Aporia vol. 21 no. 2–2011

Reductive Explanation and Qualia

Andrew Lee

Qualia are the intrinsic, subjective qualities of experience—the what-is-likeness, the raw feels. Reductive explanation is the functional explanation of a higher-level property in terms of lower-level properties. If qualia can be given a reductive explanation, then this would bridge the explanatory gap between the phenomenal and the physical. In this paper, I argue that no such reductive explanation could exist.[1] In Section I, I give an account of reductive explanation. In Section II, I argue that qualia cannot be reductively explained. In Section III, I consider what epistemological and ontological implications this has for qualia, emphasizing that failure of reductive explanation by itself establishes only an epistemological conclusion.

I.Reductive Explanation[2]

Roughly, a reductive explanation explains a higher-level property in terms of lower-level properties. I will not try to explain here what constitutes a higher-level or lower-level property; what is important here is that

[1]This paper is about epistemological issues regarding explanations of qualia. I try to remain neutral about metaphysical issues. However, I do make some observations about metaphysical issues at the end.

[2]What counts as reductive explanation is sometimes unclear. In this paper, I will only consider the functional analysis method of reductive explanation, and reductive explanation will refer to functional reductions. Other explanatory methods that might count as reductive explanations include bridge law connections and identity statements. However, I will count only functional reductions as reductive explanations, for three reasons. 1) Bridge law connections and identity statements are better regarded as reductions rather than reductive explanations, because they

Andrew Lee is a senior at Brown University pursuing degrees in philosophy, Cognitive Science, and a Masters in philosophy. His interests include philosophy of mind, particularly consciousness, representation, perception, mental causation, and personal identity. He plans on pursuing a PhD in philosophy after graduation.

reductive explanations shows how the instantiation of one property can result from the instantiation of other properties. A reductive explanation requires two steps. First, the higher-level property to be reductively explained must be given a functional analysis, or a definition in terms of its causal roles.[3] Such a definition has the following form: property G $=_{def}$ having lower-level property *F*, such that *F* fulfills causal role C.[4] C can be taken to be a specification of causal relations, such as "transmitting and encoding genetic information." Then if we can show how property *F* can transmit and encode genetic information, we have shown how *F* fulfills causal role C. Second, after establishing this functional definition, we can do the scientific, empirical work of constructing an account of how some lower-level property *L* fulfills C.[5] It should then be clear how it is that *L realizes* G. In other words, once *L* fulfills C, G comes for free—*L* fulfilling C necessitates the instantiation of G. Once we have this account, the reductive explanation is complete.

As an example, consider a reductive explanation of a gene. First, we must find a functional definition for the gene. We might come up with something like this: gene $=_{def}$ a mechanism that transmits and encodes genetic information. Then we must construct an account using lower-level property terms that satisfies this functional role. As it turns out, DNA molecules transmit and encode genetic information, and so can fulfill this functional role, giving a detailed molecular biological account of the details of the process. Once we have this, the reductive explanation of the gene in terms of DNA molecules is complete.

establish two properties that are at different levels but are coextensive. 2) Functional analyses allow for the multiple-realization of properties at the higher level, a chief reason that many have preferred reductive explanation to reduction in the first place. 3) Arguably, functional reductions have greater explanatory power than the other two methods, in that functional analyses can be applied to more phenomena in the natural world.

[3]Some, such as Block and Stalnaker (1999) hold that such functional definitions are almost never available for macroscopic phenomena. One way of seeing this is to consider a Twin Earth scenario. Even if XYZ satisfies the functional role of water, we would not say it is water, because it is an a posteriori truth that water is necessarily H2O. They argue that these a posteriori necessary identity statements such as water = H2O are required for reductive explanations of water, and that it may be a similar case for reductive explanations of pain. I will not consider this line of reply.

[4]In any kind of reductive explanation, the explanans must only use terms that are at a level lower than the terms of the explandum. Otherwise, it would not be a reductive explanation.

[5]Kim (2005) breaks the second step here into two separate steps; the first step involves identifying the physical realizers in the reductive base that can fulfill the relevant causal role, and the second step involves constructing the theory of how these realizers fulfill the causal role. I used only one step here, because it seems to me that the steps of identifying the physical realizers and constructing the theory are inevitably intertwined in the actual empirical process. Finding the physical realizers requires consideration of what the theory would look like, and constructing the theory involves looking for the physical realizers. In any case, nothing important hinges on this difference.

Note, however, that this sort of explanation is not a *reduction* of the gene to DNA molecules. Reductions are biconditional, such that a reduction of property G to property *F* shows that property G ↔ property *F*. In contrast, reductive explanations are conditional, such that a reductive explanation of property G in terms of property *F* shows that property *F* (instantiated in a certain way) →property G. Thus, the reductive explanation of the gene in terms of DNA molecules is consistent with genes being instantiated by something other than DNA molecules, as well as reductive explanations of genes in terms of things other than DNA molecules.

Before moving on to the next section, let us make some noteworthy observations about reductive explanations. First, reductive explanations are not purely a priori, but are largely empirical and scientific in nature. Successful completion of the second step requires empirical work. In order to establish that DNA molecules realize genes, that heat is molecular motion, and so on, we rely on scientific discoveries. Second, it seems that most natural phenomena in the universe could, in principle, be reductively explained in basic physical terms. After we explain the gene in terms of DNA molecules, there seems to be nothing stopping from going further down and explaining DNA molecules in terms of even lower-level properties. Of course, eventually we do hit a bottom level, at which point we can have no further reductive explanation. But it should not be surprising that we must take some things in the universe as basic. Third, and most importantly, reductive explanations remove the mystery of why certain phenomena occur. When we reductively explain the gene, we see that genetic mechanisms are not just brute facts that have no explanation. Rather, the function of DNA molecules necessitates the mechanisms of genes. This paints an attractive ontological picture. Through reductive explanation, we see that once the base-level properties are fixed, the higher-level properties automatically are instantiated.

II. Qualia

Let us now examine reductive explanation in the case of qualia. We may first consider what matters in an explanation of qualia. When we give reductive explanations of genes, heat, and so on, the explanations seem to leave nothing out. As long as we understand the relevant concepts, we see that it is inconceivable that there is a mechanism that transmits and encodes genetic information yet there is no gene, that molecular motion occurs yet there is no heat[6], and so on. There is nothing more to be explained in these cases. An adequate explanation of qualia should be likewise. If the explanation succeeds, then there should be no further

questions about qualia. An adequate explanation of qualia should explain not only relevant causal factors surrounding qualia, but qualia themselves. As Joseph Levine states, such a reductive explanation should have us saying something similar to the following:

> Suppose creature X satisfies functional (or physical) description F. I understand—from my functional (or physical) theory of consciousness—what it is about instantiating *F* that is responsible for its being a conscious experience. So how could X occupy a state with those very features and yet not be having a conscious experience? (Levine 1993)

Now, suppose we want a reductive explanation of pain. First, we need a functional analysis of pain. We might produce something as follows: "pain $=_{def}$ the property that is normally caused by tissue damage, tends to produce winces, groans, reports about pain, and so on." Call the right side of this definition *D*. We must then find a lower-level property that can fulfill *D*. Suppose that, as it turns out, C-fiber stimulation (Cfs) fulfills *D*, and we have a detailed account of how tissue damage causes Cfs, how winces, groans, and reports about pain are causally related to Cfs, and so on. Our reductive explanation of pain then shows us that Cfs is the realizer of pain (at least in humans), and our reductive explanation is complete.

If this is a successful reductive explanation, then it should remove the mystery surrounding pain. We should be able to look at how Cfs fulfills *D*, and see that the instantiation of pain necessarily follows. Like in the case of the gene, it should be inconceivable that Cfs occurs without pain. If the reductive explanation succeeds, we should not inquire further into pain. But I argue that such a reductive explanation of qualia fails. I will present two thought-experiments arguing that functional analyses cannot be applied to qualia.

First, suppose that we invite our Martian friend, Fred, over for tea. A pleasant conversation ensues about the meaning of certain human concepts, and Fred is especially curious about two concepts: "heat" and "pain." Fred does not know very much about what it is like to be a human, and so prefers having things explained in the most objective, scientific terms possible. We start with heat. Fred is told that what we call "heat" is the entity that causes metals to expand, melts ice cubes, and so on. He understands, and can even point out examples of heat on Mars.

[6]If it seems conceivable for there to be molecular motion without heat, then this is because of a conflation between heat and heat sensations. Molecular motion can occur without heat sensations, but not heat.

But we encounter a problem when we get to pain. Even after presenting Fred with *D*, our functional analysis of pain, Fred does not have the same concept of pain that we do. Fred does understand that pain is instantiated when a certain causal role is fulfilled, namely being caused by tissue damage, tending to produce winces, groans, reports about pain, and so on. But he seems to be missing something essential to our concept of pain—the subjective quality of pain sensation. The subjective quality of heat sensations is not essential to an understanding of heat, because "heat" refers to a certain property in the world—namely, molecular motion. But the case of pain is different—pain does not refer to the *cause* of our subjective sensations, but to the subjective sensations themselves. This shows that there is an asymmetry between the case of heat and the case of pain.[7]

It seems that no matter what sort of functional analysis we present to Fred, we cannot convey this essential aspect of pain. Further, it seems that there is no way for Fred to point out examples of pain, as he did for heat. He could point out examples of Martians exhibiting similar behavior, but perhaps the behavior is reflexive and there is no phenomenological sensation at all. Or perhaps Martians feel something, but it is quite different from the sensations of pain that humans feel. In any case, the functional analysis of pain fails, because any functional analysis leaves out the most essential part—the feeling of pain, or the subjective quality of it.

It may be objected that Fred does not fully understand our concept of heat either. After all, just as Fred does not know what it is like for humans to experience pain, he does not know what it is like for humans to experience heat. But this objection conflates how the two terms refer. We use heat sensations to determine what "heat" refers to, but heat sensations are not heat. Rather, heat produces heat sensations. We can imagine a world in which molecular motion exists, yet there are no beings to experience heat sensations. Even so, we would still say that heat exists in that world.[8] On the other hand, what we use to determine the referent of "pain" is pain.[9] Pain sensations are not *produced* by pain; pain sensations are pain. Fred understands what "heat" refers to in virtue of understanding the causal roles of heat, and our functional analysis of heat. But Fred does not understand what "pain" refers to in virtue of understanding *D*.

[7]On the other hand, there is no asymmetry between the case of heat sensations and the case of pain.

[8]See Kripke (1972) for a more in-depth discussion. I take Kripke's arguments about a posteriori necessities and essential properties for granted here.

[9]Cases of pain asymbolia do show that the affective and experiential aspects of pain can come apart. However, I will assume here that a proper analysis of "pain" would show that "pain" refers to the experiential aspect.

Since Fred does not understand what it is like to experience pain, he does not fully understand our concept of pain. And since there is nothing in the functional analysis of pain that could tell Fred what it is like to experience pain, the functional analysis fails.

The second argument against the functional analysis of qualia concerns the inverted spectrum thought-experiment. If qualia are to be reductively explained, then we should see from the reductive explanation why qualia have the character that they do. A reductive explanation of visual perception should explain why we experience the sensation of red (rather than the sensation of green) when seeing a tomato. But nothing in a functional analysis of visual perception could entail anything about what our visual experiences are like. This point can be contested; a functionalist might argue that in the future, we will have a more complete scientific language and we will be able to show how functional analyses entail subjective experience. But it is difficult to conceive what such a functional analysis could possibly look like. Right now, it seems clear that our best functional analyses of qualia do not entail subjective experience. If a functionalist argues that our future functional analyses will do better, then the burden of proof is on the functionalist to provide such an analysis (or even to provide an account of what such an analysis might look like.

Consider, then, inverted spectrum cases. We can imagine people that are functionally indistinguishable from humans, but who have different color experiences. Call these people inverts. When inverts see a tomato, they experience the color normal humans experience when they see a cucumber. And when inverts see a cucumber, they experience the color normal humans experience when they see a tomato. Inverts still call tomatoes "red" and cucumbers "green," but their inner experiences are different. Few people think that such inverts actually exist. But the problem here is that a functional analysis of qualia does not preclude their existence. That is, the existence of inverts produces no contradiction for the functional analysis. The intrinsic character of color experience is not something that can be defined through causal connections. Even if we have a complete characterization of a person in terms of causal connections, there seems to be no reason to regard that person as experiencing red sensations rather than green sensations. The core of the problem is that functional analyses do not entail anything about qualitative experiences.[10] But qualia *are* qualitative experiences, so functional analyses fail to reductively explain qualia.

It may be objected that the inverted spectrum case begs the question by presupposing that there is more to qualia than their causal roles. An adequate functional analysis of qualia would preclude anything like

inverts. If we have a functional definition such as experiencing red $=_{def}$ functional role *R*, then inverted spectrum cases cannot be possible because experiencing red just is the fulfillment of *R*. There is no fact of the matter over and above this, just as there is no matter of the fact over and above genes being the transmitters and encoders of genetic information. However, I argue this objection fails as well, because of a conflation between metaphysical and epistemic explanation. True, functional analyses of qualia may preclude the metaphysical possibility of inverts. But these functional analyses still fail to be explanatory in an epistemic sense. Suppose we build a robot that satisfies *R*. Is it then inconceivable that the robot experiences nothing? Or that it experiences a sensation different from the sensations humans experience when they see a tomato? Surely not; it seems that philosophical questions about the possibility of phenomenally conscious robots amount to more than just functional definitions. Our concept of "red" refers to an aspect of our experience, not to a functional role. Even if it is a metaphysical fact that experiencing red just *is* satisfying *R*, such a functional definition still fails to explain why this is so. Contrast this with the case of the gene. Our concept of "gene" is about a functional role—the role of transmitting and encoding genetic information. When we are told that genes are DNA molecules fulfilling role C, this is all the explanation we need. But when we are told that our experiences of red is the fulfillment of *R*, we may still legitimately ask why this is so. Epistemically, a functional analysis of qualia leaves much to be desired.

Another way of looking at all of this is to consider reductive explanation from the bottom up. Suppose we are given the relevant causal story about DNA molecules transmitting and encoding genetic information. Combine this with an understanding of the concept of "gene," and we would see that DNA molecules realize genes. But suppose we are given the analogous causal story about Cfs fulfilling the relevant causal roles regarding tissue damage, winces and groans, and so on. If we combine this with an understanding of the concept of "pain," we would not see that Cfs realizes pain. Our concept of "pain" refers to the subjective quality of pain. From the causal story of Cfs, we would not be able to infer that anything *hurts*. For all we know, the system that realizes this functional story feels nothing at all. This is why we find it inconceivable that DNA molecules fulfill functional role C yet there is no gene, but we find it

[10]A more vivid thought-experiment concerns zombies, who are functionally, behaviorally, and physiologically identical to us but have no conscious experience at all. The functional analysis of qualia is entirely consistent with everyone in the world being zombies—a functional analysis of pain says nothing about the phenomenological sensation of pain. I chose to use the inverted spectrum case here, because it is less extreme and also produces less complicated issues regarding how terms refer.

conceivable that Cfs fulfills functional role *D* yet there is no pain. There is an essential explanatory difference between the two phenomena.

We may conclude that functional analyses of qualia fail, unless the functional reductionist can resolve these problems. But there are reasons to think that these problems cannot, even in principle, be resolved. After all, how could a functional analysis possibly entail subjective states? What could such an account look like? It seems that this simply cannot be done. As we mentioned above, there seems to be a crucial difference between explaining genes, heat, and so on, and explaining qualia. Functional analyses always account for a property in terms of its causal relations. But it seems qualia have intrinsic qualities that prevent any causal definition. Further, this problem cannot be circumvented by tacking on "and produces sensations of type *X*" to the functional definition. Since a sensation of a quale is a quale, this add-on amounts to the same thing as "and produces quale *Q*," with *Q* replaced with whatever quale is being functionally analyzed. But to do this would be to include the definiendum in the definiens. It would result in functional definitions such as pain $=_{def}$ that which produces pain. This would violate the requirement that explanans include only terms at levels lower than the property of the explanandum. The explanans in such a functional definition of a quale would refer to the quale itself. This is clearly not satisfactory for a reductive explanation. As a result of all this, it seems that we can conclude that a functional analysis of qualia is not possible. But reductive explanations require functional analyses. Thus, qualia cannot be reductively explained.

III. The Explanatory Gap

If reductive explanation fails for qualia, then what does this entail for the explanatory gap? First, let us consider other possibilities for explanation. As noted in Section II, a successful explanation of qualia should explain why it is that qualia are they way they are. It should let us see why a lower-level property necessitates qualia, and it should leave us finding it inconceivable that qualia could not exist so long as certain states or functions are satisfied. We seek an explanation that shows that qualia are realized by the physical, that they fall out from the physical properties just like heat, genes, and so on. Reductive explanation seems to be just the right model for this sort of explanation. But as we have seen, reductive explanation, which works so well for all these other natural phenomena, seems to fail for qualia. Can we appeal to another kind of explanation? One response has been to use reductive identity statements, such as pain = Cfs. If this identity is really true, then clearly pain cannot exist without Cfs, since they are the same thing. However, any such reductions will face

the same problems that our functional reductions faced—even if they are true, they fail to explain much. In the case of other a posteriori identity statements, we are given a functional analysis that explains the identity. But as we have seen, such an analysis seems unavailable for pain and Cfs. Even if we have the identity pain = Cfs, this still fails to explain anything. This is a problem for any explanation of qualia in physical terms. It seems impossible to explain something subjective in physical terms, or to even conceive of what an adequate explanation would look like.

If the explanatory gap cannot be closed, then we must face the options that are left. Sometimes in other cases in which reductive explanation failed, we have had to take something as basic. For example, it was once thought that electricity could be reductively explained in terms of mechanics, but it turned out that a new law of nature had to be postulated, in the form of Maxwell's theory of electromagnetism. However, the problem seems to run even deeper than this. Postulating consciousness as arising from fundamental laws of nature removes little of the mystery of the phenomenon. Even if consciousness is basic, there is still no explanation of why it seems to arise only in certain functionally complex systems. Along with this come all the problems that face epiphenomenalism and interactionism, depending on the type of dualism that one wishes to take. So, one option from the failure of reductive explanation is to conclude that consciousness is not physical, but something basic in the world. But this still leaves much to be explained, and consciousness faces serious conceptual challenges that other fundamental laws like electromagnetism and gravity did not face.

However, making the metaphysical leap is not the only option. Importantly, the direct consequences of the explanatory gap are epistemic, not metaphysical. And just as there seems to be a gap between physical properties and qualia, there is a gap between epistemic consequences and metaphysical consequences, or between conceivability and possibility. Even if it is conceivable that Cfs exists without pain, it does not follow that this is a genuine metaphysical possibility. The physicalist is in a hard-pressed position to explain why it is not a metaphysical possibility that Cfs could exist without pain. But it is not incoherent for the physicalist to take such a position, even if the arguments here are accepted. The explanatory gap is certainly a deep problem for physicalism, but it does not entail that physicalism is false. The explanatory gap is an epistemic problem, but physicalism is a metaphysical thesis. We cannot jump straight from an epistemic problem to a metaphysical conclusion. Other theoretical considerations must be taken into account.

Works Cited

Block, Ned, and Robert Stalnaker. "Conceptual Analysis, Dualism, and the Explanatory Gap," *Philosophical Review* 108 (1999): 1-46.

Chalmers, David. *The Conscious Mind* (New York: Oxford UP, 1996).

Kim, Jaegwon. *Philosophy of Mind*. Boulder, Colo.: Westview, 2006. Print.

——. *Physicalism, or Something near Enough*. Princeton, N.J.: Princeton UP, 2005. Print.

Kripke, Saul. *Naming and Necessity* (Cambridge, MA: Harvard UP, 1980).

Levine, Joseph. "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354-361.

——. "On Leaving Out What It's Like," in *Consciousness*, ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell, 1993).

Stoljar, Daniel. "Physicalism (Stanford Encyclopedia of Philosophy)." *Stanford Encyclopedia of Philosophy*. 9 Sept. 2009. Web.<http://plato.stanford.edu/entries/physicalism/>.