# First-person interventions and the meta-problem of consciousness

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

**Abstract**

Chalmers' (2018) meta-problem of consciousness emphasizes unexpected common ground between otherwise incompatible positions. We argue that the materialist should welcome discussion of the meta-problem. We suggest that the core of the meta-problem is the seeming arbitrariness of subjective experience. This has an unexpected resolution when one moves to an interventionist account of scientific explanation: the same interventions that resolve the hard problem should also resolve the meta-problem.

## *1 The seeming arbitrariness of phenomenal properties*

In a vivid thought experiment, Herbert Feigl (1958) imagined an 'autocerebroscope' that 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 imply.

Vivid visual evidence often makes scientific explanations compelling. Yet many (if not Feigl) have the intuition that an autocerebroscope would still leave consciousness something of a mystery. Explaining the intuition that there would be such a residue forms the core of what Chalmers (2018) calls the *meta-problem of consciousness*.

We are realists about consciousness, like Chalmers. We are also materialists and naturalists (Barron and Klein 2016), unlike Chalmers. At a first pass, we think that the meta-problem arises due to limited access to complex brain states, along with a false belief that we have complete access (Lashley 1960; Hilbert 1987, Armstrong 1999; Pettit 2003). Because our awareness extends only to the contents of consciousness and not to the complex mechanisms that 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.

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. When science reveals their essential natures we feel satisfied, not puzzled. Consciousness is not like this.

In one sense, the lack-of-access model does predict recalcitrance of our intuitions. The core of the hard and meta- problems is a certain feeling of *arbitrariness* about subjective experiences. Why should hearing middle C feel like *that*, rather than something else? 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).

These thought experiments highlight that it doesn't seem very important *which* subjective experiences we have: were they consistent over an individual's life, phenomenal blue and green (say) could be swapped with no effect. That means that generalizations connecting one brain state to a feeling of green and another to a feeling of blue strike us as arbitrary: there's no reason (it seems) why they couldn't be otherwise.

Arbitrary relationships are anathema to scientific explanation. So even if we have a good solution to the hard problem, it won't *seem* like a good solution. That is, we think, the real bite of the meta-problem.

*2) Explanation and intervention*

The meta-problem, as we've sketched it, is deeply bound up in issues of scientific explanation. We think that the way out requires recognizing that psychophysical relationships are contingent, and that the right kinds of intervention could make them otherwise. This may seem odd to some philosophers, but it follows naturally from developments around scientific explanation.[1]

The past two decades 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). Even if we cannot

[1] For a more developed story, see Klein and Barron (ms).

perform interventions due to our contingent limitations, the sort of information that explains is precisely the sort of information that would permit intervention if we had sufficient power.

Interventionism is a departure from earlier approaches, especially the deductive-nomological (DN) theory of explanation favored by the positivists (Salmon 1989). The DN theory says that explanation comes from exceptionless covering laws. The DN model arguably lingers on in the background assumptions of consciousness studies, particularly in the search for Neural Correlates of Consciousness.

The weaknesses of the DN model of explanation are well-known (Salmon 1989). We won't recap them. Instead, we highlight two distinctive features of the interventionist model. 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 thus give different explanations (Hitchcock 1996).

Interventionism has been overlooked in consciousness studies, we suspect, because interventions have mostly been studied in the context of explanations of event-types. In the case of consciousness, this would consist of interventions that change some brain activity *B* to *B*,* thereby changing phenomenal state *P* to *P**. In that capacity, there is probably little of philosophical 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, but all parties in the current 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 of the block's acceleration as an explanation in terms of how acceleration depends on friction. But one might equally well treat that demonstration 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 another. (Conversely, fundamental physical laws might just be brute facts; they can't be explained because they couldn't be made 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* by changing *B*.[2] *But we should also be able to change the relationship between B and P.* That is, if the *B*-to-*P* relationship 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.

These are defeasible presumptions. It could turn out that there is no interesting way in which to manipulate the *B*-to-*P* relationship. Or it could be that the only interventions possible are crude ones that merely remove all consciousness. These would still be interventions on the *B-to-P* relationship, but only in the slightly degenerate sense that pulling the plug on a radio makes a difference to the relationship between the volume knob and the loudness of the music (Woodward 2010). If so, non-materialist stories would gain traction. Conversely, the more systematic and specific the control we gain over the *B*-to-*P* relationship, the more we can explain.

*3 The Upgrade*

So here's the first upshot: to solve the hard problem, we must learn to manipulate the *B-to-P* relationship. Yet that alone won't be quite enough to fix the meta-problem. So long as I cannot *experience* how the *B-to-P* relationship can be varied, it will continue to seem to me

[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.

that this relationship is as arbitrary as fundamental law of physics. We don't escape the vicious circle so easy; the meta-problem still has bite. Put another way, the hard problem may not be something we can be *reasoned* out of, no matter how compelling the evidence.

Yet so phrased, it is clear that the hard problem stems from limitations in our capabilities, rather than from the nature of consciousness itself. Those limitations could change. For this feeling of arbitrariness is a matter not of lack of knowledge but of lack of *control*: we don't have the right sort of influence over our own brains, and thereby over our phenomenal states. But that is a contingent matter.

Suppose we upgraded the autocerebroscope so that we could both observe *and change* our own neural activity. So for any *B-to-P* relationship, we could 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.

Were we to gain such control, we postulate, we would not be particularly impressed by the hard problem of consciousness. Having ceased to be a mere bystander in the production of our own experiences, they would cease to seem arbitrary. The psychophysical generalizations with which we work, having lost their arbitrary tinge, would appear to us as they are: as explanations of why and how consciousness arises from the brain.

Present technology doesn't allow this kind of 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. So the upgraded autocerebroscope will remain a philosopher's fantasy for the near future.

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 hard problem. For the key, note, is that self-intervention solves two problems at once. It solves an objective explanatory problem, by showing the sort of interventions that affect consciousness. And it solves the subjective problem of understanding why the objective interventions are not simply arbitrary, because it puts us in control. 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.

*4) Conclusion*

Where does that leave us with respect to both the hard and the meta-problem of consciousness? A brief recap. We've argued that the core of the hard problem is a feeling of arbitrariness about the mind-body relationship, born out of a certain lack of access to anything that would manipulate those grounds. The very same architectural features that give rise to hard problem also explain why we have the intuitions which constitute the meta-problem. 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 that 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. On some ways of cashing this out, this might sound like a form of illusionism.[3] Yet we agree with Chalmers that illusionism sounds odd---much more so than the prospect of 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.

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

We now know that this impression was a mistake. It stemmed from 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 how to manipulate what looked like brute facts. Indeed, demonstrations like Wöhler's 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 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.

*Acknowledgements*

Thanks to Tim Bayne, Jakob Hohwy, and Bjorn Merker for discussion, and to François Kammerer and an anonymous referee for helpful comments on a previous draft. This paper was funded by Australian Research Council Grant FT140100422 (to CK) and FT140100452 (to ABB).

**References:**

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

Barron, A.B. and Klein, C. (2016) What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences* 113(18): 4900-4908.

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.

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.

Jackson F (1982) Epiphenomenal qualia. *The Philosophical Quarterly* 32:127-136.

Klein, C. and Barron, A.B. (ms in review) How Experimental Neuroscientists Can Fix the Hard Problem of Consciousness.

Klein, C. (2017) Brain Regions as Difference-Makers. *Philosophical Psychology* 30(1-2): 1-20.

Langlitz (2016) Is There A Place For Psychedelics In Philosophy? *Common Knowledge* 22:3: 373-384.

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.

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.