events, from detection of the light by retinal receptors to activity in the visual cortex, seems to be solely a matter of *implementing certain causal-informational roles*. That is, so long as we are looking for an *information-processing story*, we have an idea how appeal to the neurological events in the visual system could explain its implementation. But when we reflect on what we want explained when considering the nature of conscious visual experience, we see that it is not exhausted by *the information concerning the external world it undoubtedly contains*. We want to know, as well, why it is like what it is like for us to have this experience. Why do red things look just that way and green things differently, and why is there not a 'way it's like' for cameras connected to computers to detect *the very same information*? (Bayne, et al., 2009)

To put that in the terms of this paper, it seems that we can provide satisfactory rational accounts ('information processing stories') of the processes of perception, but the same accounts do not satisfy the feeling (the 'what is it like?') of the experiences themselves. Why should this be? The matter might be intractable were it not that the type of account being looked for here is in fact already available to us: 'information processing stories' are not our sole representational resource, because we have other—equally compelling—modes of communication, such as the arts, which provide just the kind of fulfilling representation of certain kinds of conscious experience (such as looking at a ripe tomato) that the explanatory gap identifies as being absent from information-based accounts. Hence the incommensurability of the sciences and the arts has a direct bearing on the explanatory gap.

The explanatory gap draws attention to the need for close attention to the question of representation in our accounts of consciousness. Without that attention, not only will the explanatory gap remain unresolved, but we will be left with a fundamental asymmetry in the discussion of consciousness itself. How so? Because rational thought—to the degree that it has been taken as cognate with information-processing and the ordered representations of science—has been largely passed over as a mode of consciousness, and 'the problem of consciousness' has come to be assigned chiefly to sensory experience instead, usually under

the rubric of 'qualia' (i.e. noumenal consciousness has been overlooked, and phenomenal consciousness has been identified as the 'problem'). This is evident in Levine's account whereby the 'information-processing story' is implicitly a scientific account, and *knowing* that account is treated as inherently unproblematic, with the problem of consciousness only arising with the question of *feeling* (the 'what it is like for us'). But *knowing* is just as much a problem of consciousness as feeling is, and unless conscious thought is adequately differentiated from the unconscious information processing of machines, IP models will not only fail to provide an explanation of consciousness, but will continue to obscure the actual issues at stake.

Bibliography

Álvarez, R., Masjuan, J., 2016. Visual agnosia. *Revista Clínica Española* 216, 85–91.

Bayne, T., Cleeremans, A., Wilken, P. (Eds.), 2009. *The Oxford Companion to Consciousness*, OUP, Oxford.

Burgess, N., Hitch, G., 2005. Computational models of working memory. *Trends in Cognitive Sciences* 9, 535–541.

Carandini, M., 2014. From Circuits to Behavior: A Bridge Too Far?, in: Marcus, G., Freeman, J. (Eds.), *The Future of the Brain*. Princeton University Press, Princeton, pp. 177–185.

Clark, A., Chalmers, D., 1998. The Extended Mind. *Analysis* 58, 7–19.

Colman, Andrew M., 2015, *A Dictionary of Psychology*, 4th ed. Oxford University Press.

Crossman, A.R., Neary, D., 2000. *Neuroanatomy*, 2nd ed. Churchill Livingstone, Edinburgh.

Damasio, A., 2012. *Self Comes to Mind*. Vintage, London.

Dehaene, S., 2014. *Consciousness and the Brain*. Penguin, New York.

Dennett, D.C., 1993. *Consciousness Explained*. Penguin, London.

Farah, M.J., 2004. *Visual Agnosia*. A Bradford Book, Cambridge, Mass.

Harris, K.D., 2005. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience* 6, 399–407.

Hebb, D.O., 2002. *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, Mahwah, N.J.

Hill, S., 2014. Whole Brain Simulation, in: Marcus, G., Freeman, J. (Eds.), *The Future of the Brain*. Princeton University Press, Princeton, pp. 111–124.

Huyck, C.R., Passmore, P.J., 2013. A review of cell assemblies. *Biological Cybernetics* 1–26.

Kaas, J.H., 1999. The transformation of association cortex into sensory cortex. *Brain Research Bulletin* 50, 425.

Maguire, E.A., Intraub, H., Mullally, S.L., 2016. Scenes, Spaces, and Memory Traces What Does the Hippocampus Do? *Neuroscientist* 22, 432–439.

Marcus, G., 2014. The Computational Brain, in: Marcus, G., Freeman, J. (Eds.), *The Future of the Brain*. Princeton University Press, Princeton, pp. 205–215.

Pulvermüller, F., 1999. Words in the brain's language. *Behavioural and Brain Sciences* 22, 253–336.

Rosenbaum, R.S., Köhler, S., Schacter, D.L., Moscovitch, M., Westmacott, R., Black, S.E., Gao, F., Tulving, E., 2005. The case of K.C. *Neuropsychologia* 43, 989–1021.

Tonegawa, S., Liu, X., Ramirez, S., Redondo, R., 2015. Memory Engram Cells Have Come of Age. *Neuron* 87, 918–931.

Velmans, M., Schneider, S., 2007. *Blackwell Companion to Consciousness*. John Wiley & Sons, Oxford.

Zeki, S., 1993. The visual association cortex. *Current Opinion in Neurobiology* 3, 155–159.