Daniel C. Dennett

Welcome to Strong Illusionism

Abstract: *David Chalmers underestimates the possibility that actually answering the 'hard question' will make both the hard problem and the meta-problem of consciousness evaporate.*

David Chalmers' systematic and constructive survey of the available moves for solving the meta-problem makes many valuable contributions, including two that are unheralded:

(1) He demonstrates the importance of asking what I have called the 'hard question': 'And then what happens?' (Dennett, 1991, p. 272).
(2) He provides another example — his own — of the blind spot exhibited by the magicians who never solved the magic trick known as the Tuned Deck (Dennett, 2003).

Before turning to these contributions let me say at the outset that he has done a fine job laying out the attractions of what he calls *strong illusionism,* and I am happy to confirm that I am, and have always been, a card-carrying strong illusionist. He takes the view seriously — unlike, say, Strawson (2017; 2018), who can't bring himself to contemplate it — and candidly expresses his own temptation to leap across the abyss. I await his arrival with open arms. He has certainly obeyed the maxim to look before you leap, and his essay has — as he hoped — helped me appreciate the hidden strengths of the position I've been occupying all these years. I now consider him an ally.

We can bring Chalmers' first contribution into focus by juxtaposing two themes that he raises in several places: *acquaintance* and *robots* (casting the issue from the design stance; Chalmers, 2018, p. 20):

Correspondence:
Email: daniel.dennett@tufts.edu

Journal of Consciousness Studies, **26**, No. 9–10, 2019, pp. 48–58

> There are a couple of distinct elements to the sense of acquaintance. One is the sense of presentation: that we are somehow immediately presented with our experiences. Another is the sense of revelation: that the full nature of consciousness, and of various phenomenal properties, is fully revealed to us in introspection. (*ibid.*, p. 25)

> For example, if a computer system with both perceptual and introspective representations says that a green object is present, and one asks for its reasons, it might naturally answer that it is representing the presence of a green object. But if one asks for its reasons for saying that it is representing the presence of a green object, it may well have no further reasons. The system is simply in that state. It is not given access to the mechanisms that bring the state about. (*ibid.*, p. 24)

There is a looming but unacknowledged mismatch between these observations, brought out in another passage:

> Here we can think of a robot that visually senses the world around it, attends to certain objects, and has introspective representations of its own states. In fact the robot will stand in highly complex relations to objects and properties in the external world — a complex causal relation of seeing, an equally complex functional relation of visual representation, a complex functional relation of attending, and so on. *The robot may not have access* [my italics] to all that complexity, and there may be little need to model all the details. (*ibid.*, pp. 27–8)

Suddenly our robot has a personal level. It is not that its subpersonal states or systems have access to this or that; *it* (the whole robot or agent) has access or doesn't. Where does this fancy personal level come from and how is it constructed? That question is never even asked. Can a robot have a personal level? Does a self-driving car have a personal level? Show us, please, how this might be modelled. Is the robot *acquainted* with properties of some of its subpersonal states? Do those properties *present themselves*?

I submit that what has happened in the imagination of many theorists groping their way through the fog is that they have allowed unconscious subpersonal computational processes to generate representations that are then elevated, quick as a flash, to *personal level access*, but how that happens and what happens next is left unexamined. Instead of tackling that daunting problem of engineering, the theorist packs all the wonderfulness — all the whatever-it-is that generates problem intuitions — into the *presentation* while leaving the wonderfulness of the *consumer* of that presentation (as Millikan, 1989, would say) unanalysed. As Chalmers puts it, 'we are somehow immediately presented with our experiences' (2018, p. 25), but he doesn't pursue the details of how 'we' (somehow) are able to do this.

Chalmers (personal correspondence, commenting on an earlier draft of this essay) suggests that my problem might be called 'the access problem' and it is '*prima facie* distinct' from the hard problem and the meta-problem. Solving the access problem, he opines, is 'a reasonably straightforward research project'. All we have to do is explain or model the mechanisms involved in getting the robot to have 'states that are accessible for report, reflection, introspection, and the global control of action'. Easy? What makes Chalmers so sure that solving these easy problems won't uncover the mechanistic sources of the problem intuitions that define the meta-problem for him?

Consider *déja vu*, a ubiquitous but unsettling item of phenomenology with a relatively simple explanation (Dennett, 1979; 2003), which may not be confirmed, but which illustrates how 'phenomenal properties' in all their perplexity might be accounted for. Suppose that there is a 'familiarity detector' whose normal function is to tag current perceptual input as either novel or familiar; this would probably not be an anatomically distinct module or centre, but a general capacity of the neural networks that feed on that stream of representations. And suppose that for one reason or another the familiarity detector issues a false positive signal, which is then consumed by a wide and ever shifting variety of higher-level receivers, depending on the previous life history, sophistication, gullibility, etc. of the person. Among the many possible reports, reflections, or introspections that might be triggered by this false-positive input are 'This seems familiar, but…', 'Hang on, have I been here before?', 'Hmm. I just had a little *déja vu* episode; it's been a while since I've had one', or 'Oh wow, I think I've just recalled an experience I had in an earlier incarnation!' Any of these, in turn, would provoke further reflections, depending on the interests, roughly speaking, of the person. Now try to imagine an episode of *déja vu* with no such sequelae; the familiarity detector rings its bell, but nobody is listening. Would there be orphaned phenomenal properties, *there* but *unaccessed*? (Would a dollar bill that was the only thing in the universe still be worth a dollar?) Could a frog have a *déjà vu* experience, and what could the frog *do* with such a content?

I wish to challenge the conviction that the wonderfulness of *déja vu* is somehow packed into phenomenal — not functional — properties of some inner states instead of being embodied in the typical sequelae of some inner states. And I want to challenge the accompanying conviction, expressed by Chalmers, that modelling the machinery that could support the typical sequelae is just an easy 'access problem'. Some may be tempted to object that I've chosen a phenomenon that is

too easy; the 'phenomenology' of *déja vu* is in fact impoverished, so it can be 'explained away' in some such fashion. They may concede that *all there is* to *déja vu* is its functionally characterizable talent for generating the sorts of reactions illustrated above, but other phenomenal properties are more challenging. Are they sure? It won't do just to assert this as an unshakable intuition. Maybe this is just one of many such tricks the brain plays on us.

There is a pejorative term in computer modelling for models that don't explain how the representations get consumed: *vapourware.* Until you've specified how the information 'represented' is *used* by whatever receives it, you're just waving your hands and issuing a promissory note (as Ryle would say). And the fancier the representation, the bigger the loan. This comes out particularly vividly in Nicholas Humphrey's analogy of replacing the LED screen in the cockpit display of an aeroplane with a hologram (Humphrey, 2017). Postulating a representation with such awesome properties demands a representation-appreciator with the cognitive and aesthetic capacities to appreciate them. Feeding a hologram to the input registers of a decision-box would be like trying to impress a housefly with a serving of beef wellington. Humphrey, like other theorists, rightly wants to stress the staggering multidimensionality of the feedback loops that must exist in a conscious agent like us, but these theorists almost always concentrate on the 'input' side of the loops, exempting the receiver of the input from analysis. Humphrey makes it clear that the receiver in his analogy is no relatively simple subpersonal device, but the person:

> It only looks as if there's a third dimension. However, *you*, in the magical cockpit don't know this. To *you* it seems that the numbers really are jumping out of the screen. No wonder, then, that *you* find these sensory displays specially attention-grabbing and impressive. *You* do your best to explain to others, over the radio, just what it's like. But sadly, words often fail *you*. Still, it is your own first-person experience that *matters to you* above all. From now on *you* will go flying just to immerse yourself in these extraordinary displays. (*ibid.*, p. 5, my emphases)

It's a brilliant metaphorical description of what it's like to be a person, and Humphrey usefully draws attention to the fact that our reactions to, attitudes to, reflections upon… our experience are a major factor in *determining* the 'phenomenal content' of our experiences, but his catalogue is assembled entirely at the personal level, with no attempt to *analyse* how the knock-on effects described so vividly might be

implemented in subpersonal processes. Just how do 'words fail you' and how does 'mattering to you' get implemented, for instance? Here is where accounts of 'phenomenality' tend to stumble. If I take a pill that cures me of my dislike of cauliflower, does my experience of cauliflower thereafter have the same *phenomenal* properties (but now I don't mind them) or does not-minding-them count as itself a phenomenal property? Is the nauseating awfulness of pre-pill cauliflower a *quale* of mine, or merely a response to a quale (Dennett, 1988)? A recurrent embarrassment to philosophers is that they disagree about this fundamental issue, with the effect that there is no consensus at all about what qualia are, or about what phenomenal properties are, as Chalmers' scrupulous survey shows.

An aeroplane cockpit, like the remote control unit on a drone, has a limited number of 'buttons to push', each controlling a degree of freedom, and a limited number of 'displays', each providing information about the indirect and often distal effects of pushing the buttons. The dimensionality of the representations and the dimensionality of the representation consumers need to match, and this is the requirement that tends to get lost in the imaginings of theorists. As I have often said, all the work — and play — done by the homunculus in the Cartesian theatre (or by the pilot in the cockpit) must be distributed around in space and time in the brain, broken up and outsourced to lesser agencies that are themselves unconscious and uncomprehending.

Neuroscientists are just as guilty of shirking this obligation as philosophers, perhaps more so. Francis Crick and Christof Koch stopped half way into their model of consciousness:

> We have suggested that one of the functions of consciousness is to *present the result* [my italics] of various underlying computations and that this involves an attentional mechanism that temporarily binds the relevant neurons together by synchronizing their spikes in 40 hz oscillations.' (Crick and Koch, 1990, p. 272)

As I went on to scold:

> So a function of consciousness is to present the results of underlying computations — but to whom? The Queen? Crick and Koch do not go on to ask themselves the Hard Question: And then what happens? (Dennett, 1991, p. 272)

Chalmers' *meta-problem challenge* (2018, p. 36) goes most of the way to exposing this gap, especially in his appropriately brief critiques of Tononi's (2007) IIT and quantum theories (Hameroff and Penrose,

1996; Stapp, 1993). As he also notes, 'Global workspace and first-order representational theories can at least begin to answer the challenge' (Chalmers, 2018, p. 38).

I would add that the only paths I can see for answering the hard question invoke one version or another of a global workspace theory.

It may be that I deserve part of the blame for encouraging this abrupt truncation of theorizing when I distinguished the *personal* and *subpersonal* levels of explanation (Dennett, 1969). I had not meant the distinction to license an excuse, along the lines of 'we neuroscientists have done our job, by tracing the information subpersonally from the eyeballs up to the personal level; take it from here, you personal-level theorists!' But when neuroscientists and other cognitive scientists modestly proclaim that they are not trying to solve the hard problem, leaving that can of worms to others, they absolve themselves from addressing what might be called the *impedance match problem*: getting whatever 'outputs' are the end results of their models into the right dimensions or formats or other *functional features* necessary for consumption by whatever subpersonal mechanisms then take over. This is a generalization of a point that almost everyone[1] has accepted: no use putting a picture show in the brain if there isn't something with vision to watch and appreciate it.

By all means adopt the design stance, but then follow through, and at least sketch the design of something that can implement the imagined robot's personal-level competences. If you succeed, you will meet Chalmers' meta-problem challenge in the process, and if you fail, you may uncover, for the first time, the *functionally characterized* reason(s) why a robot made of robots made of robots could *not* be a philosophers' zombie after all: it couldn't duplicate all our personal-level competences because… That is to say, *if* there are insurmountable obstacles to providing a subpersonal model of all the competences and biases that must be present for an agent to be a 'zombie', a behavioural and cognitive duplicate of a conscious human being, this is the way to discover them. Those who are sceptical of materialism should embrace the hard question with zeal, in the same way creationists should embrace the challenge of demonstrating

[1] Yet even sophisticated theorists can lapse. Here is Christof Koch (2018): 'Indeed, the abiding mystery is how and why any highly organized piece of active matter gives rise to conscious sensation… What is it about the biophysics of a chunk of highly excitable brain matter that turns gray goo into the glorious surround sound and Technicolor that is the fabric of everyday experience?'

'irreducible complexity' in the design of organisms or their parts: it's the only honest way to prove that their guiding intuition is not just a failure of imagination.

To appreciate what I see to be Chalmers' second contribution, we first need to distinguish two different illusions: the malignant theorists' illusion and the benign user illusion. Chalmers almost does that. He asserts: 'To generate the hard problem of consciousness, all we need is the basic fact that there is something it is like to be us' (2018, p. 49). No, all we need is the fact that *we think* there is something it is like to be us. Dogs presumably do not think there is something it is like to be them, even if there is. It is not that a dog thinks there isn't anything it is like to be a dog; the dog is not a theorist at all, and hence does not suffer from the theorists' illusion. The hard problem and meta-problem are only problems for us humans, and mainly just for those of us humans who are particularly reflective. In other words, dogs aren't bothered *or botherable* by problem intuitions. Dogs — and, for that matter, clams and ticks and bacteria — do enjoy (or at any rate do not suffer from) a sort of user illusion: they are equipped to discriminate and track only some of the properties in their environment. The world, or *Umwelt* or original or manifest image, of an animal may or may not have colours, for instance. As Chalmers notes:

> It is common to observe (e.g. Chalmers, 2006) that vision presents colours as special qualities of objects that are irreducible to their physical properties. It is also common to observe that this is an illusion, and that objects do not really have those special colour qualities. Why, then, do we represent them that way? A natural suggestion is that it is useful to do so, to mark similarities and differences between objects in a particularly straightforward way. (2018, p. 25)

Curiously, for all the subtle layers in Chalmers' analysis, he never quite articulates the difference between the sort of benign illusions we and all creatures great and small enjoy and the illusions that might bedevil an autophenomenologist who goes beyond the uninterpreted catalogue of phenomena and tries to explain them with a theory. He gets very close, however:

> It is typically easy for people to accept that colours are illusions and are not really instantiated in the external world, but it is much harder for people to accept that phenomenal properties are illusions and are not really instantiated in our minds. (*ibid.*, p. 27)

Most of his essay is, of course, addressed to the latter issue, and he assays a treasure trove of different accounts of the aetiology of those problem intuitions. Which diagnosis-and-cure is *right*? He finds problems and shortcomings with all of them; none of them can handle *all* the problems. For instance, 'An obvious objection is that many people explicitly reject the sense-datum fallacy, but their problem intuitions remain as strong as ever' (*ibid.*, p. 30). Then he goes on to note that, unlike other philosophers, I have been frustratingly unwilling to settle on a single 'theory'.

> What is Dennett's account of problem intuitions? An overarching account is hard to find. Instead, he has appealed to a collection of ideas over the years, most of which are discussed elsewhere on this list. (*ibid.*, p. 31)

What does not occur to him is that there is an element of truth to *all* the diagnoses, and no one diagnosis handles all the reasons why people have been baffled by the brain's bag of tricks. That is what I have been saying since 2003: Chalmers has unwittingly fooled himself with a variation on Ralph Hull's masterpiece of card magic, the Tuned Deck.[2] Chalmers is just so sure in his heart that there is something supercalifragilisticexpialidocious about consciousness that he resists the conclusion, borne in on us by science, that we are robots made of robots made of robots… who manage, in concert, to create a user illusion of a Conscious Person, a single, unified agent, a self as a centre of narrative gravity. We outsiders need the user illusion to describe

> *others* (the robot *itself* that — who? — has access to one thing or another),

and to describe

> *each other* (*you* are *my* user illusion of your utterly complicated body and *I* am *your* user illusion of my equally complicated body)

and to describe *ourselves*.

---

2 Hull announced a 'new trick' which he called the Tuned Deck; in fact the whole trick was in the name; it was a collection of old tricks, and his audience of magicians were misled into trying to find a single explanation of all the variations he performed. For the details see Dennett (2003).

Just count all the first-person plural sentences, all the 'we's' in Chalmers' essay in addition to the standard theorists' 'we'. Here are some examples:

> …when ***we*** pick out a state indexically as 'this state', ***we*** are silent on its nature and there is no obvious reason why it should generate problem intuitions. (Chalmers, 2018, p. 22)
>
> When ***we*** recognize a cactus, ***we*** do not have problem intuitions anything like those ***we*** have in the phenomenal case. (*ibid.*, p. 22)
>
> Perhaps the best response to the belief problem appeals to an idea closely connected to the immediate knowledge idea: ***we*** have a sense that ***we*** are directly acquainted with conscious experiences and with their objects. (*ibid.*, p. 25)

The personal level is where we live, and where 'we' lives.[3] Part of meeting Chalmers' meta-problem challenge is *explaining,* subpersonally, how the user illusion of the personal level is created and maintained, answering the hard question, and we need the design stance for that. The idea that there can be a stable solution to the so-called hard problem couched in personal-level terms is a fantasy. We can, however, put first-person, phenomenological observations to good use in setting the task, doing phenomenology in its original sense, and then asking the hard questions, building subpersonal models of the personal-level phenomena we uncover.

Thanks to the design stance, we can sketch out roughly how this must go. Keith Frankish and I have embarked on a long-term thought-experimental exploration, inspired by Humphrey's (2017) discussion of replacing a pilot with an autopilot in the cockpit. We are looking at the task of taking a military drone, controlled remotely by a human pilot, and emancipating it — making it autonomous by uploading to the drone itself all the controls currently the responsibility of the pilot. Note that this exercise replaces the usual oversimplification with its inverse: whereas cognitive scientists tend to pile the 'content' into their postulated representations without specifying how this content gets discriminated by representation-users that might feed on them, we are starting with a known representation-user — a human pilot endowed with whatever knowledge and comprehension guides the uses made of the clearly described representations (video and audio of

[3] For more on the personal-level problem, and how to solve it, see Huebner and Dennett (2009).

such-and-such resolution, dial readings, labelled buttons, …) and trying to design simpler surrogate representation-users (and representations) that can accomplish the same control with less comprehension and no consciousness — just like the subpersonal representations and reactions that must underlie our own personal-level prowess. We are sketching out the specs, starting from the simplest modules needed, and then sketching the systems that would consume or use the signals from those base units. We are not engineers ourselves, but we have enough familiarity with the way engineers think about such projects to reason about the range of practical solutions. We want our robot drone to be able to tell us what it is like to be it, what bothers it, and what it notices and doesn't (or can't) notice. For this, it will have to notice its own noticing, and be able to do something with this noticing, if only to store it in memory or blurt out a report, for starters. It is going to have to have some way of getting from contentful states involved in responding and controlling its non-verbal activities to terms — verbs, nouns, adjectives, … — and combinations of terms that approximately capture or refer to or describe what is going on inside. This might be intelligently designed 'top-down' by a genius language-inventor who understands every level and feature of the robot's control system, or — much more practically and plausibly — it could be a system designed in the bottom-up, Darwinian way, by a competitive trial-and-error process that permits myopic, local pruning and focusing. For some opening moves in this research programme — answering the hard questions about how we are able to *say* (to others and to ourselves) what it is like to be us — see Jackendoff (1996), Carruthers (2009), Huebner and Dennett (2009).

I think Chalmers is ideally equipped and situated to make major contributions to this fledgling research effort, bringing both phenomenological and computational insights to the task at hand. If instead he stays with the hard problem, all he will ever create is vapourware. As Ralph Hull said, when he explained why his colleagues went off on a wild goose chase, 'the boys have all looked for something too hard' (Hilliard, 1938, p. 518).

Acknowledgments

I am indebted to David Chalmers, Keith Frankish, Nicholas Humphrey, François Kammerer, and Charles Rathkopf for their advice and criticism.

References

Carruthers, P. (2009) How we know our own minds: The relationship between mindreading and metacognition, *Behavioral and Brain Sciences*, **32**, pp. 121–182.

Chalmers, D.J. (2018) The meta-problem of consciousness, *Journal of Consciousness Studies*, **25** (9–10), pp. 6–61.

Crick, F. & Koch, C. (1990) Towards a neurobiological theory of consciousness, *Seminars in the Neurosciences*, **2**, pp. 263–275.

Dennett, D.C. (1969) *Content and Consciousness*, London: Routledge & Kegan Paul.

Dennett, D.C. (1979) On the absence of phenomenology, in Gustafson, D. & Tapscott, B. (eds.) *Body, Mind, and Method: Essays in Honor of Virgil C. Aldrich*, Dordrecht: Reidel.

Dennett, D.C. (1988) Quining qualia, in Marcel, A. & Bisiach, E. (eds.) *Consciousness in Modern Science*, pp. 42–77, Oxford: Oxford University Press.

Dennett, D.C. (1991) *Consciousness Explained*, Boston, MA: Little Brown.

Dennett, D.C. (2003) *Sweet Dreams: Obstacles to a Science of Consciousness*, Cambridge, MA: MIT Press.

Hameroff, S.R. & Penrose, R. (1996) Conscious events as orchestrated space-time selections, *Journal of Consciousness Studies*, **3** (1), pp. 36–53.

Hilliard, J.N. (1938) *Card Magic*, Minneapolis, MN: Carl W. Jones.

Huebner, B. & Dennett, D.C. (2009) Banishing 'I' and 'we' from accounts of metacognition, *Behavioral and Brain Scien*ces, **32**, pp. 148–149.

Humphrey, N. (2017) The invention of consciousness, *Topoi*, 9 August 2017, [Online], doi:10.1007/s11245-017-9498-0.

Jackendoff, R. (1996) How language helps us think, *Pragmatics and Cognition*, **4** (1), pp. 1–34.

Koch, C. (2018) What is consciousness?, *Scientific American*, June 2018, pp. 61–64.

Millikan, R. (1989) Biosemantics, *Journal of Philosophy*, **86** (6), pp. 281– 297.

Stapp, H. (1993) *Mind, Matter, and Quantum Mechanics*, Berlin: Springer Verlag.

Strawson, G. (2017) One hundred years of consciousness, *Isaiah Berlin Lecture*, Wolfson College, Oxford, 25 May.

Strawson, G. (2018) The consciousness deniers, *New York Review of Books*, [Online], https://www.nybooks.com/daily/2018/03/13/the-consciousness-deniers/ [14 March 2018].

Tononi, G. (2007) Integrated information theory, in Velmans, M. & Schneider, S. (eds.) *The Blackwell Companion to Consciousness*, Oxford: Blackwell.