First-person interventions and the meta-problem of consciousness

Colin Klein (The Australian National University)
Andrew Barron (Macquarie University)

1 Introduction

In a vivid thought experiment, Herbert Feigl (1958) imagined an 'autocerebroscope' which would allow us to view, in real time, the brain processes responsible for particular phenomenal states. Much of the subsequent debate over consciousness can be rephrased in terms of how satisfying the autocerebroscope would be, and what that answer would show.

Vivid visual evidence often makes scientific explanations compelling. Yet many people (if not Feigl) have the intuition that an autocerebroscope would still leave something of a mystery. It feels like there is something odd about consciousness that wouldn't be captured, even if we could see whatever we'd like about the processes that underlie it. Explaining these problem intuitions forms the core of what Chalmers (2018) calls the *meta-problem of consciousness*.

The meta-problem is distinct from the more traditional hard problem of consciousness. The hard problem (Chalmers, 1996) requires explaining how material stuff can give rise to phenomenal experience, given what seems like a deep metaphysical mismatch between the two. The meta-problem, by contrast, requires explaining why there seems like a deep metaphysical mismatch, regardless of whether there actually is one.

The 'meta-problem' is a handy device for bringing many disparate positions into conversation, and to provide further constraints on what a solution to the hard problem of consciousness could be like. We are realists about consciousness, like Chalmers. We're also materialists and naturalists, unlike Chalmers. Yet we agree that a good solution to the meta-problem will be key to getting to grips with the hard problem.

One way of emphasizing the difference is to note that the hard problem really has two aspects. There is a *third-person hard problem*: that requires figuring out the actual mechanisms which support subjective experience. Many scientists are engaged in that. Then there is a *first-person hard problem*: making the solution to the third-person hard problem feel convincing, given that we have problem intuitions that there isn't any such solution to be

found. Of course, we agree with Chalmers that the meta-problem intuitions depend in part on the mechanisms that support consciousness. So a solution to the third-person problem should also detail why we have the problem intuitions, which support the first-person hard problem.

In what follows, we offer a higher-level diagnosis of the meta-problem as depending on a certain feeling of *arbitrariness* about phenomenal features. This is meant to be a very general story, compatible with many particular answers to the meta-problem. It also emphasizes that, without doing something about the meta-problem, the first-person aspect of the hard problem will remain pressing, regardless of whether we actually hit on a good third-person explanation.

We then suggest that this arbitrariness is a matter not of lack of knowledge but of lack of *control*: we don't have the right sort of control over our phenomenal states. But that is a contingent matter. Suppose we upgraded the autocerebroscope so that we could both observe *and change* neural activity. We argue that this would be sufficient to eliminate the meta-problem. For having gained the right kind of control over our conscious states, we would no longer feel as if they were arbitrary. Further, we suggest, the removal of the meta-problem in this way would be intimately tied up with a solution to the hard problem.

Most of this is written from a naturalist point of view. Indeed, we intend this to outline both a position and a research programme. As such, we take materialism as a working hypothesis. It is a working hypothesis that might fail, however: if the programme sketched does not work, that will be valuable evidence against materialism.

2) The seeming arbitrariness of phenomenal properties

We prefer a view on which the apparent simplicity and arbitrariness of conscious states is due to limited access to a more complex state, along with a false belief that we have complete access (Hilbert 1987, Pettit 2003). This is often the case in the sciences. As Armstrong put it, it is part and parcel with the general scientific worldview that:

> …everything in the world, *everything*, every property of things and events, every relation that things and events have to each other, are each one of them an epistemological *iceberg*. Our knowledge and rational beliefs about all these things, though real, is selective and limited. (Armstrong 1999; 129)

This is no less true in the case of consciousness. Our access to the facts upon which consciousness depends is similarly restricted. As Lashley poetically put it:

> *No activity of mind is ever conscious*. This sounds like a paradox, but it is none the less true. There are order and arrangement, but there is no experience of the creation of that order (Lashley 1960; 532).

Because our awareness extends only to the contents of consciousness and not the complex mechanisms which support it, it is natural to *feel* that the hard problem is hard. That's where the meta-problem comes in, and what supports the first-person aspect of the hard problem.

There is an instructive parallel between this way of thinking of things and Hume's remarks in the *Treatise* on free will. As he puts it:

> There is a *false sensation* or *experience* even of the liberty of indifference; which is regarded as an argument for its real existence... We may imagine we feel a liberty within ourselves; but a spectator can commonly infer our actions from our motives and character; and even where he cannot, he concludes in general that he might, were he perfectly acquainted with every circumstance of our situation and temper, and the most secret springs of our complexion and disposition. Now this is the very essence of necessity... (§2.3.2)

In other words, we feel like we have unfettered free will because we lack of access to the causes of our own actions. Similarly, we think, with subjective experience.

Yet as Chalmers rightly notes (2018; 23), lack of access can only go so far as an explanation. There are many things in the world to which we lack complete access. Yet when science reveals their essential natures we feel satisfied, not puzzled.

We think that the core of the meta-problem is a certain feeling of *arbitrariness* about subjective experiences. Why should hearing middle C feel like *that*, rather than something else? In a vivid illustration, David Foster Wallace imagines a budding philosopher who

> ...is struck by the ghastly possibility that, e.g., what he sees as the color green and what other people call "the color green" may in fact not be the same color or experience at all: the fact that both he and someone else call Pebble Beach's fairways green an a stoplight's GO signal green appears to guarantee only that there is a similar consistency in their color experiences of fairways and GO lights, not that the actual subjective quality of those color experiences is the same; it could be that what [he] experiences as green everyone else actually experiences as blue, and

> what we "mean" by the word *blue* is what he "means" by *green*, etc...
> (Foster-Wallace 2001, fn23)

This thought experiment highlights that it doesn't seem very important *which* subjective experiences we have: it seems as they were consistent over an individual's life, phenomenal blue and green could be swapped with no effect. And in general, a whole family of thought experiments emphasize the degree to which it seems that phenomenal properties could be swapped, eliminated, or otherwise messed with, all with no deep consequence to how we get around in the world (Jackson 1982; Chalmers 1996).

Arbitrary relationships are anathema to scientific explanation. Good explanations explain why things are one way rather than another. If there is a materialist explanation of consciousness, it will explain why token brain experiences give rise to *this* sort of experience rather than *that*: that is, it will explain why that particular psychophysical generalization holds *rather than* some other generalization.

3) Explanation and intervention

This implies, note, that psychophysical relationships are contingent, and that the right kinds of intervention could make them otherwise. A brief note about this, as it may seem odd to some philosophers.[1] The past fifteen years of philosophy of science have emphasized the importance of direct intervention for explanation (Woodward 2003). Most scientific disciplines care about intervention, and the hunt for control variables is key (Campbell 2007).

Interventionism is a departure from earlier approaches, especially the deductive-nomological (DN) theory of explanation favored by the positivists (Salmon 1989). The DN theory emphasizes the importance of exceptionless covering laws for explanation. Descendants of the DN model crept into the philosophy of mind both directly (Klein 2013) and indirectly through discussions of psychophysical reduction (Klein 2009). They arguably linger on in the background assumptions of consciousness studies, particularly in the search for Neural Correlates of Consciousness.

Yet the weaknesses of the DN model of explanation are well-known (Salmon 1989). We won't recap them here. Instead, two contrasting features of the interventionist model will do

[1] For the full story, see Klein and Barron, (under review) .

some work for us. First, interventionism provides for explanation without appeal to exceptionless universal laws: the invariant generalizations it appeals to hold only over a certain range of conditions (Woodward 2003). Second, interventionism is fundamentally contrastive. One doesn't explain why X holds *tout court*, but only why X holds *rather than* not-X, or *rather than* {Y or Z or…}. Different contrast classes give different explanations (Van Fraassen 1980, Hitchcock 1996).

Interventionist explanations have mostly been studied in the context of explanations of event-types. In that capacity, there's probably little of interest to consciousness studies. Knowing that stimulating *this* part of the brain gives rise to the taste of a ham sandwich might be some evidence against the crudest forms of substance dualism. That has some rhetorical advantage, but all parties in the present debate can accommodate it.

However, while less emphasized, it seems clear that there should also be interventionist explanations of invariant generalizations themselves. Consider Woodward's (2003; 12-13**)** example of a block sliding down a ramp. Woodward presents the standard derivation as an explanation of the block's acceleration and how it counterfactually depends on friction. But one might equally well treat it as an explanation of the invariant relationship itself. Given the explanation, we can *also* show how this generalization would vary given changes in (say) wind resistance on the block, or if the block ceased to be a solid object and acted as a viscous liquid. That is, the standard derivation is not explanatory simply because it is a *derivation* (this is the lesson of attacks on the DN model). Rather, it is explanatory because it shows why the generalization is one way *rather than other ways it could be*.

This explanatory strategy is, note, a completely natural extension of the two interventionist principles above. First, generalizations are not exceptionless and universal, so it makes sense to ask why a generalization is one way rather than other ways they might be. (Conversely, fundamental physical laws are just brute facts; they can't be explained because they couldn't be otherwise.) Second, because explanation is fundamentally contrastive, we can equally well ask about contrasts between generalizations as we can between event-types.

We think this logic follows over to the explanation of consciousness. Consider some brain state *B* that's responsible for a particular experience *P* of pain. At a minimum, the

interventionist says, we should be able to change *P* in some way by changing *B*.[2] *But we should also be able to change the relationship* between *B and P*. That is, if the relationship between *B* and *P* is not just a brute law of nature – and the experimentalist is committed to the idea that it isn't – then it should also be a target of intervention.

Note, finally, that these are defeasible presumptions. It could turn out that there is no interesting way in which to manipulate the *B* to *P* relationship (that is, it can't be manipulated at all, or it can only be manipulated in a boring 'switch-like' way (Woodward 2010)). Again, if so, non-materialist stories would gain traction.

4 The Upgrade

Suppose we went through the explanatory project sketched in the previous section. Would that eliminate the feeling of arbitrariness about conscious experience? We think not. The psychophysical generalizations *could* be different, and we could even have good evidence and fully believe that. Yet for all we experience, they *don't* differ. Further, because we have no control over the relevant psychophysical generalizations, they will still seem brute and unexplained. For although the *B-to-P* relationship might be different, we cannot make it different. And without being able to *experience* how the *B-to-P* relationship can be varied, it will continue to seem as arbitrary as a law of nature.

So a solution to the third-person hard problem won't necessarily touch the meta-problem. Put another way, the meta-problem may not be something we can be *reasoned* out of, no matter how compelling the evidence is otherwise. Yet we have suggested that this is more of a limitation of our capabilities than of science. That could change.

To make the point vivid, imagine that we not only have an upgraded autocerebroscope, but that we've become skilled in its usage. That is, for any *B-to-P* relationship, we can alter it smoothly and seamlessly, in real time, and experience the results. We can imagine that we've become so fluent at this that the move from textbook description to altered experience is as smooth as the move from score to notes seems for a skilled pianist.

[2] Note that there might be multiple regions which affect *P*, and multiple ways in which *P* can be affected (Klein 2017). We elide this complication here, though we think it is an important feature of the contrastive aspect of explanation.

Were we to gain such control, we postulate, we would not be particularly impressed by the meta-problem of consciousness. Having ceased to be a bystander in our own experiences, they wouldn't seem arbitrary; neither would the relationship between brain and subjective experience

Present technology doesn't allow this kind of desirable tight coupling between specific brain intervention and specific first person experience, of course. The available methods for intervening on the brain tend to be crude, relatively slow, and relatively dangerous. There has been a recent revival of the idea that that experience with psychedelic drugs (for example) allows access to normally unconscious properties of the mind (Lethby 2015)**.** If true, that would be handy – but we are skeptical, and we think it's worth avoiding the well-documented mistakes of the past (Lattin 2010).

The upgraded autocerebroscope will remain a philosopher's fantasy. Yet the tight coupling envisioned is probably unnecessary: the ability to observe and change psychophysical links, even at a relatively coarse grain, might be all we need to tackle the meta-problem. For the key, note, is that self-intervention solves two problems at once. It solves a third-person explanatory problem, by showing the sort of interventions that affect consciousness. And it solves a first-person problem of understanding why the third-person interventions are not simply arbitrary, because it gives us control over them. Thus it would solve the meta-problem and the hard problem in one go—emphasizing, as Chalmers does, that the two problems are also intimately linked.

5) Conclusion

Where does that leave us with respect to the meta-problem of consciousness? A brief recap. We've argued that the core of the meta-problem is a feeling of arbitrariness about the mind-body relationship, born out of a certain lack of access to anything which would manipulate those grounds. This is (we hope) a characterization of the hard problem that is compatible with Chalmers' topic-neutral criterion (2018; 15ff). It is, at least, the sort of thing which many different theorists can subscribe to: the dualist thinks that the lack of access is a deep metaphysical feature of the world, the materialist realist can tell a variety of stories about why it might be, and the illusionist thinks that the seeming arbitrariness is part and parcel with the overall illusory features of consciousness itself.

We assume, again, that this story is compatible with a realist position about consciousness. We take ourselves to be realists about consciousness in the same sense that Bohr was a realist about atoms. That is, we think conscious states exist, but we're mostly wrong about their properties and maybe entirely wrong about their essential properties. The meta-problem stems from such an error. That's true both from the first-person and third-person point of view On some ways of cashing this out, this might count as a form of illusionism.[3] Yet we agree with Chalmers': it sounds *odd* to assert that we don't feel pain. One of the authors, on the other hand, has made a heap of true but nonobvious claims about the essential nature of pain (Klein 2015). So we're ok with surprises about the nature of the phenomenal.

We think that there is a useful analogy between the experimental program we suggest and other historical advances in science. The concept of *life* was once as fraught as that of consciousness. It seemed to many (including many philosophers) that there was a fundamental mismatch between the properties of living and nonliving matter, severe enough that the latter could never give rise to the former on its own. We now know that this impression was a mistake, due to inadequate concepts of both life and matter. Yet the history by which this mistake was corrected is primarily one of *experimentation* rather than mere observation or rational reflection. Wöhler's synthesis of urea and Bernard's dramatic manipulations of homeostatic mechanisms were important demonstrations precisely because they showed that what looked like brute facts about the world and its laws were in fact manipulable. Indeed, demonstrations like Wöhlers' were important not because they provided decisive evidence against vitalism (which lingered on for a long time afterwards) but because they gave evidence that the search for non-vitalist explanations might even be possible (Ramberg 2000).

The position we sketch also has some antecedents in analytic philosophy, usually in oblique discussions of the effect of psychedelic drugs (Langlitz 2016). The closest parallel might be with claims by Thomas Metzinger, for example that:

> … scientific research programs on consciousness and its neurofunctional correlates *could* be greatly optimized if researchers were well traveled in phenomenal state space, if they were cultivated in terms of the richness of their own inner experience as

[3] It's often surprising to non-philosophers that questions that appear to be about consciousness often turn on debates about meaning and reference (Block and Stalnaker 1999; Chalmers and Jackson 2001). Unsurprisingly, we are the sorts of naturalists who also like causal theories of reference.

> well. But not because this would give them a mysterious kind of first- person "data"—more likely, because it would thoroughly shatter their folk-phenomenological intuitions and endow them with completely new *theoretical* intuitions. What is right is that first-person approaches possess an enormous *heuristic* potential, and that we are currently far from realizing it. (Metzinger 2006; 2-3)

We think there is much to endorse in this. In particular, we think Metzinger is right to stress both the value of first-person experiences and to eschew the idea that the *content* of these experiences is of primary explanatory value.

Instead, we think that the primary value of first-person experience is best considered in terms of its effect on the meta-problem. Sometimes first-person demonstrations *that* something can be done are more powerful than *what* is actually accomplished: the first-person value of direct interventions, we suspect, will be most useful in that regard. Indeed, that usefulness might obtain even before we have a full science of consciousness sorted—for although the meta-problem and the hard problem are intertwined, it may take less to fix the former than it will to solve the latter.

Ultimately, we think Chalmers' framework should be of great interest to the experimentalist and the philosopher alike. So long as the meta-problem remains pressing, experimental research will seem unsatisfying. Conversely, experimental research might itself hold the key to fixing the meta-problem, and thereby making real progress on the science of consciousness.

Insert Thanks and Funding Info later

References:

Armstrong D (1999) *The Mind-body Problem: An Opinionated Introduction.* Westview Press: Boulder.

Block, N and Stalnaker, R. (1999) Conceptual analysis, dualism, and the explanatory gap. Philosophical Review 108(1): 1-46.

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

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

Chalmers, D. (2018) The Meta-Problem of Consciousness. *Journal of Consciousness Studies* 25(9-10): 6-61.

Chalmers, D. and Jackson, F. (2001) Conceptual analysis and reductive explanation. *Philosophical Review. 110(3): 315-360.*

Feigl, H. (1958) "The 'mental' and the 'physical'" *Minnesota studies in the philosophy of science*. 2(2): 370-497.
Hilbert DR (1987) *Color and color perception: A study in anthropocentric realism* (Center for the Study of Language and Information, Stanford).
Hitchcock CR (1996) The Role Of Contrast In Causal And Explanatory Claims. *Synthese* 107:395-419.
Hume D (2000) *A Treatise of Human Nature* (Oxford University Press, Oxford).
Jackson F (1982) Epiphenomenal qualia. *The Philosophical Quarterly* 32:127-136.
Klein, C. (2009) Reduction without Reductionism: A Defence of Nagel on Connectability. Philosophical Quarterly 59(234): 39-53.
Klein, C. (2013) Multiple realizability and the semantic view of theories. Philosophical Studies 163(3): 683-695.
Klein, C. (2015) *What the Body Commands: The Imperative Theory of Pain. Cambridge: MIT Press.*
Klein C. (2017) Brain Regions as Difference-Makers. *Philosophical Psychology* 30(1-2):1-20.
Klein, C and Barron, A. (draft ms) "How Experimental Neuroscientists Can Fix the Hard Problem of Consciousness"
Langlitz (2016) Is There A Place For Psychedelics In Philosophy? *Common Knowledge* 22:3: 373-384.
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(9-10):170-193.
Lashley KS (1960) The *neuropsychology of Lashley: Selected papers of KS Lashley*. ed F. A. Beach, D. O. Hebb, C. T. Morgan & H. W. Nissen. New York: McGraw-Hill**.**
Metzinger T (2006) "Reply to Hobson: Can There Be a First-Person Science of Consciousness?" *Psyche* 12, no. 4: 2.
Pettit P (2003) Looks as powers. *Philosophical Issues* 13(1):221-252.
Ramberg (2000) The death of vitalism and the birth of organic chemistry: Wohler's urea synthesis and the disciplinary identity of organic chemistry. *Ambix* 47(3): 170-195.
Salmon, W. (1989) *Four Decades of scientific explanation.* Minneapolis: University of Minneapolis press.
van Fraassen BC (1980) *The Scientific Image* (Oxford University Press, New York).
Wallace DF (2001) Tense Present Democracy, English, and the Wars over Usage. *Harper's Magazine*, April 2001.
Woodward J (2003) *Making Things Happen* (Oxford University Press, New York).
Woodward, J (2010) Causation in biology: Stability, specificity, and the choice of levels of explanation. Biology and Philosophy 25(3): 287-318.