How Experimental Neuroscientists Can Fix the Hard Problem of Consciousness

Colin Klein[1] and Andrew B. Barron[2]

1. School of Philosophy, The Australian National University, Canberra, ACT 0200 Australia
2. Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

Abstract

The contemporary search for the neural correlates of consciousness (NCCs) largely avoids the so-called 'hard problem' of conscious experience. This is due to an old, and outdated, view on which such questions are reserved for philosophers. The appearance of a hard problem is plausibly a limitation of our own relationship to underlying neural realizers. Moving past the hard problem will require (among other things) the development of safe, systematic techniques for self-manipulation of conscious experience. Much of the needed work is still hypothetical. Yet we outline a way in which advances in both the neuroscience of consciousness and in the philosophy of explanation provide a clear path forward for an integrated experimental neuroscience of consciousness.

Introduction

The modern neuroscience of consciousness begins with a division of territory. In a seminal work, the philosopher David Chalmers distinguished the *easy* and *hard* problems of consciousness (Chalmers, 1998; Chalmers, 2003). The easy problems involve sorting out the mechanisms that mediate conscious perception and action. The hard problem requires saying why activity in these mechanisms is accompanied by any subjective feeling at all. Why, in Nagel's (Nagel, 1974) evocative phrase, is there *something it is like* for you to be you, while there's nothing it's like for a rock to be a rock? As Chalmers put it:

> Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? ... Why should physical processing give rise to a rich inner life at all? (Chalmers, 1996, 201)

Chalmers and others argued that this was a deep philosophical mystery, upon which empirical evidence could have no bearing.

Meanwhile, in several influential pieces Crick and Koch argued that it might still be possible to meaningfully work around the hard problem even if. They noted that “at any one moment some active neuronal processes correlate with consciousness, while others do not.” (Crick and Koch, 1990, 263) Thus there is a viable scientific project that searches for the *neural correlates of consciousness* (NCCs). Chalmers did not disagree---the hard problem lies in *explaining* the existence of NCCs, not doubting their existence in the first place. So the search for NCCs could progress whether or not the hard problem had a solution.

Thus was the territory divided. Philosophers inherited the hard problem, scientists the easy problems and the search for NCCs. Despite occasional defectors on both sides, this truce has held for a quarter century. Yet this effective stalemate has meant that there has been no serious attempt to explain why putative NCCs actually give rise to subjective feelings. This strikes many as unsatisfactory. Further, philosophy divorced from neuroscience has begun to endorse a variety of counterintuitive views. At one extreme, there is frequent defense of panpsychism, the position that consciousness is *everywhere* (Chalmers, 2003) At the other, some philosophers assert that consciousness is simply an illusion (Dennett, 1991; Irvine, 2013). Such extremes do not feel like progress. Conversely, the search for NCCs continues to hit an impasse that looks more philosophical than empirical.

Outside of the study of consciousness, meanwhile, both fields have made striking advances. The interventionist revolution in philosophy of science has overturned the background picture of explanation on which Chalmers’ arguments depend (Craver, 2007). Comparative neuroscience has made great strides investigating the evolutionary origins of the capacities that support consciousness (Feinberg and Mallatt, 2016; Klein and Barron, 2016b; Ginsburg and Jablonka, 2019) Improvements in techniques for neurobiological interventions and observations offer the possibility for studying more than mere NCCs.

These developments have made possible a *rapprochement*. We believe that experimental neuroscience can make real progress on the hard problem of consciousness. Conversely, contemporary philosophy of science has useful tools that show how to move beyond the search for NCCs. What follows outlines what that research project might look like. We argue that development of interventions on capacities for experience, including the capacity for safe self-intervention, is ultimately necessary for moving beyond the hard problem of consciousness. This is a difficult task---but it is ultimately one that falls within the traditional bailiwick of experimental neuroscience.

Intervening on Consciousness

The core notion will be that of a *direct intervention* on a subject's conscious experience. A direct intervention is one that targets the neural underpinnings of an experience, rather than indirectly via (e.g.) perception. A neural system that allows one to alter some aspect *E* of experience will be called a *difference-maker for E.*

Despite the name, the search for neural correlates of consciousness is typically a search for difference-makers rather than correlates as such. The notion of a 'content-specific NCC', for example, is often glossed in explicitly difference-making terms (Koch et al., 2016; Boly et al., 2017)**.** Koch et al note that this means were content-specific NCCs for face perception artificially activated, "…the participant should see a face even if none is present…" (Koch et al., 2016, 208). Even if the primary evidence is correlational (i.e. through neuroimaging), this is better taken as evidence *about* potential loci of intervention (Klein, 2017).

Similarly, recent debates (Boly et al., 2017; Odegaard et al., 2017) over whether the NCCs for visual experiences are more anterior or posterior hinge on which neural activity represents the working of the NCC and that is distinct from the activity of mere precursors, background conditions, and downstream effects. Each of these notions are well-studied in the interventionist literature. Mere background conditions, for example, cannot be altered without altering many other facets of experience (and more besides).

Building up an understanding of how specific interventions on neural systems change what we experience, in specific ways, will yield an operational understanding of subjective experience. We will be mostly concerned with controlled interventions done on awake, healthy, neurotypical adults. However, many interventions that do not fit that bill still provide evidence about what *would* happen were one to intervene on healthy adults. We briefly survey experimental studies that have shown it is possible to intervene on the brain and change a person's experience in highly specific ways. There is much more to do, of course. Yet what has been done serves as useful proof of principle as well as emphasizing different types of interventions that might be made.

Most of the difference-makers studied so far have involved interventions on particular contents of consciousness. Conscious contents are features of experiences like colors, sounds, shapes, or pains. Since Penfield's pioneering work (Penfield and Rasmussen, 1950; Penfield and Jasper, 1954), we have known that direct electrical stimulation of the cortex can produce a wide variety of distinct experiences, constituting interventions on capacities. Noninvasive stimulation (e.g. by TMS) can produce similar results, though with less specificity.

In addition to intervention on contents, it is also possible to alter the broader *capacities* necessary for conscious experience on the other. The capacities necessary for conscious experience are (plausibly) functions supporting conscious experience such as selective attention, integrative and interactive processing of exteroceptive and interoceptive information, a unified spatial and temporal framework for sensory information, or unlimited associative memory (Ginsburg and Jablonka, 2007; Merker, 2007; Ginsburg and Eva, 2010; Barron and Klein, 2016).

Capacities for experience are important, we suggest, because they often correspond to structural features of consciousness. We experience visual and auditory sensations as occurring within a common, external space, for example. Yet while these structural features are the conditions for the possibility of conscious experience, they are arguably not objects of experience themselves (Kant, 1999). While we experience external sensations as occurring in a unified frame of reference, for example, we do not experience space as such, independent of the objects within it.

Broadly speaking, different brain regions appear responsible for capacities and contents of experience. Intervention on the cortex tends to change the contents of experience (Penfield and Rasmussen, 1950; Penfield and Jasper, 1954). Work on the capacities for consciousness, by contrast, has often focuses on subcortical structures such as the midbrain and diencephalon. These are evolutionarily basal structures for vertebrates (Striedter, 2005; Merker, 2007). Further, the sorts of activities that the midbrain/diencephalon support are plausibly the sorts of activities that are conditions for consciousness.

Bjorn Merker, for example, suggests that the integration of self-motion, exteroception, and internal valuation (supplanted by memory) are the key to bringing together a first-person perspective on the world (Merker, 2005, 2007, 2013; Barron and Klein, 2016). Eva Jablonka similarly focuses on the role of integrative structures in supporting the capacity for unbounded associative learning (Ginsburg and Jablonka, 2007; Ginsburg and Eva, 2010). We argue elsewhere that the development of integrative structures combining sensory percepts into a unified neural representation of the mobile animal within its environment is a form of major transition in neural evolution, which enabled a fundamental shift in behavioural capacity (Barron and Klein, 2016).

Capacities themselves can also be intervened upon. An intervention on a capacity should have a broad and systematic effect across experience as a whole. So, for example, the same sensations can vary in how they feel given systematic alterations in broader underlying states. In the phenomenon known as pain asymbolia, patients with anterior insula damage will report that they continue to feel pain but no longer care about it (Schilder and Stengel, 1931). (Similar effects occur with a range of other dissociative drugs (Keats and Beecher, 1950).) There is debate about whether this effect is due to a sensory-limbic dissociation (Grahek, 2007) or to a general breakdown in processes of bodily ownership and concern (Klein, 2015), but in either case there appears to be alteration to the character of individual sensations by changes in the background conditions of experience.

Complex and subtle changes in the structural features of experience have also been reported. In a coda to *Awakenings,* Oliver Sacks (Sacks, 1999) describes the subjective distortions of space and time that coexist with degeneration of the substantia nigra, suggesting an interesting relationship between the basal ganglia, the general perception of space and time, and the specific motor impairments of Parkinsonian patients. Drugs such as dextromethorphan can also produce striking alterations in the perception of motion and time (Wolfe and Caravati, 1995).

Finally, there are ways to alter experience even more profoundly. Work on anesthetics, or on patients with severe impairments of consciousness also looks more directly at the basic capacities that support subjective experience (Alkire et al., 2008; Mashour and Alkire, 2013; Klein and Barron, 2016b). Of course, many unsubtle and uninteresting interventions can cause unconsciousness. But as several authors have recently urged, the capacities that underlie consciousness, and hence the broader modes of variation possible, are probably numerous and heterogenous (Bayne et al., 2016). There is no simple, well-ordered scale of 'degree' of awareness, for example; instead, there are numerous dimensions along which conscious experience as a whole might vary, and that plausibly link to the underlying functional capacities.

Intervention and an experimental approach to the Hard Problem

Focus on direct intervention is not new to neuroscientists. Nevertheless, we think it makes a substantial difference to the tractability of the hard problem. Individual types of experience and brain states are connected via *linking generalizations.* These have two important features. First, like traditional psychophysical bridge laws (Davidson, 1970), they connect physical and phenomenal states. Like all laws, they are counterfactual-supporting and change-relating, not simply descriptive. If an experience of blue corresponds to brain process B1 and an experience of red to B2, then changing from B1 to B2 should change the experience. Second, linking generalizations are a species of Woodward's invariant generalizations (Woodward, 2003). They need hold only under a limited range of circumstances and interventions (as is, in fact, the case for nearly all invariant generalizations, special science or otherwise).

The challenge posed by the hard problem is to explain why linking generalizations hold. This means that the hard problem is, at heart, a puzzle about scientific explanation (Levine, 1983; Irvine, 2013). Over the past two decades, philosophers of science have come to broad agreement about the centrality of interventions in scientific explanation.

This notion of interventionism has its roots in Judea Pearl's seminal work on causation (Pearl, 2000) and has been developed most notably by James Woodward(Woodward, 2000, 2003, 2010). The notion of a difference-maker is widely applicable, finding special traction in analyzing explanation in special sciences like genetics, neuroscience, and economics. Though originally conceived of as a way to explicate causal relationships, it has also found important use in analyzing synchronic relationships such as those between cognitive states and the neural states that realize them (Craver, 2007; Woodward, 2010; Klein, 2017).

The same is true for consciousness. Different interventions one might make on brain states have the capability to explain different aspects of experience. Systematic surveys of both the first- and third-person consequences of different types of intervention can thus address different aspects of consciousness.

Figure 1 shows a progression of hypothetical interventions we might make on experience. To begin we might (1a) have evidence that activity in certain brain region *B* is associated with seeing a red object. Crucially, the claim is not that *B* alone is sufficient to give rise to a red experience. Rather, difference-makers always act against a large background of causal factors. Similarly, many brain regions might make a difference to the same aspect of experience, and interventions on the same brain region might have many effects on experience.

Explanation is always *contrastive*: we explain why the world is one way rather than another. Different contrast classes can require different explanations (van Fraassen, 1980; Hitchcock, 1996; Woodward, 2003). The same property may receive different explanations when we consider different contrast classes. This is especially common when a property depends on combinatorial properties of its realizer. What makes a pixel white *rather than yellow* is not the same as what makes it white *rather than cyan*.

Thus, one explains why the subject sees red *rather than* some other color, given that they have one of several possible regional patterns of brain activation. Other contrasts might invoke other brain regions: seeing saturated versus desaturated red, say, might be influenced by something other than *B*.

The linking generalization in 1a says more than that pattern B_1 is reliably correlated with seeing red. As shown in 1b, a direct intervention on B that changed its activation from B_1 to B_2 (and left everything else the same, as much as possible) would make the subject experience blue in the presence of the same stimulus. Capacities could be intervened upon in the same way. Figure 1c shows a single intervention that warps and distorts the perception of space, creating correlated changes across a variety of experiences.

Some interventions will be relatively crude. Many interventions will simply eliminate consciousness altogether by eliminating a necessary background condition (Figure 1d); intervention on the claustrum appears to act as a kind of on-off switch for experience (Koubeissi et al., 2014). There may be many such interventions, and in general we are interested in interventions that give a more selective (Woodward Hopf and Bonci, 2010; Griffiths et al., 2015) or systematic (Klein, 2017) handle on phenomena we care about. These are interventions where there are many states of the control variable and many of the target, linked in a roughly one-to-one fashion, allowing for fine-grained control of the target.

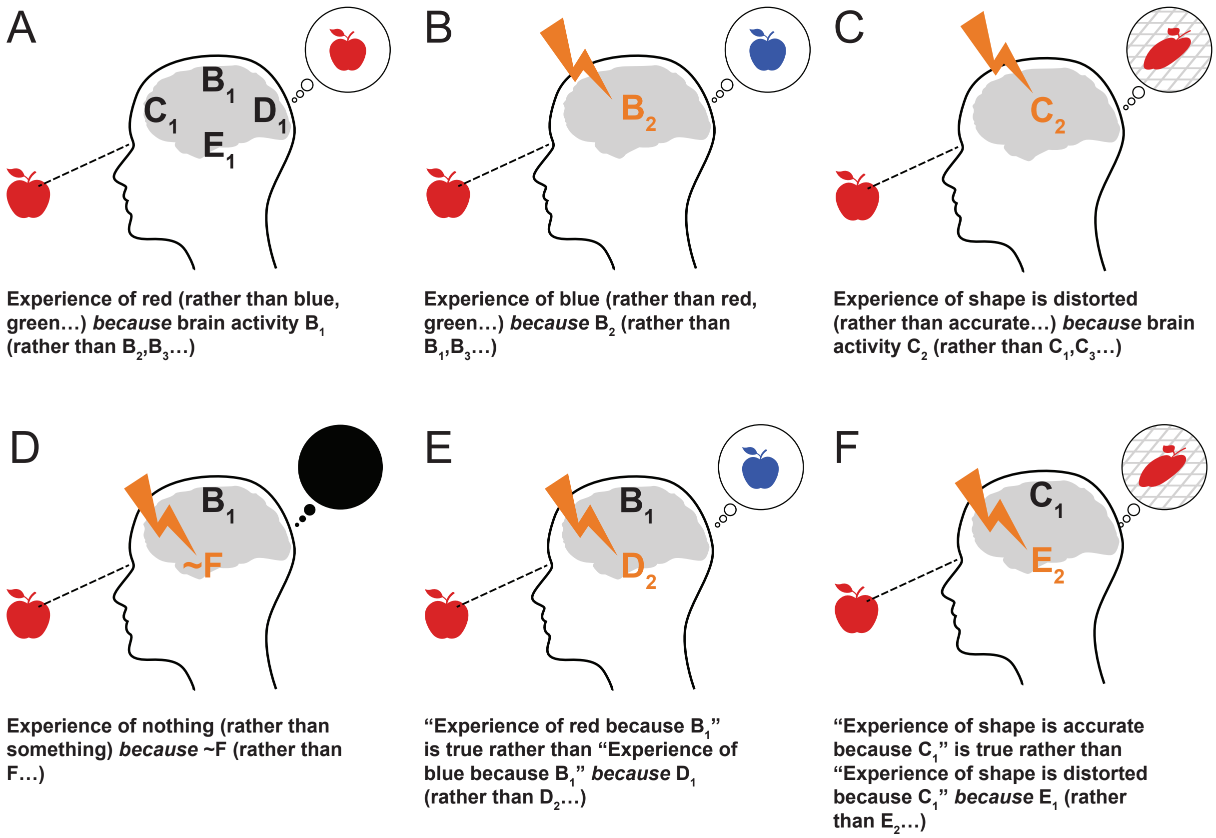

Figure 1: Different possible interventions of conscious states. (a) The simple case, explaining a token experience. B1 is a pattern of brain activation associated with seeing red, while C_1, D_1, E_1 etc are background conditions. The caption expresses a linking generalization. (b) An appropriate intervention on B will change the felt color given the same stimulus. This provides evidence for the claim in (a). (c) An intervention on a structural capacity. The metric of perceived space itself is distorted by an intervention of C, causing a variety of linked changes in experience of the stimulus. (d) A non-specific intervention on consciousness by eliminating a necessary condition. (e) An intervention on a linking generalization. D is part of what makes the laws in (a) and (b) true. Note that given the intervention on D, the same brain pattern B_1 that gives rise to a red experience in (a) gives rise to blue sensation. (f) A similar intervention on structural capacities. Cases (a) and (b) are entirely compatible with property dualism (SI 1), whereas the remaining cases would be problematic.

Figure 1d also illustrates another important point. The intervention eliminates consciousness even in the presence of activity B_1 that gave rise to a red experience in ordinary circumstances. Another way to understand this is that the linking generalization in 1a holds only contingently; appropriate intervention changes the law from holding to not holding.

This way of thinking about linking generalizations is only possible by moving to an interventionist picture of explanation. The hard problem, and the official statement of the NCC project, both rely on an older picture of scientific explanation. The positivists argued that explanation was the derivation of a proposition from a statement of a general law plus particulars (the so-called ‘Deductive-Nomological’ account); explanation thus showed why the state of the world was *sufficient* for the effect (Hempel, 1965). The official definition of an NCC is defined as the "minimal neural mechanisms jointly sufficient for any one conscious percept" (Koch et al., 2016; Boly et al., 2017 p9604) suggesting adherence to sufficiency as an explanatory principle. Much of Chalmers’ fame in philosophy is due to his revival of a contentious connection between *a priori* conceptual knowledge and metaphysical necessity (Chalmers, 1996; Soames, 2009). General, exceptionless laws are explicable only from even *more* general laws (Hempel, 1965). It is difficult to see what would fit the bill in the case of consciousness.

On the other hand, the deductive-nomological theory of explanation performs poorly outside of fundamental physics. The move to interventionism within philosophy of science was triggered in part by examples showing that sufficiency and explanatory purchase come apart (Salmon, 1989). Tophi are pathognomic for gout, and so sufficient evidence to deduce the presence of gout---yet they do not explain it. Having a genes for brown eyes explains why you have brown eyes, even though those genes, on their own, cannot actually make an eye (Sterelny and Kitcher, 1988). The demand for explanation of linking generalizations was thus virtually guaranteed to fail: not because physicalism is false, but because the story about explanation was not fit for purpose.

The deductive-nomological picture had other issues. In neuroscience, as in most disciplines, there are no exceptionless laws: laws are merely invariant across a range of circumstances. Hence the deductive-nomological account was particularly poorly suited for explanation in complex systems like the brain (Craver, 2007).

The potential variation of linking generalizations is the key to their explanation. Most work on difference-makers focuses on the relationship between event types. However, the same logic can be expanded to the explanation of invariant generalizations themselves. That is, we can explain some generalization *L* by showing variables that could be varied in order to vary *L*.

Figures 1e and 1f show interventions that change linking generalizations in more systematic ways. In 1e, intervention on D changes the relationship between B and color experience. Activity in D gives us an explanation of why the linking generalization in 1a holds, *rather than* some other linking generalization. Figure 1f shows a similar intervention on linking generalization regarding structural features of experience.

But of course, on the interventionist picture, what it *is* to explain a phenomenon is to demonstrate how it can be made to vary in replicable, systematic ways. Systematic intervention on linking generalizations themselves is thus the way to explain them. And that is what the hard problem demands.

The Meta-Problem and Self-Manipulation

Many balk at this point. Neuroscience that focuses on manipulation of structural features of awareness has the basic ingredients for tackling the hard problem of consciousness. Yet there is something about the hard problem that *feels* different than other problems. Chalmers has recently dubbed this the *meta-problem of consciousness* (Chalmers, 2018). A satisfying solution to the hard problem ought to explain why it seemed like there was a hard problem in the first place---why linking generalizations *seem* arbitrary and inexplicable, even if they aren't. Many otherwise promising accounts clearly fail to fit the bill.

We think this is a serious challenge. To begin, we note that the challenge has a perfectly objective answer. There is a view, tracing back at least to Leibniz, on which the apparent simplicity and arbitrariness of conscious states is merely an introspective confusion about a complex underlying state (Hilbert, 1987; Armstrong, 1997; Pettit, 2003). As Lashley famously put it: "No activity of mind is ever conscious...There is order and arrangement, but there is no experience of the creation of that order" (Lashley et al., 1960). The hard problem arises because we lack access to the relevant goings-on. There may be other sources of trouble as well, such as our relatively limited capacity for introspection and discussion of our conscious states compared to the richness of conscious experience itself (Block, 2011). Each of these mechanisms is a fact about us and how we are constituted, rather than a deep metaphysical feature of the world.

In short, our subjective experience is underpinned by a great number of mechanisms to which we have no conscious access, and which are not themselves represented in conscious experience. As we are aware only of the products of a complex mechanism and not its actual workings, we feel an arbitrariness of, and passivity towards, those products. The unconscious workings that give rise to conscious experience do not require effort of will and do not admit of first-person control. That is why conscious states feel arbitrary: subjectively, they simply appear out of nowhere.

Yet knowing all of this does not, by itself, make conscious experience feel any less arbitrary. *That* is the sense in which the hard problem is a unique scientific problem: just knowing the explanation does not remove the sense of mystery. Conversely, the remaining sense of mystery persists and undermines attempts to search for explanations, by making the problem seem harder than it is.

Nevertheless, this tangle should properly be seen as a problem with *us,* stemming from how we are constituted. We think that this problem can only be met head on: that is, via *self*-intervention. Interventions on brain states have both an objective and a subjective component. By intervening on brains, we don't simply discover *that* certain experiences can be evoked, or that they depend on certain interventions. The first-person, subjective experience of that intervention is critical as well.

This is not just proof of principle, though the proof of principle is important. (it is one thing to read about (say) the experience of alien hand, and quite another to *feel* your fingers jump around under the influence of TMS.) We believe that by *feeling* how subjective experience is altered by altering brain activity the impression of arbitrariness should vanish. As with the objective aspect, problem, the most telling alterations are likely to be specific, systematic alterations of the capacities that underlie conscious experience itself.

Thus self-manipulation of brain activity thus has the unique possibility not just to solve the hard problem, but to *fix* the passivity that leads to both the hard and the meta-problem. The point of interventions is to give us points of mastery over the world (Campbell, 2007, 2010). Self-mastery will be, and probably must be, the key to pushing past the lack of understanding that holds back effective research.

Making Progress

We recognize that this is an unusual solution. Self-manipulation should be seen as a necessary but possibly remote step in a long process. Nevertheless, to make progress, we will ultimately need safe, specific, selective techniques that allow us to intervene on experience in awake adults. Such interventions are difficult to come by in humans; at present they are either pharmacological, invasive, or non-invasive. We have noted some possibilities above, and we consider each in turn.

Pharmacological intervention is the most familiar and accessible way to intervene upon consciousness. Anesthetics remove consciousness altogether, and the specific ways and mechanisms by which consciousness breaks down already provides useful data about the capacities underlying experience.

There is an old tradition by which more specific interventions via psychedelic drugs have been thought to reveal interesting structural features about experience. There has been a recent revival of interest in psychedelics given their promising results in treating conditions like PTSD. That said, we think there is serious danger of repeating the mistakes of the past. In particular, we think it is worth being wary of returning to the uncritical pharmacological investigations that were popular in an earlier generation of research (Jay, 2009; Lattin, 2010). Some authors have been tempted to claim that the psychedelic experience itself is interesting precisely because it allows normally unconscious properties of the mind to be made manifest as objects of consciousness (Letheby, 2015). This is an old idea, embodied in the etymology of 'psychedelic' itself. We are skeptical. Despite decades of citizen science, we note few lasting contributions of such work to modern understanding of cognitive mechanisms.

Part of the problem is that psychedelics tend to have widespread and complex effects on consciousness. Less common drugs with more limited effects may be more useful. For example, reports suggest that low doses of diisopropyltryptamine (DiPT) have effects primarily limited to nonlinear distortions of audition (Shulgin, 2000). Limited and well-defined phenomena may also be fruitfully investigated, as for example in work done using LSD to investigate the central mechanisms of binocular rivalry (Carter and Pettigrew, 2003).

Invasive interventions involving direct electrical stimulation of the brain have been important to understand conscious function (Penfield and Rasmussen, 1950). Invasive work presents obvious ethical and practical concerns, and so is typically only done concurrent with some medical need. Much of the direct intervention work has focused on the effects of cortical stimulation on the contents of consciousness. However, there is increasing evidence that direct stimulation of the posterior cingulate/precuneus can produce more profound alterations in global experience (Herbet et al., 2014; Balestrini et al., 2015; Herbet et al., 2015). This would be consonant with these regions purported role in consciousness and mediating cortical-subcortical interactions (Vogt and Laureys, 2005; Cavanna and Trimble, 2006).

As for subcortical interventions, deep brain stimulation (DBS) has shown intriguing evidence of effects on consciousness. Much of this evidence takes the form of alleviation (Krack et al., 2010; Lyons, 2011) or induction (Bejjani et al., 1999) of psychiatric conditions such as obsessive-compulsive disorder and depression. Thalamic DBS has also led to promising improvements in minimally conscious patients (Schiff et al., 2007). The variety of possible stimulation parameters, and the variability of results between microstimulation and direct electrical stimulation (Vincent et al., 2016) suggests a fruitful experimental program in this area. We note that many case reports present no or only minimal data about a patients' subjective experience, even when this would clearly be accessible. We think that this ought to be more routinely and systematically collected.

Finally, noninvasive brain stimulation such as transcranial electrical stimulation (tES) may avoid the practical problems associated with invasive interventions. There have been initial indications that tES can improve responsiveness of patients in Minimally Conscious States (Thibaut et al., 2014). Perhaps the most interesting applications of tES, using either DC or AC current, is the possibility of entraining underlying circuits and thereby altering temporal dynamics of brain activation (Filmer et al., 2014; Tavakoli and Yun, 2017). tES has had problems showing specificity and replicability**,** but recent techniques using EEG/MEG to guide stimulation timing (Thut et al., 2017) may help ameliorate these concerns.

Most of the existing interventions we have discussed are still relatively broad and uncontrolled. The ability to make more systematic interventions would make subjective experiences seem less like passive and fleeting epiphenomena; they could be controlled, evoked, and altered at will. Ultimately, the requirement for specific interventions will demand developing new ways to intervene on the brain.

Invasive interventions also occur in research on brain-machine interfaces, though it is early days for this field. The current focus is on developing devices that can interact with neural circuits in such a way that they can become part of the system of information representation (sensuClark, 1995); the aim being to supplement or replace memory, or even add new information representations (Berger et al., 2011; Deadwyler et al., 2013; Deadwyler et al., 2017). We suspect that these particular forms of brain-machine interface will not help us much with the hard problem. Working out how information is represented in the brain remains an easy problem. On the other hand, if (say) an artificially implanted memory somehow *felt* distinctive, we might be able to learn more about the hard problem by using this as a contrast case.

Much of the work on developing new forms of brain-machine interface is currently happening with animals (Berger et al., 2011; Deadwyler et al., 2013; Deadwyler et al., 2017). This is the norm for experimental interventionist neuroscience. It is unethical to develop new methods on humans, but the reality of the deep homology of brain system functions across vertebrates (Striedter, 2005), and of neuron functions across most animal phyla (Kristan, 2016), means that methods developed in one species can usually be translated (with informed modifications) to another.

There is, however, a unique tension in using animal systems to study the nature of conscious experience. There remains a lively debate around which animals have any conscious experience at all, precisely because we don't know what neural circuits are necessary to support conscious experience (Klein and Barron, 2016b, a). Further, solving the subjective hard problem ultimately requires self-intervention, so animal models can only ever do part of the job.

That said, we envision research with animal models to play a key role for developing the interventionist tools, methods, and approaches needed for an experimental investigation of the hard problem in humans. Indeed, even very simple animals such as insects might provide a useful test-bed for developing more complex interventions (Barron and Klein, 2016; Klein and Barron, 2016b).

7. Conclusion: Fixing the hard problem

We have outlined an ambitious program for solving the hard problem. The hard problem of consciousness has two roots: a mistaken philosophy of science, and a deep (but not insuperable) limitation in our own ability to understand the roots of our experiences. Having identified these, neuroscientists must fix those shortcomings. This will require direct intervention, and a mix of third-person and first-person techniques.

Our proposal may strike authors from certain philosophical traditions as odd. Surely the hard problem was about explaining *consciousness*, not why the laws connecting brain to experience have one feature rather than another. Furthermore, contrastive explanation is by its nature pluralist: there will be not one grand explanation but many interlinking explanations. That might feel like something of a letdown.

Yet we suggest that the history of science provides numerous optimistic parallels. A closer look reveals that what initially appear to be grand, singular explanatory projects always end up dissolving into an array of specific, contrastive explanations as science advances. In the 18th century, there was a grand philosophical challenge to explain *Life* (Nassar, 2016). Considered as such, little progress could be made.

The advance of physiology in the 18th century did not attempt to explain life as a whole. Rather, it explained why this inorganic process could give rise to urea, why that process kept blood pH within reasonable limits, while that process cleared carbon dioxide rather than letting it accumulate, and so on (Bernard, 1865/1949). The march of progress ends up dissolving the original grand problem into an array of contrastive explanations, leaving even the project of *defining* 'Life' as a questionable one (Machery, 2012). We have not explained *Life* as it preoccupied the early modern philosophers. Instead, we can explain a great variety of things about living beings.

Similarly so, we envision, with consciousness. Successful interventionist research projects will alter and vary the relationship between brain activity and subjective experience. This will elucidate important mechanisms, and allow ever-finer control of experience. In the limit case, we will find consciousness just as grand, but no more mysterious, than life.

Finally, and crucially, we emphasize that this is a research program that is fundamentally falsifiable. That is, we might find that there are no systematic ways to intervene on linking generalizations: they are in fact like the brute laws of fundamental physics. Were the evidence to go that way, then non-physicalist theories of consciousness would gain plausibility.

Such a project obviously faces a host of practical problems. We do not pretend it will be easy. Many of the techniques and frameworks that will be required are only dimly understood at present. We can do a lot of science, and indeed will have to, before embarking on such a project. Rather, the claim is that until we reach this final step, the appearance of a hard problem will persist. Our discoveries about consciousness will always have a whiff of the arbitrary. The open question—why *this*?—will linger in the air.

Yet we think it is worth being optimistic. The idea that the hard problem might be a *practical* problem rather than a philosophical one has an unexpected pedigree. When Nagel argued that we do not know what it is like to be a bat, his point was not to argue against physicalism (Nagel, 1974, pp 447). Though often overlooked, Nagel closes his discussion with a positive proposal. Part of our difficulty in understanding consciousness, he says, is reliance on imagination when we try to take up the point of view of another subject. Imagination is an inherently limited faculty. Hence, Nagel tells us, his argument should be seen as "a challenge to form new concepts and devise a new method" of approaching experience (Nagel, 1974, pp 449). We agree. We just think that a more direct approach is required.

The hard problem will not be solved by philosophical discussion of positions relative to the problem alone, be they illusionist, dualist, physicalist or panpsychist. The difference between our proposal and the present philosophical impasse is akin to the difference between Freudian psychoanalysis and modern pharmaceutical approaches to mental illness. Increased understanding is important, and conceptual change is inevitable. But there is no 'talking cure' for the hard problem: some degree of direct intervention will be necessary.

Acknowledgements

This paper was funded by Australian Research Council Grant FT140100422 (to CK) and FT140100452 (to ABB).

References

Alkire MT, Hudetz AG, Tononi G (2008) Consciousness and anesthesia. Science 322:876-880.

Armstrong D (1997) What is consciousness? In: The nature of consciousness: Philosophical debates, pp 721-728: MIT press.

Balestrini S, Francione S, Mai R, Castana L, Casaceli G, Marino D, Provinciali L, Cardinale F, Tassi L (2015) Multimodal responses induced by cortical stimulation of the parietal lobe: a stereo-electroencephalography study. Brain 138:2596-2607.

Barron AB, Klein C (2016) What insects can tell us about the origins of consciousness. Proc Nat Acad Sci USA 113:4900-4908.

Bayne T, Hohwy J, Owen AM (2016) Are there levels of consciousness? Trends Cogn Sci 20:405-413.

Bejjani B-P, Damier P, Arnulf I, Thivard L, Bonnet A-M, Dormont D, Cornu P, Pidoux B, Samson Y, Agid Y (1999) Transient acute depression induced by high-frequency deep-brain stimulation. New England Journal of Medicine 340:1476-1480.

Berger TW, Hampson RE, Song D, Goonawardena A, Marmarelis VZ, Deadwyler SA (2011) A cortical neural prosthesis for restoring and enhancing memory. Journal of Neural Engineering 8:046017.

Bernard C (1865/1949) An introduction to the study of experimental medicine: Henry Schuman.

Block N (2011) Perceptual consciousness overflows cognitive access. Trends Cogn Sci 15:567-575.

Boly M, Massimini M, Tsuchiya N, Postle BR, Koch C, Tononi G (2017) Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. Journal of Neuroscience 37:9603-9613.

Campbell J (2007) An interventionist approach to causation in psychology. In: Causal Learning: Psychology, Philosophy and Computation, pp 58-66: Oxford: Oxford University Press.

Campbell J (2010) II--Control Variables and Mental Causation. In: Proceedings of the Aristotelian Society, pp 15-30.

Carter OL, Pettigrew JD (2003) A common oscillator for perceptual rivalries? Perception 32:295-305.

Cavanna AE, Trimble MR (2006) The precuneus: A review of its functional anatomy and behavioural correlates. Brain 129:564-583.

Chalmers D (1996) The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press.

Chalmers DJ (1998) On the Search for the Neural Correlate of Consciousness. Toward a Science of Consciousness II: The Second Tucson Discussions and Debates 2:219.

Chalmers DJ (2003) Consciousness and its place in nature. In: Blackwell guide to the philosophy of mind (Stich SP, Warfield TA, eds), pp 102-142: Blackwell.

Chalmers DJ (2018) The meta-problem of consciousness. Journal of Consciousness Studies 25:6-61.

Clark A (1995) Moving Minds: Situating Content in the Service of Real-Time Success. Philosophical Perspectives 9:89-104.

Craver CF (2007) Explaining the brain. New York: Oxford University Press.

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

Davidson D (1970) Mental Events. In: Essays on Actions and Events (Foster L, Swanson JW, eds), pp 207-224: Clarendon Press.

Deadwyler SA, Hampson RE, Song D, Opris I, Gerhardt GA, Marmarelis VZ, Berger TW (2017) A cognitive prosthesis for memory facilitation by closed-loop functional ensemble stimulation of hippocampal neurons in primate brain. Experimental Neurology 287:452-460.

Deadwyler SA, Berger TW, Sweatt AJ, Song D, Chan RHM, Opris I, Gerhardt GA, Marmarelis VZ, Hampson RE (2013) Donor/recipient enhancement of memory in rat hippocampus. Frontiers in Systems Neuroscience 1:120.

Dennett DC (1991) Consciousness Explained. Boston: Little, Brown, & co.

Feinberg TE, Mallatt JM (2016) The Ancient Origins of Consciousness: How the Brain Created Experience. Cambrige Mass.: MIT Press.

Filmer HL, Dux PE, Mattingley JB (2014) Applications of transcranial direct current stimulation for understanding brain function. Trends in neurosciences 37:742-753.

Ginsburg S, Jablonka E (2007) The transition to experiencing: the evolution of associative learning based on feelings. Biol Theor 2:231-243.

Ginsburg S, Eva J (2010) The evolution of associative learning A factor in the Cambrian explosion. J Theor Biol 266:11--20.

Ginsburg S, Jablonka E (2019) The Evolution of the Sensitive Soul : Learning and the Origins of Consciousness: MIT Press.

Grahek N (2007) Feeling pain and being in pain. Cambridge: MIT Press.

Griffiths PE, Pocheville A, Calcott B, Stotz K, Kim H, Knight R (2015) Measuring Causal Specificity. Philosophy of Science 82:529-555.

Hempel C (1965) Aspects of Scientific Explanation. New York: Free Press.

Herbet G, Lafargue G, Duffau H (2015) The dorsal cingulate cortex as a critical gateway in the network supporting conscious awareness. Brain 139:e23.

Herbet G, Lafargue G, de Champfleur NM, Moritz-Gasser S, le Bars E, Bonnetblanc F, Duffau H (2014) Disrupting posterior cingulate connectivity disconnects consciousness from the external environment. Neuropsychologia 56:239-244.

Hilbert DR (1987) Color and color perception: A study in anthropocentric realism. Stanford: Center for the Study of Language and Information.

Hitchcock CR (1996) The Role Of Contrast In Causal And Explanatory Claims. Synthese 107:395-419.

Irvine E (2013) Consciousness as a scientific concept: A philosophy of science perspective. New York: Springer.

Jay M (2009) The atmosphere of heaven: the 1799 nitrous oxide researches reconsidered. Notes and Records:rsnr20090006.

Kant I (1999) Critique of Pure Reason: The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

Keats AS, Beecher HK (1950) Pain relief with hypnotic doses of barbiturates and a hypothesis. Journal of Pharmacology and Experimental Therapeutics 100:1-13.

Klein C (2015) What pain asymbolia really shows. Mind 124:493-516.

Klein C (2017) Brain Regions as Difference-Makers. Philosophical Psychology 30:1-20.

Klein C, Barron AB (2016a) Insect consciousness: commitments, conflicts and consequences. Animal Sentience 9:21.

Klein C, Barron AB (2016b) Insects have the capacity for subjective experience. Animal Sentience:100.

Koch C, Massimini M, Boly M, Tononi G (2016) Neural correlates of consciousness: progress and problems. Naure Reviews Neuroscience 17:307-321.

Koubeissi MZ, Bartolomei F, Beltagy A, Picard F (2014) Electrical stimulation of a small brain area reversibly disrupts consciousness. Epilepsy & Behavior 37:32-35.

Krack P, Hariz MI, Baunez C, Guridi J, Obeso JA (2010) Deep brain stimulation: from neurology to psychiatry? Trends in neurosciences 33:474-484.

Kristan WB (2016) Early evolution of neurons. Curr Biol 26:R937–R980.

Lashley KS, Beach FA, Hebb DO, Morgan CT, Nissen HW (1960) The neuropsychology of Lashley: selected papers of KS Lashley.

Lattin D (2010) The Harvard psychedelic club: How Timothy Leary, Ram Dass, Huston Smith, and Andrew Weil killed the fifties and ushered in a new age for America: Harper Collins.

Letheby C (2015) The philosophy of psychedelic transformation. Journal of Consciousness Studies 22:170-193.

Levine J (1983) Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly 64:354-361.

Lyons MK (2011) Deep brain stimulation: current and future clinical applications. In: Mayo Clinic Proceedings, pp 662-672.

Machery E (2012) Why I stopped worrying about the definition of life... and why you should as well. Synthese 185:145-164.

Mashour GA, Alkire MT (2013) Consciousness, anesthesia, and the thalamocortical system. Anesthesiology 118:13-15.

Merker B (2005) The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. Consciousness and Cognition 14:89-114.

Merker B (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. Behavioral and Brain Sciences 30:63-81.

Merker B (2013) The efference cascade, consciousness, and its self: naturalizing the first person pivot of action control. Frontiers in Psychology 4:501.

Nagel T (1974) What is it like to be a bat? The Philosophical Review 83:435-450.

Nassar D (2016) Analogical reflection as a source for the science of life: Kant and the possibility of the biological sciences. Studies in History and Philosophy of Science Part A 58:57-66.

Odegaard B, Knight RT, Lau H (2017) Should a few null findings falsify prefrontal theories of conscious perception? Journal of Neuroscience 37:9593-9602.

Pearl J (2000) Causality: Models, Reasoning and Inference Cambridge university press.

Penfield W, Rasmussen T (1950) The cerebral cortex of man: a clinical study of localization of function. New York: Macmillan.

Penfield W, Jasper H (1954) Epilepsy and the functional anatomy of the brain. London: J.&A. Churchill, Ltd.

Pettit P (2003) Looks as powers. Philosophical Issues 13:221-252.

Sacks O (1999) Parkinsonian Space and Time. In: Awakenings, pp 339-349: Vintage Books.

Salmon W (1989) Four Decades of Scientific Explanation. Minneapolis: University of Minnesota Press.

Schiff ND, Giacino JT, Kalmar K, Victor J, Baker K, Gerber M, Fritz B, Eisenberg B, O□connor J, Kobylarz E, others (2007) Behavioural improvements with thalamic stimulation after severe traumatic brain injury. Nature 448:600.

Schilder P, Stengel E (1931) Asymbolia for pain. Archives of Neurology & Psychiatry 25:598-600.

Shulgin A (2000) Tryptamines I have known and loved. Berkeley: Mind Books.

Soames S (2009) Reference and description: The case against two-dimensionalism: Princeton University Press.

Sterelny K, Kitcher P (1988) The return of the gene. The Journal of Philosophy 85:339-361.

Striedter GF (2005) Principles of brain evolution. Sunderland, MA: Sinauer Associates.

Tavakoli AV, Yun K (2017) Transcranial alternating current stimulation (tACS) mechanisms and protocols. Frontiers in cellular neuroscience 11:214.

Thibaut A, Bruno M-A, Ledoux D, Demertzi A, Laureys S (2014) tDCS in patients with disorders of consciousness Sham-controlled randomized double-blind study. Neurology 82:1112-1118.

Thut G, Bergmann TO, Fröhlich F, Soekadar SR, Brittain J-S, Valero-Cabré A, Sack AT, Miniussi C, Antal A, Siebner HR, others (2017) Guiding transcranial brain stimulation by EEG/MEG to interact with ongoing brain activity and associated functions: a position paper. Clinical Neurophysiology 128:843-857.

van Fraassen BC (1980) The Scientific Image. New York: Oxford University Press.

Vincent M, Rossel O, Hayashibe M, Herbet G, Duffau H, Guiraud D, Bonnetblanc Fc (2016) The difference between electrical microstimulation and direct electrical stimulation-towards new opportunities for innovative functional brain mapping? Reviews in the Neurosciences 27:231-258.

Vogt BA, Laureys S (2005) Posterior Cingulate, Precuneal & Retrosplenial Cortices: Cytology & Components of the Neural Network Correlates of Consciousnes. Progress in Brain Research 150:205-217.

Wolfe TR, Caravati EM (1995) Massive dextromethorphan ingestion and abuse. The American journal of emergency medicine 13:174-176.

Woodward Hopf F, Bonci A (2010) Dnmt3a: addiction's molecular forget-me-not? Nature Neuroscience 13:1041-1043.

Woodward J (2000) Explanation and Invariance in the Special Sciences. British Journal for the Philosophy of Science 51:197-254.

Woodward J (2003) Making Things Happen. New York: Oxford University Press.

Woodward J (2010) Causation in biology: Stability, specificity, and the choice of levels of explanation. Biology & Philosophy 25:287-318.