The neural basis of consciousness

Chris D. Frith

Abstract

Consciousness has evolved and is a feature of all animals with sufficiently complex nervous systems. It is, therefore, primarily a problem for biology, rather than physics. In this review, I will consider three aspects of consciousness: level of consciousness, whether we are awake or in a coma; the contents of consciousness, what determines how a small amount of sensory information is associated with subjective experience, while the rest is not; and meta-consciousness, the ability to reflect upon our subjective experiences and, importantly, to share them with others. I will discuss and compare current theories of the neural and cognitive mechanisms involved in producing these three aspects of consciousness and conclude that the research in this area is flourishing and has already succeeded to delineate these mechanisms in surprising detail.

Introduction

The scientific study of consciousness remains as controversial as ever. It is widely taken for granted that this is a ‘hard’ problem (Chalmers, 1995) and some thinkers (known as mysterians McGinn, 2012) claim that the problem is sufficiently hard that we will not be able to solve it. Nevertheless, accounts of the neural basis of consciousness continue to be developed and arguments between proponents can be fierce (e.g. Odgaard et al., 2017). There is an interesting distinction between the theories that get media exposure and those that are discussed at scientific meetings such as the Association for the Scientific Study of Consciousness (see e.g. Michel et al., 2019). The media like philosophical approaches such as panpsychism, leading to headlines such as ‘Everything from spoons to stones is conscious’ (Goldhill, 2018). In contrast, neuroscientists and psychologists like experimental data and there is now a great deal of such data to discuss. In these circles, the dominant approach is global workspace theory (Baars, 1988; Dehaene, 2005) to which I will return later in this essay.

The standard definition of being conscious, which I will use here, is having subjective experiences. Two aspects of this definition are particularly worth considering. The first is sometimes known as the level of consciousness. At one extreme, when I am in a coma or deep sleep, I am not having any subjective experiences. At the other extreme, when I am wide awake, I am having many rich and complex experiences. In relation to this aspect of consciousness, we can ask ‘what neural systems are necessary and sufficient for being conscious, rather than being in a coma?’

The second aspect is known as the content of consciousness, that is, what it is that I am subjectively experiencing. This is a specific conscious state (sometimes called state consciousness). The key insight driving research on this topic is that much precise and well-adapted behaviour in a fully awake person can occur without consciousness (Jacob and Jeannerod, 2003; Goodale and Milner, 2004). Furthermore, there are cases where performance actually deteriorates when people become conscious of what they are doing (Bolick et al., 2002). It is therefore possible to contrast brain activity associated with neural processes that lead to consciousness and those that do not. The paradigm example of this approach is binocular rivalry. When different stimuli are presented to the left and right eye (e.g. a house and a face) we do not see a mixture of a house and a face. Our conscious perception alternates every few seconds between a house and a face, even though the sensory input does not change. We can use this paradigm to track the neural activity that changes with the conscious experience (Leopold and Logothetis, 1996). This strategy is often referred to as the search for the neural correlates of consciousness (Rees et al., 2002a; Kim and Blake, 2005).

In this review, I shall treat consciousness as a problem for biology rather than physics. I assume that consciousness is a feature of living things and that it has evolved (Carruthers, 2000; Bronfman et al., 2016). There was unconsciousness before consciousness, and humans have richer conscious experiences than many other creatures. I will also set aside the hard problem, making no attempt to explain how physical processes in the brain can lead to subjective experiences. Instead, I shall consider some of the easier
problems, exploring the neural processes that seem to be necessary and sufficient for different aspects of consciousness.

Markers of subjective experience

To identify a neural correlate of consciousness we have to link neural activity with subjective experience. But how do we access subjective experience when it is not our own? The most direct way is via some form of a report. This approach is particularly useful for studying the contents of consciousness. So, in the case of binocular rivalry, we ask participants to press the left button when they see a house and the right button when they see a face. Such reports depend upon introspection and can vary in sophistication. The participant might simply report whether they saw a stimulus or not or they might indicate how visible the stimulus was on, say, a four-point scale (Ramsey and Overgaard, 2004). Introspection used to be the mainstay of experimental psychology but fell into disrepute with the rise of behaviourism. Nevertheless, simple introspective reports have been used in studies of psychophysics from the work of Fechner (1860) until today. Such studies are among the most replicable in psychology (Read, 2015). Used carefully, introspection is a robust source of data about subjective experience.

One problem with the use of report as a marker of subjective experience is that the neural activity associated with the experience may be confounded with the neural activity associated with reporting. ‘No report’ paradigms have been developed, but they cannot avoid the problem that they have to be validated via introspective reports (Overgaard and Fazekas, 2016). My own view is that reportability is not a nuisance, but an important feature of certain kinds of subjective experience (Frith and Metzinger, 2016).

Another problem with the use of reports is that these largely, although not entirely, depend on language. Of course, this is not to claim that non-human animals and preverbal infants are not conscious. But the study of subjective experience in such groups is far more difficult.

A non-verbal sign that someone is conscious could be that they engage in deliberate, intentional action rather than automatic behaviour. This sign is particularly relevant to the study of levels of consciousness. For example, a patient is typically considered to be conscious if they can perform an arbitrary action in response to a command. This is the basis of the test developed by Owen et al. (2006) to demonstrate signs of consciousness in ~20% of patients apparently in a persistent vegetative state. Here, one such command is to ‘imagine that you are playing tennis’. When such a command is obeyed, activity is seen in specific brain regions concerned with planning for action (Jeanerod, 2001). The idea is that, while the patient is unable to produce any overt movements, she can still deliberately imagine making a movement and the brain activity associated with such imagining will be generated.

As with introspection, this approach is not fool proof. There will be cases where it is difficult to tell whether an action is automatic or deliberate (see e.g. Nachev and Husain, 2007). This problem is not only of academic interest. There is a great need for a robust marker that tells us whether or not a comatose patient still has some subjective experiences. Also, the distinction between automatic and deliberate actions plays a critical role in the application of the automatism defence in law. This defence can allow a person who commits a crime to go free if the act was committed while unconscious (Rolnick and Parvizi, 2011). This defence was successfully used in the case of a man who strangled his wife while asleap (de Bruxelles, 2009). It is widely considered that people should not be held responsible for acts performed in the absence of consciousness (Shepherd, 2012).

The neural control of the level of consciousness

Our understanding of the neural control of the level of consciousness is largely based on studies of brain activity in patients in coma and related states. Through such studies, different levels of consciousness can be compared and changes in activity associated with recovery from coma can be recorded. In coma proper, the patient is entirely unresponsive. In the vegetative state, the patient goes through the sleep wake cycle (shown by EEG and other signs), but does not respond to stimuli or show any signs of intentional behaviour. A patient in a minimally conscious state shows some responsiveness to stimuli and will sometimes follow simple commands.

An important characteristic of the brain activity associated with these states of reduced consciousness is that metabolic activity is reduced. It has been estimated that consciousness will not be found where metabolic activity has fallen below ~40% of normal cortical activity (Stender et al., 2016). The reduction of metabolism in the vegetative state is most pronounced in the frontal and parietal cortex (Stender et al., 2015), two brain regions that are strongly connected anatomically. Reduction of activity is accompanied by a loss of functional connectivity between brain regions. In other words, distant brain regions become disconnected, leading to a disruption of large-scale information integration. This loss of connectivity is correlated with the degree of clinical impairment, the loss being greatest in coma patients and vegetative state, less in the minimally conscious, while patients with locked-in syndrome resemble healthy controls (Soddu et al., 2009). A similar loss of connectivity and lack of integration is seen during the unresponsiveness associated with anaesthesia (see Hudetz and Mashour, 2016 for a review).

Of particular importance is the connectivity between intralaminar thalamic nuclei and prefrontal and anterior cingulate cortices. A lack of such connectivity is seen in the vegetative state but returns to normal after recovery of consciousness (Laureys et al., 2000). Furthermore, after severe traumatic brain injury, behavioural improvements in patients in the minimally conscious state can be achieved via thalamic stimulation (Schiff et al., 2007). This stimulation has the effect of reactivating thalamocortical connectivity.

Connectivity in the cortico-cortical and cortico-thalamic brain networks runs in both directions: forwards, for example, from early sensory processing regions to prefrontal cortex (bottom-up) and back again (top-down). The feedback connections are more numerous and extensive than the feedforward connections (Salin and Bullier, 1995). Boly et al. took account of this distinction in a study of vegetative state patients. They found that the reduced connectivity in these patients was restricted to the feedback (top-down) connections, while feedforward connections were preserved (Boly et al., 2011).

My conclusion from these studies is that wakeful, contentful consciousness is associated with long range functional connectivity between distant brain regions. Of particular importance are
thalamo-cortical loops and large-scale cortical networks involving frontal and parietal cortex. In particular, it is the top-down control instantiated in these networks that is important.

**Relation to the current major theories of consciousness**

These results provide a useful framework for introducing four major theories currently available for understanding the neural bases of consciousness. All of these theories receive partial support from these studies of the neural activity associated with different levels of consciousness. I suspect that, ultimately, all these approaches will be combined in any viable account of the neural basis of consciousness.

**Integrated Information Theory** (Tononi, 2008) assumes that conscious experience requires a specific kind of integration of information from many brain regions. Long-range brain connectivity seems necessary for consciousness, but is it sufficient?

**Recurrent Processing Theory** (Lamme and Roelfsema, 2000) assumes that conscious perceptual experience requires recurrent processing (i.e. top-down feedback) in sensory areas. Top-down neural feedback seems necessary for consciousness, but is it sufficient?

**Global Neuronal Workspace** (Dehaene et al., 2003) assumes that the contents of conscious experience can be equated with the contents of a work space, a form of working memory, containing ‘momentarily active, subjectively experienced’ events that can be compared and manipulated (Baddeley, 1992; Baars, 2002). While consciousness does not seem to be necessary for contents maintained in working memory, it does seem to be required for manipulation of the content (Trübutschek et al., 2019).

On this account, consciousness requires the involvement of the fronto-parietal network associated with working memory (see, e.g. Collette and Van der Linden, 2002). The loss of activity in the fronto-parietal network associated with the vegetative state is consistent with the global workspace hypothesis. But can consciousness occur without working memory processes and the associated brain regions being engaged (see, e.g. Moutoussis and Zeki, 2002)?

**Higher-Order Theories** (Lau and Rosenthal, 2011) assume that consciousness involves thinking about (i.e. representing) first-order mental states. This requires a higher-order re-representation of, for example, seeing a face. Such higher-order representations involve frontal cortex. This theory is also compatible with the observation that frontal activity is reduced in vegetative state patients. But is an intact frontal cortex necessary for consciousness?

**Altered states of consciousness**

The characterisation of the level of consciousness as a single dimension varying from coma to alert wakefulness has been criticised as inadequate (Bayne et al., 2016). Consider, for example, the conscious experiences of a patient with dense amnesia. One such patient kept a note book in which he repeatedly wrote ‘I have just woken up’ (Wilson et al., 1995). This patient was alert and awake, but his subjective experience was very different from that of most people since it lacked any temporal continuity. This is an example of an altered state of consciousness. Bayne et al. (2016) suggest that the state of consciousness can vary along with a number of dimensions, so that altered states of consciousness should be characterised by their position in this multi-dimensional space rather than being ordered along a single dimension of the level of consciousness. It remains to be seen what these dimensions might be, but temporal continuity might be one of them.

Altered states of consciousness have more typically been associated with the effects of psychoactive drugs and questionnaires have been developed for characterising these states (e.g. Studerus et al., 2010). This questionnaire provides scores on a number of dimensions, including experience of unity, disembodiment, and complex imagery. There is evidence that different substances relocate users in different parts of a multi-dimensional space of conscious experience. For example, experiences after taking LSD are dominated by changes in visual perception (e.g. complex imagery and audio-visual synaesthesia). ‘Blissful state’ is also elevated, while ‘anxiety’ is least elevated (Carhart-Harris et al., 2016). In another study, psychedelics (LSD and psilocybin) increased the phenomenological richness of experience, while reducing the sense of bodily self (Millière et al., 2018). There is also evidence that even micro-doses of LSD can alter the perception of time (Yanakieva et al., 2018).

It is probable that the precise location in the ‘space’ of consciousness created by different substances can be related to the mode of action of the drug, for example, whether it targets serotonin receptors (e.g. mescaline, LSD, psilocybin) or opioid receptors (e.g. opium, morphine) (Zamberlan et al., 2018). This role for neurotransmitters in sculpting the space of conscious experience confirms the importance of long-range brain connectivity. Neurotransmitters alter the weighting of the various modes of connectivity in the brain (e.g. Hahn et al., 2012) and the rising monoamine pathways have distinct, modulatory effects on sensory processing (Jacob and Nienborg, 2018) (Fig. 1).

**The contents of consciousness**

The majority of studies on the neural basis of consciousness are concerned with the contents of consciousness. A classic paradigm, mentioned already, is a binocular rivalry. Here we can look at the brain activity when a face is presented to one eye and a house to the other. During binocular rivalry, brain activity shifts to match the current contents of consciousness (but see Giles et al., 2016). When we perceive the face, activity is higher in the face area of the fusiform cortex. When we perceive the house, activity is higher in the place area of the para-hippocampal cortex (Tong et al., 1998) (see also Moutoussis and Zeki, 2002).
Visual masking is another widely used paradigm for identifying brain activity associated with the content of consciousness. Using this paradigm, it is possible to identify brain activity associated with stimuli that influence behaviour in the absence of any subjective experience. This is because visual masking terminates neural processing before the stimulus is finally represented in consciousness. For example, if a fearful face is presented for 33 ms and immediately replaced by a neutral face, participants report seeing only the neutral face. Nevertheless, activity is elicited in the amygdala in response to the unseen fearful face (Whalen et al., 1998). This technique has been used in series of studies of word reading by Dehaene et al. who have shown that unseen masked words alter the processing of subsequently presented words (semantic priming) (Dehaene et al., 1998). These masked words elicit activity in the visual word form area of the fusiform cortex. When the same words are perceived consciously, activity in the visual word form area is greater and additional activity is seen in the parietal and frontal cortex (Dehaene et al., 2001). A similar result was obtained using the phenomenon of change blindness, where participants often fail to notice a change in rapidly repeating stimuli (Jensen et al., 2011). Using this paradigm, it is possible to compare three well matched conditions: no change occurs, undetected change occurs, detected change occurs. Beck et al. (2001) used this approach with faces as the changing stimuli. In comparison to no change an undetected change in the identity of the face elicited activity in the face area of the fusiform cortex (FFA). When detected changes were compared with undetected changes greater activity was seen in the FFA as well as activity in the parietal and frontal cortex. The parallels with the results of Dehaene et al. are striking (but see Moutoussis and Zeki, 2002).

A role for parietal cortex in specific conscious experiences was already suspected on the basis of the spatial neglect that can occur after damage to the right parietal cortex (Driver and Vuilleumier, 2001). In patients with such lesions, awareness of a stimulus (e.g. a face) on the left side of space can be extinguished if a second stimulus is presented on the right side. However, an extinguished face stimulus on the left still robustly activates the fusiform face area in the right cortex. When the left visual stimulus is consciously perceived, greater activity is seen in the frontal and parietal areas of the intact hemisphere (Rees et al., 2002b). These results suggest that activity in the FFA is necessary for the conscious experience of a face, but not sufficient.

**Essential nodes for specific conscious experiences**

The FFA is just one of many specific brain regions that are essential for different specific conscious experiences. For example, area V5/MT is necessary for the experience of visual motion (Zeki, 1991). But, to confirm that these nodes are indeed essential for specific conscious experiences, we need to show that conscious experience is affected when activity in these areas is directly manipulated. This can be achieved by electrical stimulation or as the result of lesions.

Electrical stimulation of discrete regions of visual cortex can elicit a variety of visual experiences the precise nature of which depends upon the location of the stimulation. Simple visual forms are elicited by stimulation of the striate cortex, experiences of colour and motion by stimulation of distinct regions of extrastriate visual cortex, and complex forms by stimulation of the fusiform gyrus (Lee et al., 2000). Visual hallucinations of complex scenes can be elicited by stimulation of the para-hippocampal place area in the inferior temporal cortex (Mégevand et al., 2014). These results also demonstrate that visual input from the retina and subcortical structures is not necessary for a conscious visual experience.

Conversely, loss of the conscious experience of colour or visual motion occurs after circumscribed damage to extra-striate areas V4 (achromatopsia, Zeki, 1990) and V5/MT (akinetopsia, Zeki, 1991).

**Not just visual experience**

The search for the neural correlates of consciousness has been dominated by studies of vision because the visual system in the brain is more amenable to study than any other system. As a result, we have much less information about the location of the essential nodes for other aspects of conscious experience. In the case of audition, secondary auditory cortex, in particular, the planum temporale (Griffiths and Warren, 2002), is probably an essential node for the experience of sound (Dykstra et al., 2017). In the rare cases where auditory cortex has been damaged bilaterally, profound deafness results. One such patient was entirely unable to interpret sounds, but could detect onsets and offsets of sounds when cued to attend (Engelien et al., 2000). More or less complicated auditory hallucinations occur when this region is activated during temporal lobe epilepsy (Florindo et al., 2006) or by direct electrical stimulation (Penfield and Perot, 1963).

The primary somatosensory cortex is likely to be the essential node for the experience of touch (Schwartz et al., 2004), while the anterior insula is likely to be an essential node for interoceptive experience (awareness of bodily states such as one’s own heartbeat, thirst, &c. Craig, 2009).

The conscious experience that accompanies voluntary actions has been repeatedly linked to the pre-SMA (Altene et al., 2013). Neuronal activity in the pre-SMA is related to the awareness of intentions to act (Lau et al., 2004; Fried et al., 2011), whereas direct electrical stimulation of this area can elicit the experience of an ‘urge’ to move a specific body part (Fried et al., 1991).

The nodes are necessary, but not sufficient: the role of parietal and frontal cortex

Although these nodes are essential for various types of content of conscious experience, they are not sufficient. Several studies suggest that long range coupling of these regions with frontal and parietal cortex are also necessary.

The parietal cortex has a role in a range of different conscious experiences (see, e.g. Bor and Seth, 2012). After the damage to this region, visual cortex activity alone is not sufficient to result in awareness. For example, after parietal lesions, patients suffer from a loss of awareness of stimuli on one side of space (unilateral neglect). In these cases, the visual cortex is intact and processing can still take place for neglected stimuli, without reaching the patient’s awareness (Driver and Vuilleumier, 2001). The same phenomenon can be observed for tactile stimuli. After parietal damage, activity in the somatosensory cortex may not lead to tactile awareness (Valenza et al., 2004). Similarly, amygdala and orbitofrontal cortex can be activated by emotional stimuli without awareness after parietal damage (Vuilleumier et al., 2002). Patients with parietal lesions can no longer report the time at which they had the intention to move, although they can still report the time at which they initiated a movement (Sirigu et al., 2004).
Disruption of parietal or prefrontal cortex in normal volunteers receiving transcranial magnetic stimulation impairs change detection (e.g. Beck et al., 2006) and alters the dynamics of binocular rivalry (Zarotezkya et al., 2010). The prefrontal cortex also has a role in awareness of action. Patients with prefrontal lesions can make normal adjustments to their movements when performing tasks in which there is sensory-motor conflict, while at the same time being unaware of the conflict (Slacheysky et al., 2001).

The limited capacity of conscious experience. These results suggest that an interaction between frontal and parietal cortex with stimulus-specific information in sensory cortices may be necessary for specific conscious experiences. But what is the role of parietal and frontal cortex in conscious experience? Prefrontal cortex and parietal cortex are the major neural substrates for the cognitive processes required for working memory, especially for the selection of the content that enters the work space (Quentin et al., 2019) and for the manipulation of the items held there (Wager and Smith, 2003). So, the role of these regions in consciousness is consistent with the idea that we can equate the contents of consciousness with the contents of working memory.

A well-established feature of working memory is that it has a very limited capacity (Cowan, 2001). It would follow, therefore, that, like working memory, our conscious experience also has a very limited capacity. At first glance, this idea seems incompatible with the extra-ordinary richness of our conscious experience. However, phenomena, such as change blindness (Jensen et al., 2011), suggest that our visual experience is actually rather sparse. It comes as a surprise to learn that very little of the information that is currently striking our senses is included in our experience. Rather, the richness of our visual experience is confined to the centre of our gaze and attention (Kouider et al., 2010), and we know that this rich experience will be carried over to wherever we may shift our gaze (Noe et al., 2000).

There is, however, another aspect of our conscious experience that makes it rich. If I am gazing at a strawberry, I am not simply experiencing its colour, I am also experiencing its shape, its smell, what it would feel like in my hand and in my mouth, and whether I can reach for it. All these lower level features are bound together to form the rich experience of the strawberry. This binding together of low-level features to form a highly compressed single representation is a form of chunking. Chunking is a long-established feature of working memory, often put forward as a mechanism for overcoming its limited capacity (Miller, 1956; Gobet et al., 2001). For example, it is noted that we can remember four letters, four words or four sentences, revealing the requirement to chunk items has been shown to robustly activate the prefrontal-parietal network (Bor and Seth, 2012).

I suggest that this process of compression is not simply a mechanism to compensate for a limited capacity system. The right kind of compressions can create representations that allow better communication of information within the space of working memory and also when reporting to others (Shea and Frith, 2019). Coarse graining, for example, provides a compressed representation in which the effective information is greater than would be the case if unlimited capacity allowed for representation at a microscale level of detail (Hoel, 2017). Such simplifications are also necessary in order to solve complex, multi-step planning problems (Huys et al., 2015).

Parietal and frontal cortex are the likely substrates in which these compressed representations are created. For example, parietal cortex has been implicated in sensory feature binding (e.g. Friedman-Hill et al., 1995; Rohe and Noppeney, 2015), while the requirement to chunk items has been shown to robustly activate the prefrontal-parietal network (Bor and Seth, 2012).

In summary, these studies offer a more precise definition of the neural basis of the contents of consciousness. This aspect of consciousness relies on an interaction between two systems, the essential nodes (e.g. the fusiform face area), which determine the specific content, and a frontal and parietal network, which binds the content into the compressed and enriched form that we experience. This formulation is compatible with Global Workspace Theory (Dehaene et al., 2003), but also suggests that the other theories provide additional relevant explanations. Critical here is long-range connectivity. This is required so that information can be integrated in a manner that optimises compression. This would be consistent with a version of the Integrated Information Theory (Tononi, 2008). Top-down feedback from higher-level cortical regions to sensory cortical areas is involved, implying that recurrent processing is necessary (Lamme and Roelfsema, 2000). There is clearly a role for frontal cortex in the conscious experience via some sort of interaction with lower-level sensory information, but whether this involves the kind of re-representation proposed by Higher-Order Theories (Lau and Rosenthal, 2011) remains unclear.

A higher level of consciousness: explicit metacognition

The contents of subjective experience discussed so far are sometimes referred to as being at the object level (or C1 Dehaene et al., 2017), as, for example, in my experience of the strawberry: a representation in consciousness of an object in the outside world. But there is another level of consciousness, the meta level (C2), which monitors the representation at the object level (Nelson and Narens, 1990). For example, I might suddenly realise that I have been so fixated on the strawberry, that I have stopped listening to what my dining companion is saying. I have become aware that my mind has wandered from my intended focus (Schooler, 2002). This higher level of consciousness, sometimes referred as meta-consciousness or self-monitoring, often involves a subjective experience of error (Dehaene et al., 2017).

Another way of distinguishing C1 and C2 is that C1 involves awareness of representations (contents), while C2 involves the additional awareness of processes (Shea and Frith, 2016). This distinction relates to dual process theory as applied to problem solving (Evans and Stanovich, 2013). In intuitive problem solving (system 1), we are typically aware of the problem and of the answer (C1), but not of the process by which the answer is derived. This process occurs rapidly and automatically without consciousness. In contrast, system 2 reasoning involves deliberate attention to the processes by which the problem is solved (C2).

This monitoring of cognitive processes, thinking about thinking, is often referred to as metacognition. Early studies of metacognition mostly concerned memory. A striking example is the tip-of-the-tongue phenomenon (Brown and McNeill, 1966) an example of the feeling of knowing. This is the experience of not being able to retrieve some item from memory, while, at the same time, being confident that the item is available and will soon be retrieved (Proust, 2010). Metacognitive judgement can be made about various aspects of memory. For example, prospective and retrospective confidence in one’s learning of new material.
and the ordering of one’s confidence about learning different kinds of material (Metcalfe and Shimamura, 1994). The experience of confidence has also been extensively studied in perceptual decision tasks. In such tasks, participants are asked not only object level questions about the stimulus ‘Were the dots moving left or right?’, but also meta-level questions about their perception, ‘How well could you see the stimulus?’ or ‘How confident are you in your decision?’.

Metacognition, since it involves the monitoring and control of lower-level cognitive processes, is a component of executive function (Shalllice, 1982) and therefore prefrontal cortex is likely to have a critical role (Shimamura, 2000). A recent meta-analysis examined imaging studies of metacognition in perceptual decision making and memory tasks (Vaccaro and Fleming, 2018). In these studies, brain activity was elicited by requiring participants to make judgements of confidence. The medial and lateral prefrontal cortex, precuneus (medial parietal), and anterior insula were identified as important for such judgements for both decision-making and memory.

More recently, a computational approach derived from signal detection theory has been applied to the measurement of confidence in perceptual decision tasks (Fleming and Lau, 2014). The signal-detection approach assumes that there are two aspects to metacognition in these tasks: bias of confidence and accuracy of confidence. Bias measures a general level of confidence, such that an over-confident person over-estimates the probability of being correct on every trial of a perceptual task. The person with accurate confidence is more confident in trials where she is correct and less confident in trials where she is wrong. This sensitivity of confidence to the correctness of trials is independent of the overall level of confidence (bias).

At the level of signal-detection, we can measure objectively how good people are at recognising when a signal is present or absent. At the metacognitive level, we can measure objectively how good people are at distinguishing between their correct and incorrect decisions independently of any bias they may have. In other words, people know, to some extent, when they are guessing. Accuracy at this meta-level can be independent of accuracy at the object level. A person might be good at detecting the signal, while, at the same time, not very good at knowing when they were right or wrong.

Fleming et al. looked at individual differences in metacognitive sensitivity that were independent of objective signal detection performance (Fleming et al., 2010). They found that greater metacognitive accuracy was associated with greater grey matter volume in the anterior prefrontal cortex (Brodmann area 10). A subsequent study (McCurdy et al., 2013) confirmed this result and indicated that the precuneus might also have a role in metacognitive sensitivity. The role of PFC in metacognitive accuracy has been confirmed in a lesion study and a brain stimulation study. Anterior PFC lesions had no effect on signal detection performance at the object level, but resulted in a ∼50% loss in metacognitive sensitivity (Fleming et al., 2014). These patients’ confidence reports no longer related to their performance. Application of TMS to the prefrontal cortex in healthy volunteers produces very similar results. TMS has no effect on signal detection, but it impairs metacognitive accuracy (Rouns et al., 2010).

**Monitoring our actions**

We have a very little subjective experience of the precise details of our actions (e.g. Fourneret and Jeannerod, 1998). But we are very much aware when things go wrong. We experience a loss of control when the movement that we intended is not the one that occurred. We experience a loss of agency when the effect that we intended in the outside world does not occur.

After an error has occurred in a choice reaction time task, the next response is typically slower (Rabbit, 1966). This post-error slowing indicates that the performance is being monitored. The actor detects an error and slows down as to avoid further errors. Post-error slowing is therefore a consequence of metacognition, although, in this case, no report of confidence is involved. However, this post-error slowing can occur in the absence of awareness. In a study of skilled typists, real errors of which the typist was unaware, were followed by post error slowing, while fake ‘errors’, inserted by the experimenters, were reported as errors, but were not followed by post-error slowing (Logan and Crump, 2010). A similar distinction between conscious and unconscious monitoring of errors was found when visual masking was used so that participants sometimes made decisions about targets that they classified as ‘unseen’ (Charles et al., 2013).

Even on unseen trials, actors performed better than chance. However, the error related negativity, an EEG response associated with error detection, was only elicited by errors occurring on trials in which the target was seen. These results suggest that there are two forms of meta-cognition; implicit (without awareness) and explicit (with awareness).

Source localisation based on magneto-encephalography (MEG), suggested that the unconscious detection of errors was associated with a short-lived response confined to anterior cingulate cortex. In contrast, conscious detection of errors was associated with robust activity in anterior and posterior cingulate cortex. Recording activity in neurons in the posterior cingulate cortex of the monkey has also revealed a robust response to errors during the learning of a conditional association task (Heilbronn and Platt, 2013). These responses were especially marked when performance was poor and the task was difficult. These are conditions in which the detection of errors would be highly relevant for learning.

An important aspect of the experience of action is the ease or fluency with which the action can be selected. Reduced fluency is associated with a reduced feeling of control (e.g. Chambon and Haggard, 2012) and increased likelihood of errors. When participants are asked to monitor the fluency of action selection greater activity is seen in anterior cingulate and anterior insula (Teuchies et al., 2019). When actors are required to make metacognitive judgments about the extent of their own control compared with when they are required to make judgments that are not about control (i.e. judgments about performance), increased activity is observed in the anterior PFC (Miele et al., 2011). This area is located close to the region identified in studies of metacognitive sensitivity in perceptual decisions (e.g. Fleming et al., 2010).

Imaging studies of the experience of agency have explored the ability to distinguish between whether an outcome was caused by the self or another. Activity in temporo-parietal junction (TPJ), pre-SMA, precuneus and dorsomedial prefrontal cortex (dMPFC) was higher during the experience of external-agency, while insula activation was associated with self-agency (Sperditi et al., 2011).

**Explicit metacognition and the brain: awareness of the self and the other**

The results reviewed above identify a discrete network of brain areas associated with explicit metacognition (C2, meta-
consciousness), which is largely independent of whether memory, decision making, or action is involved. This network includes anterior prefrontal cortex (Brodmann area 10), anterior insula, and midline structures, including anterior cingulate, posterior cingulate and precuneus.

The anterior insula and various midline cortical structures have previously been proposed as critical for the experience of the self. The anterior insula has been associated with interoception and the experience of the bodily self (Craig, 2009). More abstract aspects of the self, such as autobiographical memory, are associated with midline cortical structures; ventro-medial prefrontal cortex, dorso-medial prefrontal cortex, anterior cingulate, and posterior cingulate/precuneus (Northoff et al., 2006).

There is also a striking overlap between the network of brain regions associated with explicit metacognition and that associated with the ability to mentalise, suggesting that these regions are required for thinking about the thoughts of others as well as about one’s own thoughts (Vaccaro and Fleming, 2018).

**The evolution of consciousness**

At the beginning of this essay, I suggested that consciousness is a biological phenomenon found in many living creatures. In particular, I proposed that consciousness has evolved (see e.g. Carruthers, 2000). This assumption raises two questions: When did consciousness emerge in the history of evolution? What is the survival value of consciousness that drove its evolution?

The question about the survival value of consciousness is probably easier to answer, since it should be possible to identify tasks which can be achieved without consciousness and contrast these with tasks that require consciousness, or, at least, are performed much better with consciousness. But we also need a principled way of specifying the differences between the tasks so identified. In the brief speculations presented here, I shall attempt to distinguish between the classes of a task in terms of the underlying computations.

With regard to C1 or primary consciousness (sometimes called *sentience*), it is widely held that this is a feature of many animals, and not just humans (e.g. Low et al., 2012). This means that tasks that depend upon a report will not be a useful basis for our considerations. As I mentioned at the beginning of this essay, one non-verbal sign that someone is conscious is that they engage in deliberate, intentional action rather than automatic behaviour. But what is the computational distinction between automatic behaviour and deliberate, intentional behaviour?

**Automatic, unconscious behaviour and model-free learning**

Automatic behaviours can be precise and well adapted (Jacob and Jeannerod, 2003; Goodale and Milner, 2004), but are inflexible and can sometimes be difficult to suppress when inappropriate (Aarts and Custers, 2010). These automatic behaviours, sometimes referred to as habits, are acquired through a form of association learning in which direct connections are developed between stimuli and responses so that presentation of the stimulus automatically elicits a habitual response (Daw et al., 2002). They are essentially reflexes that can be modified by experience. In more recent computational formulations, they are considered to arise through *model-free* learning. Model-free learning is a slow process by which we evaluate how good (valuable) actions are through a process of trial and error. This requires no knowledge of the state of the world or of the consequences of actions. Furthermore, since the values are continually updated, no information is retained about the past. Model-free learning does not make complex computational demands, but does require extensive experience and is therefore slow (Huys et al., 2013). Studies in humans show that such automatic behaviours and the learning associated with them can occur in the absence of conscious experience of the stimuli (e.g. Eimer, 1999) or of the outcomes of actions (e.g. Pessiglione et al., 2007).

**Deliberate, intentional action and model-based learning**

In contrast to model-free learning, *model-based* learning produces rapid, flexible goal directed behaviour. While model-free systems learn about rewards, model-based systems learn about states of the world and their relationships. Model-based learning uses models of the world to evaluate possible actions on the basis of anticipated future outcomes (Doll et al., 2012). These internal models release the organism from subservience to the environment and to habitual behaviour. A prototypical example of such a world model would be the cognitive map of a maze, which enables an animal to select a novel route through it when established routes are blocked (Tolman et al., 1946). Models of the immediate past enable one of the most basic functions of working memory, that is, the ability to respond on the basis of a stimulus which is no longer physically present. Models of the more distant past, episodic memory, can be used when a new situation is encountered and a decision must be made concerning what action to take. A model of the current situation is compared with models of past situations. The action chosen is the one associated with the highest value, based on the outcomes of the past situations that are most similar to the present (Botvinick et al., 2019).

A model of the task currently being performed, for example, knowing about the set of appropriate responses, permits a simple form of *counterfactual learning*. For example, in serial reversal learning, where there are only two options, a drop in the value of one option (go left) implies an increase in the value of the other option (go right). A model-free system would be blind to such a structure since it can only update the value of the action performed. A slow process of trial and error learning would be needed to respond to the change in the state created by the reversal. In contrast, a model-based system recognises that, if the chosen option has not been rewarded, then the other option would have been rewarded and its value should be increased. This requires counterfactual reasoning, considering what would have happened if the other option had been chosen (Lee et al., 2005). At the same time a model of the task, which recognises that there are two possible states of the task, allows an immediate switch to the alternative action when the reversal occurs. The ‘learning to learn’ (Harlow, 1949) that is possible with model-based learning enables a much more rapid response to environmental changes than is possible for model-free learning.

The studies reviewed above show that model-based learning enables fast and flexible responses to changes in the environment and is plausibly identified as the computational basis for the deliberate, intentional action that is a non-verbal sign of consciousness. In this framework, the contents of consciousness would be identified as the mental models currently active for decision-making, essentially a reformulation of the idea that conscious contents can be equated with the content of working memory.
Brain imaging studies implicate subcortical areas, in particular, the basal ganglia, as underlying model-free learning. For example, learning through outcomes of which the participants were unaware (Pessiglione et al., 2007) was associated with activity in the ventral pallidum. In another brain imaging study, a distinction could be made between reward prediction errors, associated with model-free learning and state prediction errors associated with model-based learning (Gläscher et al., 2010). Activity associated with reward prediction errors (model-free) was observed in the ventral striatum, while activity associated with state prediction errors (model-based) was observed in the inter-parietal sulcus and the lateral prefrontal cortex.

These results are consistent with the claim that, in mammals, conscious experience is associated with cortical activity, with a special role for parietal and frontal cortex. However, this conclusion cannot be applied to sentient animals such as birds or cephalopods, with very different nervous systems. Rather than linking consciousness to particular brain regions, it should be linked to specific, cognitive mechanisms at Marr’s algorithmic level of analysis (Marr, 1982). The precise instantiation of the relevant algorithms in the CNS will vary from one species to another.

It has been argued that primary consciousness (minimal consciousness or sentience) emerged ∼520 million years ago during the Cambrian explosion. For Bronfman et al. (2016) the argument was based on the emergence of a CNS that could perform ‘unlimited association learning’, a version of model-based learning. For Feinberg and Mallatt (2016) a critical event in the development of sentience was the emergence of sensory maps (mental images) in the CNS, such as, for example, the topographical maps in the optic tectum of the fish. The limited consciousness experienced by fish must be very different from the subjective experience of mammals. However, this difference is quantitative rather than qualitative. All these different species are alike in having subjective experiences. But the conscious experience of the mammal is richer as a result of the evolution of, among other things, more complex sense organs and a greater information processing capacity.

**Meta-consciousness: a higher level of consciousness**

With the evolution of the processes underlying explicit metacognition, however, there was a qualitative change in consciousness. It is reasonable to assume that C2 has evolved much more recently than C1 and explicit metacognition is much more developed in humans compared to other animals (Metcalfe, 2008). While there is some evidence that non-human animals can take account of their confidence in a decision (Smith et al., 2003), it remains to be shown that this involves the kind of second-order computations that I will discuss below (see Fleming and Daw, 2017 for a discussion of how the kind of metacognition observed in non-human animals can arise from first-order computations).

Meta-consciousness provides a number of advantages over and above primary consciousness. For example, a well calibrated feeling of confidence can act a signal for learning when no external feedback is available (Guggenmos and Sterzer, 2017), since a confident decision is likely to be correct, while a non-confident one will often be wrong. Lack of confidence in our model of the world will lead us to collect more information (to explore, rather than exploit Boldt et al., 2019). Monitoring our decision processes can also help us to identify the source of an error (Purcell and Kiani, 2016). Is it a consequence of inadequate sensory processing, indicated by low perceptual confidence, or is it the result of poor understanding of the task, indicated by low response confidence? Distinguishing between these two sorts of error enhances the efficiency of model-based learning.

However, an even greater advantage arising from this higher level of consciousness is the possibility for sharing our experiences with others that it makes possible. Sharing experiences may be a uniquely human ability which can optimise social interactions (Shea et al., 2014). For example, discussions of confidence make it possible for two people to make perceptual decisions that are better than either person working on their own (Fusaroli et al., 2012). Even more important is that the ability to share experiences, such as the feeling of agency, can give rise to cultural concepts such as personal responsibility and, more generally, can enable the emergence of cumulative cultural norms (Frith, 2014).

The model-based system associated with primary consciousness probably evolved on top of an earlier model-free system, rather than separately and in parallel (Doll et al., 2012). It freed organisms from subservience to the immediate environment and from habit. In the same way, the system subserving meta-consciousness evolved on top of the model-based system and freed us from subservience to our models of the world. As a result, we can now question the validity of our models of the world and discuss them with others (Fig. 2).

Fleming and Daw (2017) have recently provided a Bayesian framework for the computation of confidence which highlights this second order aspect of meta-consciousness. They argue that correctly judging confidence requires inferences about the causes of one’s own behaviour. In other words, while the decision mechanism makes use of a model of the world, the metacognitive mechanism that estimates confidence in the decision makes use of a model of the decision-maker. Wokke et al. (2019) also concluded, on the basis of an EEG study, that explicit metacognition is a second-order process, integrating sensory evidence with estimates of the state of the decider during decision-making.

As yet we do not know precisely how the components of these computational models map onto underlying brain activity. Are...
there critical nodes that are necessary but not sufficient for the experience of confidence and other metacognitive feelings? Can we identify the higher-order brain regions with which these nodes must interact for the experience to emerge? Fleming and Daw (2017) suggest that fronto-polar cortex (FPC, Brodmann area 10) is one potential high-order convergence zone for integrating state and action information to enable the second-order computations necessary for meta-consciousness. This is consistent with its anatomical positioning at the top of a cognitive hierarchy, and the observation that it receives information from dorsolateral prefrontal, cingulate and anterior temporal cortex (Ramnani and Owen, 2004). It has been proposed that the contribution of FPC to metacognition may be to represent internal states in a format suitable for explicit communication (Fleming et al., 2012).

A role for FPC in meta-consciousness is consistent with the more recent evolution of this area. In the human brain, this region is larger relative to the rest of the brain than it is in the other apes, with more space available for connections with other higher-order cortical areas (Semendeferi et al., 2001, Donahue et al., 2018). Another brain region implicated in meta-consciousness, the precuneus, has also expanded recently during human evolution (Bruner et al., 2017).

Conclusions and possible relevance for psychiatry

While writing this review I was struck by how many empirical studies there are concerned with the neural basis of consciousness. Only a small fraction of these studies is cited here and, inevitably, the choice will reflect my biases and preferences as to the nature of consciousness. Given the wealth of information, many alternative stories can be told. With regard to the four accounts that I alluded to above, I believe that, while all have something to offer, none, on their own, provide an adequate explanation of consciousness. The brain systems underlying consciousness require long range integration of information from many brain regions (Tononi, 2008) with top-down, re-entrant processes having a critical role (Lamme and Roelfsema, 2000). Global workspace theory (Baars, 2002; Dehaene et al., 2003) remains a leading contender through its greater specificity. Brain regions associated with various aspects of working memory (frontal and parietal cortex) are consistently implicated in experimental studies of consciousness. However, higher-order thought theory (Lau and Rosenthal, 2011) also fits very nicely with the hierarchical nature of the computational approaches I have outlined above. In particular, the second-order account of meta-consciousness with its models of models resembles high-order thought theory and strongly implicates prefrontal cortex in its instantiation.

I mentioned earlier that delineating the neural basis of consciousness has already had important implications for clinical and legal practice. I believe that the more recent concept of meta-consciousness (C2) could have important implications for the study of psychiatric patients. Abnormalities of conscious experience are a defining feature of many psychiatric disorders, especially schizophrenia (Frith, 1979). But it is now possible to be much more precise about the nature of some of these abnormalities. For example, lack of insight, in the form of denial of one’s disabilities, has long been recognised as a feature of psychosis (David, 1990; Lysaker et al., 2010). Lack of insight is a failure of metacognition.

Experimental studies have linked specific failures of metacognition with positive psychotic symptoms such as delusions of control (Farrer and Franck, 2007) and auditory hallucinations (Brebion et al., 2012). More detailed examination, however, suggests that these metacognitive failures relate only to explicit metacognition and not to implicit metacognition. For example, schizophrenia is associated with failures of explicit source memory (‘remember’ responses), in the absence of failures of implicit source memory (‘know’ responses) (Danion et al., 1999). Patients in the first episode of paranoid schizophrenia are not impaired on the first-order detection of stimuli but show reduced metacognitive accuracy for the most visible stimuli (Bliksted et al., 2017). A large-scale study found that psychiatric symptoms were associated with impairments in metacognition but not decision performance, while old age was associated with changes in decision performance but not metacognition (Rouault et al., 2018).

These results suggest schizophrenia is a disorder of metacognition, rather than of consciousness more generally.

An important feature of explicit metacognition is that it enables us to reflect on our conscious experiences and describe them to others. Furthermore, our experiences and beliefs are altered by discussions with others (Shea et al., 2014). Schizophrenia is associated with great difficulty in describing even ordinary experiences to others (e.g. Cohen, 1976) and the beliefs and experiences of these patients are influenced very little by what other people say. Their delusions are defined not so much by bizarreness, but by lack of consensus with the beliefs of others. ‘A false belief [] that is firmly sustained despite what almost everyone else believes’ ‘The belief is not one ordinarily accepted by other members of the person’s culture or subculture’ (American Psychiatric Association, 2013, Psychosis & Explicit Metacognition p. 819).

Experiments developed for the study of meta-consciousness and computational accounts of the underlying neural mechanisms have the potential to greatly enhance our understanding of psychiatric disorders such as schizophrenia.

References


Notes

1 Not all reports reflect introspection. We might be reporting what we think is out there in the world, rather than our subjective experience.

2 The possibility of reaching would have two components: physical – is it within my reach?, social – is it appropriate for me to take it?


