From Sensor Variables to Phenomenal Facts

Wolfgang Schwarz*

Final version, 16 May 2019

Abstract. Some cognitive processes appear to have "phenomenal" properties that are directly revealed to the subject and not determined by physical properties. I suggest that the source of this appearance is the method by which our brain processes sensory information. The appearance is an illusion. Nonetheless, we are not mistaken when we judge that people sometimes feel pain.

1 Dualist intuitions

I will outline a physicalist explanation of certain intuitions that are often thought to cast doubt on physicalism. The intuitions I have in mind are more or less those Chalmers discusses in [Chalmers 2018]. My explanation will recapitulate ideas from [Schwarz 2018], but here I will present the case more directly, abstracting away from the Bayesian framework assumed in that paper. I begin by describing the relevant intuitions, since I want to highlight a few points that Chalmers passes over.

First of all, some physical processes seem to give rise to conscious experience. There is something it is like to see a red flower, or to burn one's fingers. Moreover, what it's like to see the flower is different from what it's like to burn one's fingers: the two processes seem to be associated with different "phenomenal properties".

Only a small range of physical processes seem to involve phenomenal properties. Paradigm examples are certain brain processes related to perception. Most other things that happen in the brain appear to happen in the dark, without any accompanying phenomenal properties. The same is true (intuitively) for processes outside the brain: the digestion of food in our intestines, the recording of a burglary on a security camera, or the flow of a river down a valley.

A second, more puzzling fact about phenomenal properties is that, to some extent, they seem to be directly and infallibly revealed to the subject of a relevant perceptual experience. For example, earlier today I was cycling through the rain. As I recall what that was like, I am certain that I am not right now having an experience with that phenomenal character. I may be hallucinating the desk in front of me, I may be a brain

* Thanks to Andy Clark, Justin d'Ambrosio, Frank Jackson, François Kammerer, Colin Klein, Alexander Sandgren, and Daniel Stoljar for comments on an earlier version, and to David Chalmers for helpful discussion.

in a vat, but I can conclusively rule out that my present experience has *those* phenomenal properties.

A third important feature of the association between phenomenal properties and physical processes is its apparent contingency. An experience of seeing red could have been associated with different phenomenal properties, or with none at all. This appearance of contingency gives rise to the anti-physicalist intuitions that are the focus of Chalmers's "meta-problem". For example, since phenomenal properties seem to be independent of physical properties, an agent could intuitively know all physical facts about colour perception without knowing what it is like to see red. Even the totality of all physical facts, it seems, is a priori compatible with the hypothesis that there is nothing it is like to see red. As a corollary, it is hard to see how physical facts could truly explain the existence and distribution of phenomenal properties: one could always accept the physical facts, it seems, and still wonder why the phenomenal side should be one way rather than another.

That's how things *seem* – to me, at least, and I suspect to many others. In short, perceptual processes (along with a few other things that happen in the brain) seem to have an irreducible phenomenal aspect that is directly revealed to the relevant subject. It is this seeming that I want to explain.

Chalmers identifies his target as a class of dispositions to think or assert sentences like 'consciousness is irreducible'. I think of my target as causally upstream of these dispositions. For example, it seems to me that any physical hypothesis about an agent is compatible with a total absence of phenomenal properties. I am here reporting a seeming in English, but the report is not identical to the seeming, nor is the seeming a brute disposition to make the report.

I mention this because I want my explanation to support a particular answer to the puzzle of consciousness, and I don't think that puzzle is primarily a puzzle about linguistic dispositions. I want to explain why certain physical processes appear to have an independent phenomenal character, not why some people are disposed to think or utter certain words.

Taking the intuitions I have described at face value leads to dualism. We would have to accept that some physical processes really are associated with special properties that aren't fixed by physical properties, except by epistemically contingent laws. Like many others, I worry how such properties fit into a naturalistic picture of the world. But I also think there is something deeply counter-intuitive about the dualist picture. Consider a world physically like ours but in which the phenomenology of seeing a red flower has traded places with that of burning one's fingers. In that world, people classify the sensation known to us as a kind of pain as neutral or even pleasant, and a popular opinion among philosophers is that it has intrinsic representational properties related to colours and flowers. Such a world seems conceivable. But it is strange – much stranger

than a world in which the laws of physics are different. Intuitively, the phenomenology of burning one's fingers somehow involves bad things happening to one's fingers; something has gone wrong if that phenomenology is associated with looking at a flower.

My intuitions here pull in opposite directions. On the one hand, the physical facts seem to radically underdetermine the phenomenal facts: physical information about an agent, it seems, can never conclusively rule out any hypothesis about what it's like to be that agent. On the other hand, the actual association between phenomenal and physical/functional properties does not seem arbitrary; unlike the laws of physics, it is not a brute empirical coincidence.

I will try to explain both of these conflicting intuitions. Since I don't think the source of the intuitions is essentially linguistic or conceptual, I don't think the explanation lies in special features of our phenomenal concepts or words, as some have suggested. Even a being without a (public or private) language could be puzzled about consciousness. What gives rise to the intuitions, I think, is something more elementary. It is the functional architecture of our perceptual system.

2 From stimulus to belief

It used to be easy. When a sensory stimulus arrived, our ancestors responded with a fixed pattern of behaviour. Same stimulus again, same behavioural response. That was some hundred million years ago. Since then, we have grown intermediate neural layers between sensory input and behavioural output – layers which allow us to store an internal representation of the world, detached from current sensory input. We can navigate around an obstacle even if we don't constantly perceive it. When a sensory stimulus arrives, our brain updates our internal model of the world and then chooses a behavioural response which makes sense in light of that model and our goals.

But now a problem arises that our distant ancestors didn't face. How should our world model be updated when a sensory stimulus arrives?

There are well-known methods for updating a model in response to new *information*. But here the input isn't information; it's a (proximate) stimulus: a certain pattern of activity at our neural periphery. To apply the well-known methods, we need to convert the stimulus into information.

It would help if every possible stimulus could only be caused in one relevant way – for example, if a certain pattern of photoreceptor activity could only be caused by a snake. Evolution could then have selected for systems whose internal world model represents the presence of a snake whenever that stimulus arrives. In fact, however, the correlation between receptor activity and relevant external (or internal) circumstances is far from perfect. An oddly shaped stick in the grass may produce the same receptor activity as a snake. A red surface under white light can cause the same activity as a white surface

under red light. Conversely, due to microscopic "noise", one and the same state of the macroscopic environment can lead to different patterns of receptor activity.

So suppose a stimulus S arrives that could have been caused by a variety of environmental conditions $E_1, E_2, E_3, \ldots$. How should a cognitive system update its world model, if the goal is to reliably achieve an accurate representation?

In philosophy, some have argued (in effect) that the system's world model should be sensitive not just to the proximal stimulus, but also to the distal causes of the stimulus: if S was caused by E_1, the system should represent E_1; if S was caused by E_2, the system should represent E_2. One problem with this proposal is that if E_1 and E_2 call for sufficiently different behavioural reactions, then such a system is physically impossible. Besides, the proposal has highly implausible consequences. Suppose an organism has placed a snake-like stick in a certain patch of grass, but worries that the stick has been removed, and so comes back to check. Before looking at the patch of grass, the organism is confident that the patch contains either a snake-like stick or nothing snake-like at all. When it then receives the kind of stimulus (S) that could be caused either by a snake (E_1) or by a snake-like stick (E_2), it ought to become confident that the patch contains a snake-like stick – even if someone in fact took away the stick and a snake happened to crawl in its place.

In what follows, I will assume without further argument that we need to find a method that takes a system's prior world model and a sensory stimulus as input and outputs a posterior world model, without direct access to what caused the stimulus. What should that method look like?

A simple but sub-optimal method would select one of the conditions ($E_1, E_2, \ldots$) that could have caused the stimulus – the most common one, perhaps – and update the system's internal model so that it assigns high probability (say) to that condition. Ideally, however, the update should take into account earlier information, as in the above example of the snake and the stick. Depending on the system's prior world model, the incoming stimulus S should sometimes be taken to carry information E_1, sometimes E_2, sometimes a certain probability of E_3, and so on. How could that work?

The standard solution in artificial intelligence is to expand the system's world model by extra variables corresponding to the sensory stimuli. A *variable*, in this context, is simply a part of the world model. Let's assume a system has variables representing things like distance to the next tree or inclination of the floor. (Variables typically have numbers as values.) To process sensory input, extra *sensor variables* are added to the model. These new variables are causally associated with incoming stimuli so that any relevant stimulus causes the sensor variables to take a certain value, independently of the rest of the world model. Since there are only probabilistic connections between sensor variables and other variables, and these probabilities can be adjusted, the same stimulus can lead to very different values of the ordinary variables depending on the circumstances

and the system's background beliefs.

Intuitively, the values of the sensor variables (which, again, might be numbers) represent the stimuli with which they are associated. The system starts out with certain views about which stimuli are likely to arrive given various configurations of the ordinary variables. Initially, the system might assume that E_1 is more likely to produce stimulus S than E_2. But this assumption can be revised. For example, after surprisingly receiving stimulus S when the ordinary variables were in such-and-such a configuration, the system may be less surprised to find S again when its ordinary variables are again in that configuration.

In principle, we wouldn't need to add a new dimension of sensor variables if a system already had a sufficiently rich world model. Suppose a system's ordinary variables allow it to make fine-grained distinctions about the electrochemical events taking place at its computational periphery. When a given stimulus arrives, we could then design the system to set its ordinary variables in such a way that they correctly describe the stimulus, in electrochemical terms. However, this method would be highly inefficient (as we will see). It also doesn't work for simpler systems like our not-so-distant ancestors whose ordinary variables weren't sufficiently rich to identify all possible stimuli by their physical or functional profile.

So a more general and more efficient strategy is to use designated sensor variables in the system's world model. These variables *represent* sensory stimuli in the sense that every suitable stimulus systematically causes the system to update its world model by a corresponding value of its sensor variables. But this causal association between sensor variables and electrochemical events need not be transparent to the system. If a certain stimulus involves the release of glutamate, say, and the system is unsure about whether glutamate is presently being released at its computational periphery, then merely receiving the stimulus will not resolve its uncertainty. Computationally, sensor variables are just further variables, logically independent of ordinary world variables.

From the system's own perspective, then, receiving a sensory stimulus will seem to make it certain of a special kind of fact that is only contingently associated with ordinary hypotheses about the world.

3 A bag of tricks

You can see where this is going. Evolution had to find a method for updating our internal world model in response to sensory stimuli. The method it stumbled upon, I suggest, is the method of sensor variables. That's why perceptual experiences appear to reveal to us a special kind of information – information about a primitive, non-physical aspect of the experience.

To flesh out this explanation, we need to look at another challenge our brain faces

when processing sensory input: the sheer volume of information. Consider a visual sense organ with 1000×1000 photoreceptors, each of which can distinguish 10 kinds of electromagnetic waves. There are 10^{10^6} possible activation patterns of the photoreceptors. It isn't practical to discretely store for each of these patterns, perhaps encoded as a 1000×1000 pixel matrix, how it affects any possible hypothesis about the external world.

Some tricks are needed to make the update of a system's world model tractable. An obvious first move is to pre-process and compress the sensory input. The pixel matrix could be re-encoded in terms of edges, movements, shapes, and colours, factoring out the influence of ambient luminance or movements of the system's own sense organs. Again, the task is well known in cognitive science. There are attractive computational models showing how a hierarchy of processing steps can efficiently convert raw sensory input of different kinds into a more useful representational format involving feature spaces for sounds, colours, orientation, and the like (perhaps influenced by attention and other top-down processes).

Plausibly, then, the "input" to our general, amodal world model does not consist of sensor variables tracking raw sensory activity. Rather, the amodal input is the output of perceptual pre-processing, with different perceptual systems using different feature spaces which help to extract useful information about the world.

The problem from the previous section now returns in a modified form: relevantly different states of the world can cause the same end result of visual pre-processing. So how should a prior world model be updated in response to a given pre-processing output (a point in a feature space, say)?

The answer, I suggest, is the same as before. When a system looks at a snake in a patch of grass, and its visual system generates a visual representation R, the system should not become certain that it is looking at a snake, irrespective of its background beliefs. The impact of R on the system's world model should depend on the prior state of the model, and even on the history of earlier experiences. So the system's world model should be extended by extra variables for perceptual input.

Further tricks are needed. Updating an entire world model in response to a given input is computationally taxing. If it looks like a rock is about to hit your head, the first priority is to duck; later you can figure out whether the object is really a rock, where it came from, and what it reveals about the geology of your environment. So if a certain representation R is typically caused by an environmental condition E_1, it could make sense to implement a fast and frugal process that infers salient E_1 aspects from R, checks if these call for immediate behavioural response, and then lets a slower background process figure out whether endorsing E_1 is really sensible in light of other information available to the system, and how other aspects of the system's world model should be revised in response to R.

Such a fast and frugal process puts further constraints on the format of perceptual

representation (the output of perceptual pre-processing): it should be easy to read off relevant features of the environment from a given representation. If the system's world model represents snakes and rocks in terms of certain variables, then it should be easy to connect the output of visual pre-processing to these variables.

Here it may help to "import" features from perceptual representations into the system's model of the environment. So far, I have assumed that a system's world model is neatly divided into "ordinary" variables, representing physical features of itself and its environment, and sensor variables, representing the pre-processed sensory input. The sensor variables provide the basis for updating the system's model of the external world, and that is all they do. The present suggestion is that a system could re-use the feature space of sensor variables in its model of the external world, to represent objects in its environment. For example, a system might represent objects in the environment as having colour properties that are defined not in physical or functional terms, but through their association with features in the space of visual representations. This would make it almost trivial to infer the relevant external-world variables from the output of perceptual pre-processing.

All these tricks, I suspect, are used in our nervous system. The success of multi-cellular organisms largely rests on their capacity to adequately update an internal world model in response to sensory input. Evolution won't have missed any opportunities.

If a system works the way I described, then its world model will have an extra dimension: an extra degree of freedom whose functional purpose is not to represent special properties in the environment but to allow for an efficient and context-sensitive update of the rest of the world model. In [Schwarz 2018], I called the extra dimension "imaginary" and the others "real", in analogy to the complex numbers, and to highlight the fact that points in the imaginary dimension are logically independent of any genuine hypothesis about the world as represented by the real part of the agent's world model. When a perceptual experience makes the system certain of some point or region in the imaginary dimension, flexible probabilistic links determine the update of the real part of the model.

None of that requires a language. But let's suppose some such system has developed a language, and it has started theorizing about its own perceptual experiences. It would be no surprise if the system were inclined to judge that there are special phenomenal facts to which perceptual experience provides direct access and which are only contingently related to ordinary physical and functional facts.

The system may also realize that it closely resembles certain other systems in its environment. If its world model represents its own experiences as having special, phenomenal properties, it would be natural to represent the experiences of other systems as also having these properties – properties that are revealed only to the relevant system at the relevant time. From the outside, such a thinker might conclude, we can never be sure whether other systems experience green and red the way we do, and we can never know

what it is like to be a bat.

To be clear, a system like that is not condemned to dualism. A system's world model need not match its considered judgements about metaphysical reality. If a system suspects that its world model has an extra dimension for the processing of sensory input, it may even write an article arguing that we should resist the temptation to postulate an extra dimension in metaphysical reality.

4 Is consciousness an illusion?

Perceptual experiences appear to have special properties, independent of all physical facts and directly revealed to the subject of the experience. But there are no such properties. The appearance is an illusion – an artefact of the way our brain processes sensory input. Does this mean that phenomenal consciousness itself is an illusion? Does it mean that no-one ever feels pain?

Imagine a community of completely physical agents in a completely physical world, whose cognitive system works the way I described. When these agents burn their fingers, they update their world model by a certain "imaginary proposition" – a certain setting of their amodal sensor variables. Over time, let's imagine, the agents have developed a language in which that setting of the sensor variables has become associated with a certain sentence Q, so that when an agent burns their fingers, they become disposed to utter and assent to Q. Some members of the community might hold that Q describes a basic dimension of metaphysical reality, but that is not built into the meaning of Q. On the other hand, Q also need not be equivalent to any physical or functional proposition, at least not in a transparent way: for any physical/functional proposition P, a competent speaker could coherently assert or entertain $P \wedge \neg Q$.

This is, roughly, how I think 'I am feeling pain' works in English. You are not committed to dualism by uttering that sentence, but the sentence also seems to convey more than straightforward physical or functional information.

Now suppose one of our imaginary agents, Bob, has burnt his fingers, updates his world model by the relevant imaginary proposition, and utters Q. Is his utterance true or false?

If pressed, I'd say it is true. Bob is in precisely the kind of state in which it is appropriate by the rules of his language to assert Q. And if it is appropriate in a certain context to assert Q, then it is also appropriate to assert that an utterance of Q in that context is true.

On the other hand, you might complain that an utterance is *genuinely true* only if (1) it represents the world as being a certain way, and (2) the world is that way. To assess whether Bob's utterance is "genuinely true", we would therefore need to know its representational content – how it represents the world as being.

The answer will depend on how we understand 'representational content'. In some sense, perhaps, Bob's utterance of Q represents certain patterns of electrochemical activity in his brain, due to suitable causal connections between these events and productions of Q. On that approach, Bob's utterance would again be true. Like Chalmers, I don't think brute causal conceptions of content can do much explanatory work. I prefer conceptions of content on which ideally rational and competent speakers can't be ignorant of the fact that two sentences of their language have the same content (in a given context). On such a conception, the content of Bob's utterance can't be any physical proposition, since even an ideally rational and competent Bob need not recognize $P \wedge \neg Q$ as a contradiction.

So I'm willing to admit that there is a strict and philosophical sense in which Bob's utterance is not "genuinely true", since it does not represent the physical world – which is all the world there is – as being a certain way. But the utterance also isn't "genuinely false", which would mean that (1) it represents the world as being a certain way, and (2) the world is not that way.

The upshot is that the situation is complicated. If we accept the explanation I have outlined, we certainly don't have to deny that people feel pain when they burn their fingers. We don't have to say that consciousness is an illusion. We don't have to revise our practice of expressing and attributing phenomenal properties. At most, we might have to say that in a certain strict and philosophical sense, attributions of phenomenal properties are neither true nor false.

References

David Chalmers [2018]: "The Meta-Problem of Consciousness". *Journal of Consciousness Studies*, 25(9–10): 6–61

Wolfgang Schwarz [2018]: "Imaginary Foundations". *Ergo*, 29: 764–789