Materialism and the Psychology of Explanation

Brian Fiala
Philosophy Program
The University of Arizona
fiala at email dot arizona dot edu

Materialism is a thesis that all the facts about consciousness are determined in virtue of physical facts. According to materialism, once we have given a complete physical story about the world, there is no further story about consciousness to be told: consciousness itself can be explained in physical terms. However, it is plausible to think that even if we had a complete physical explanation of consciousness in hand, an air of mystery would still remain. Even if a materialist theory yielded perfect correlations between neural states and phenomenal states, we could still ask the question, “Why is that neural state correlated with that phenomenal state—or any phenomenal state at all, for that matter?” In other words, there would be an “explanatory gap.” The explanatory gap poses a prima facie problem for materialist theories of consciousness, insofar as the existence of an explanatory gap prima facie suggests the existence of an ontological gap. Hence, the materialist must either deny that there is a gap, or find a way of rendering the gap consistent with materialism.

In what follows I will put forth a new strategy for rendering the explanatory gap consistent with materialism. Whereas extant strategies emphasize the unique characteristics of our concepts of phenomenal consciousness, my proposed strategy will focus on the unique cognitive profile of explanation. The hope is that we can account for the explanatory gap as a by-product of the relatively less mysterious psychological features of the explanation, thus de-mystifying the gap.

I will proceed as follows. In §1, I introduce the explanatory gap and its relation to materialism. I also introduce the standard phenomenal concept strategy in §1, in order to contrast it with the novel strategy under consideration. In §2, I give a brief tour of the psychology of explanation. In §3, I sketch the new strategy, and point out some of its current limitations. Finally, I conclude by discussing some implications for future research.

1This paper is a work in progress, drafted in March 2007.
2Here I use the term “explanatory gap” in the way introduced by Levine in his (1983).
1 Background from Philosophy

1.1 The Explanatory Gap

The explanatory gap is, roughly, the puzzlement that remains once we’ve drawn psycho-physical identities like “pain is nociceptor firing.” After such an identity has been drawn, it’s still easy to wonder “why is nociceptor firing pain, as opposed to some other feeling (or no feeling at all)?” In particular, such questions arise when we draw identities between phenomenal states (or properties) and physical states (or properties).\(^3\) Even if the identity relations hold, it’s hard to see why they must hold. That is to say, even if the identities hold, the identities themselves call for explanation.\(^4\)

Thus, the explanatory gap is an epistemic gap. Although some ultra-hard-core materialists deny that there is any epistemic gap, for the purposes of this paper we will suppose that there is one.

Does the explanatory gap preclude materialism? This depends on whether there can be an epistemic gap without a corresponding ontological gap. Dualists take the explanatory gap as an indicator of an ontological gap.\(^5\) On their view, the explanatory gap shows that consciousness can’t be naturalized, because ontological reduction demands explanatory reduction. According to the dualist, we would be unjustified in in positing an ontological identity without positing a corresponding explanatory connection.

Even some materialists hold that the existence of an explanatory gap entails that consciousness can’t be naturalized. McGinn, for example, holds that consciousness will forever remain a mystery.\(^6\) He argues that we can’t naturalize consciousness without bridging the explanatory gap, and further argues that we can’t bridge the gap. Bridging the gap would require the acquisition of radically different concepts. We could acquire these concepts in principle, but as matter of fact

---

\(^3\)Psycho-physical identity statements involving other sorts of mental states may be considerably less problematic.

\(^4\)Although I characterize the explanatory gap in terms of identity relations, similar gaps arise from other explanatory relations. Suppose, for example, that a materialist posits law-like supervenience relations relation between physical states/properties and phenomenal states/properties. In this case, we would be left wondering why that physical state/property sub-venes on that phenomenal state/property, rather than sub-vening on some other phenomenal state, or no phenomenal state at all.

\(^5\)See Chalmers (1996), for example.

\(^6\)McGinn (1989)
we are unable to acquire them due to limitations on human cognitive architecture. Despite all this, McGinn is unable to give up materialism. Hence his view: ‘mysterianism.’ Mysterians agree with the dualists that the mind-body connection is unexplained. It’s just that mysterians are so taken with materialism that they are willing to postulate the unjustified (mysterious) psycho-physical identities, in spite of the explanatory gap.

There is a third sort of position that sees the explanatory gap as a genuine barrier to materialism, but as a barrier that we may be able to surmount. Stoljar adopts a position in this vicinity.⁷ The position is similar to mysterianism, but is ultimately more optimistic. Although there is an explanatory gap now, we may be able to dissolve the gap in the future by developing radically new concepts: we’ll just have to wait and see if we’re able to develop the appropriate concepts.

Levine, who introduced the “explanatory gap” nomenclature into the recent literature, agrees with McGinn that the existence of an explanatory gap does not entail the existence of an ontological gap.⁸ Levine further emphasizes that even if there is an epistemic gap that is unbridgeable in principle, it still doesn’t follow that phenomenal states are not identical to physical states. In other words, pain might be identical to nociceptor firing even if this fact leaves us bewildered and craving further explanation. Many materialists capitalize on this failure of entailment to defend their thesis: they acknowledge the epistemic gap, but deny that there is any ontological gap. For the remainder of the paper, we will pursue this sort of response to the explanatory gap. That is, we will pursue options for treating the explanatory gap as a merely epistemic gap. The goal of this strategy is to explain the origin of the epistemic gap in a way that is materialistically kosher— that is, in a way that does not entail the existence of an ontological gap.

⁷Stoljar (2001)
⁸Levine (1983)
1.2 The Phenomenal Concept Strategy

The phenomenal concept strategy attempts to account for the explanatory gap in terms of human psychology. The idea is that once we have the appropriate psychological story in hand, the gap will no longer seem mysterious. That is, the hope is to ‘psychologize away’ the gap. Chalmers describes the general structure of the strategy in terms of the following theoretical schema:

The general structure of the phenomenal concept strategy can be represented as follows. Proponents put forward a thesis $C$ attributing certain psychological features—call these the key features—to human beings. They argue (1) that $C$ is true: humans actually have the key features; (2) that $C$ explains our epistemic situation with regard to consciousness: $C$ explains why we are confronted with the relevant distinctive epistemic gaps; and (3) that $C$ itself can be explained in physical terms: one can (at least in principle) give a materialistically acceptable explanation of how it is that humans have the key features.\(^9\)

Note that in Chalmers’s formulation, there is no mention of concepts. Hence the label “phenomenal concept strategy” is something of a misnomer. That said, the great majority of proponents of the phenomenal concept strategy do hold a thesis about the concepts we possess. That is, they tell a story about concepts that serves as an instance of (C) in Chalmers’s schema. Standard phenomenal concept theorists hold that our concepts of the physical and our concepts of the mental are (i) distinct from one another, and (ii) incommensurable with one another. The concepts are ‘incommensurable’ in the sense that propositions conceived under phenomenal concepts don’t \textit{a priori} entail any propositions conceived under physical concepts. More generally, there are no interesting \textit{conceptual connections} between physical concepts and phenomenal concepts. Any connection between propositions conceived under phenomenal concepts and propositions conceived under physical concepts is psychologically contingent, and does not fall out of the concepts themselves. In other words, there is a \textit{conceptual gap}.

Perhaps the most well-known approach to the phenomenal concept strategy is to say that phenomenal concepts are a certain sort of \textit{recognitional concept}.\(^{10}\) The basic idea behind the strategy

\(^9\)Chalmers (2006: 8)
\(^{10}\)Both Loar (1997) and Tye (2003) take this sort of approach. Other approaches include treating phenomenal concepts as \textit{indexical} concepts (Ismael 1999, Perry 2001), as \textit{quotational} concepts (Block 2006, Papineau 2006), or as
runs roughly as follows. Recognitional concepts are deployed during acts of perceptual recognition, as when one recognizes a particular dog as being a certain sort of dog. A subject can apply a recognitional concept to an object or property without having a scientific (or physical) concept of that object or property. For example, one might be able to recognize a schnauzer as that kind of dog, even if one doesn’t know that it’s called a ‘schnauzer,’ that it’s a German breed, that was bred to hunt rats, and so forth. Conversely, one might know an awful lot about schnauzers without being able to recognize one.\(^{11}\) Hence, there is a sort of epistemic gap between the recognitional concept and the scientific (or physical) concept. The phenomenal concept strategist goes on to say that the case of consciousness is analogous to the case of the schnauzer: we pick out the same property (phenomenal redness, for example) via two distinct concepts, one recognitional and one physical.

I won’t attempt to evaluate the phenomenal concept strategy here. However, I will point out that even if successful, the phenomenal concept strategy stands in need of empirical support. Advocates of the phenomenal concept strategy have typically not said much to support the thesis that we actually possess the relevant concepts. If we do have the right sort of concepts, then the strategy may turn out to be successful. But if we don’t have the relevant concepts, then the strategy is a non-starter. It seems to me that this way of responding to the explanatory gap puts the cart before the horse. We would be better off proceeding by first noting some psychological features that we actually possess, and then using those features to devise a strategy for resisting the explanatory gap.\(^{12}\) Minimally, it would be desirable for the materialist to bind up his strategy more intimately with empirical psychology.

\(^{11}\)This is rather implausible in the case of schnauzers, since its likely that anybody who knows a lot about schnauzers has seen a picture of a schnauzer.

\(^{12}\)Though see Robbins & Jack (2006), who seem to put the horse before the cart. Robbins & Jack leverage empirical data about our concepts of agents and experiencers in order to motivate a version of the phenomenal concept strategy.
2 The View from the Cognitive Science of Explanation

2.1 Explanation and Motivation

We humans like to explain things, and moreover it’s *good* for us to explain things. Explanations help us to navigate the world more effectively, and ultimately help us to reproduce our genes more effectively. That is, the production of explanations is *evolutionarily beneficial*.\(^\text{13}\) This isn’t to say that we generate explanations *because* it is evolutionarily beneficial to do so—at least, it need not seem this way from the perspective of the explainer. Rather, we generate explanations because it is intrinsically rewarding. Generating explanations *feels* good. Consider the analogy with hunger. Eating is extremely important for survival and reproduction. But we don’t eat in order to survive and reproduce. Rather, we eat because we feel *hungry*, and eating food satisfies this hunger. Beyond merely satisfying hunger, it *tastes good* to eat food. In general, agents are not directly motivated by considerations of evolutionary advantage, but rather by intrinsic rewards and good feelings. The motivation to generate explanations is no exception in this regard. In the following sections, I will develop and expand upon this theme.

2.1.1 The Phenomenology of Explanation

Alison Gopnik has an interesting take on the coupling between explanation and motivation. She argues that there is a distinctive *phenomenology* of explanation, and that this phenomenology signifies the operation of an underlying motivational system.\(^\text{14}\) The reader might be suspicious whether we can make any progress by merely observing *what it’s like* to generate an explanation. After all, it is notoriously difficult for theorists to reach mere *agreement* as to what a particular phenomenon is like, and more difficult still to leverage phenomenological considerations in support of substantive theses.\(^\text{15}\) Gopnik’s phenomenological observations about explanatory phenomenological are

---

\(^\text{13}\)&ndash; Though this much is pretty clear, it’s less clear *why* explanation is evolutionarily beneficial. I’ll say a bit more about this in §2.2 and §2.3.

\(^\text{14}\)&ndash; Gopnik (1998)

\(^\text{15}\)&ndash; For a more optimistic view of phenomenological disputes, see Siewert (1998) and Siewert (*forthcoming*).
relatively uncontroversial, however, and should be familiar to those who are in the business of generating explanations.\footnote{Really, the phenomenological observations should be familiar to anybody who is paying attention. But people who generate explanations for a living (scientists, academics, perhaps lawyers, and so forth) should be especially sensitive to the observations.}

[T]here is a distinctive phenomenology associated with explanation. The phenomenology involves both the search for explanation and a recognition of the fact that an explanation has been reached. We might call them the ‘hmm’ and the ‘aha’. In English they seem to be expressed by ‘why?’ and ‘because’. These experiences are close to what we more broadly call curiosity or interest but they are not identical with them.\footnote{Gopnik (1998: 108)}

Gopnik’s suggestion is highly plausible. Introspectively, there seems to be a distinctive feeling associated with grasping an explanation. It feels good to grasp an explanation. But this feeling goes beyond mere pleasure, as it is distinctly cognitive (or content-involving).\footnote{It may be the cognitive element of explanatory phenomenology that led John Stuart Mill to hold that intellectual pleasure is in some sense more worthwhile than mere sensory pleasure of the sort that could be had by (e.g.) a pig.}

The story of Archimedes illustrates the aha! feeling nicely. According to the story, the King commanded Archimedes to determine whether his new crown was made of pure gold, or whether it was adulterated with other metals.\footnote{Apparently the King was worried that he’d been ripped off by the crown-jeweler!} Archimedes knew that in order to solve the problem, he would have to determine the density of the crown. But he couldn’t figure out a way to determine the crown’s density without melting the crown (which would undoubtedly upset the King). While drawing a bath one day, Archimedes noticed that the water level rose by a measurable amount as he climbed into the tub. He realized that this effect could be exploited to measure the volume of the crown, and along with a weight measure, the density of the crown could be determined sans melting. Archimedes exclaimed, “Eureka, I’ve found it!” He was so excited about the solution that he jumped out of the bath and sprinted directly to the King’s chamber in order to inform him of the good news. When Archimedes jumped out of the tub, he was experiencing the aha! feeling.

The hmm... feeling should be familiar to anybody with children. Young children apparently have an insatiable appetite for explanations. They demand explanations with a seemingly endless
stream of “Why?” questions. When you (the parent) just can’t take it anymore, and attempt to stop the flood of questions by answering “Because!”, the child remains curious, and craves further explanation. She is not satisfied with a mere “Because.” Plausibly, the child is experiencing the hmm... feeling that is characteristic of curious behavior. Of course, the hmm... feeling is not limited to children; we all experience the feeling from time to time. One experiences the hmm... feeling whenever one is curious or puzzled about a phenomenon, and “formulating a unified, consistent story is one way to eradicate that peculiar feeling of inward unrest. An explanation pleasantly discharges that feeling of intellectual unease.”\(^{20}\) Consider doing a crossword puzzle, for example. Once one begins working on the puzzle, one experiences a strong craving to complete the puzzle. This craving goes beyond the mere desire to ‘defeat’ the puzzle or to ‘achieve victory’—one also wants to know the solution to the puzzle. If one succeeds in the finishing the puzzle, the hmm... feeling is discharged, and one normally experiences the aha! feeling.\(^{21}\) If one is unable to complete the puzzle, then the hmm... feeling remains and one feels frustrated.

### 2.1.2 Two Kinds of Affective State

The aha! feeling clearly has a positive affective valence. Trout observes that “Few products of intellectual life are more pleasing to give and receive, than a good explanation.”\(^{22}\) As such, the aha! feeling is plausibly regarded as a reward mechanism. The purpose of the mechanism is to give the subject an immediate, intrinsic reward when an explanation has been generated.\(^{23}\) The intrinsically rewarding explanatory phenomenology offers the subject a short-term incentive to generate explanations, so that she will be more likely to accrue the long-run reproductive benefits. By contrast, the hmm... feeling has a negative affective valence, and signals the need to generate an explanation. The hmm... feeling is unpleasant, and is plausibly a kind of anxiety that functions...

---

\(^{20}\)Trout (2002: 217)

\(^{21}\)The feeling may be particularly strong at the moment it becomes clear how to fit the last few letters into place.

\(^{22}\)Trout (2002: 212)

\(^{23}\)Gopnik’s (1998) slogan that “explanation is like orgasm” emphasizes the positively valenced reward component of explanatory phenomenology. The analogy to orgasm should be clear!
to motivate the subject to generate or seek an explanation. If the *aha!* feeling is like orgasm, then the *hmm...* feeling is like *lust*.

These sorts of considerations suggest that there may be a *dual-component* motivational system prompting us to generate explanations. That is, the motivational system involves both a *positive* affective component, and a *negative* affective component. Either the positively valenced component or the negatively valenced component alone would be sufficient to yield some motivation to generate explanations. They are considerably more powerful, however, when used in combination. The complimentary components function *together* in order to yield *maximum* motivation.

Hunger, an extremely powerful motivational system, shares the same structure. Here the good taste of food is the positive affective component, and stomach pangs are the negative affective component. The dual-affect structure is also exploited by the powerful drive for sex. Orgasm has a positive valence, and lust has a negative valence. Finally, cases of substance addiction standardly exhibit the same sort of motivational structure.\(^{24}\) Consider smoking. It feels good to smoke. When you smoke a cigarette, you get a “buzz” that is intrinsically rewarding.\(^{25}\) But when a cigarette addict fails to smoke, she enters a state with a negative affective valence; namely, the state of craving a cigarette. Lighting up a cigarette is sufficient both to banish the craving and to yield a positive reward. This is perhaps one reason why it is difficult to break a cigarette addiction.

In all of the above cases, the combination of positive and negative affective states results in the creation of a powerful motivational system. Considerations from the phenomenology of explanation make it plausible to think that there is a similar motivational system underlying our drive to generate explanations. Since nature endowed us with such a powerful drive to generate explanations, we should expect to find a substantial adaptive benefit attached to that function; otherwise generating all those explanations would be a waste of precious cognitive resources.

---

\(^{24}\) I’ve often thought of philosophers as *explanation-addicts*; at least, this seems plausible in my own case. Maybe this sort of consideration explains why I find the analogy plausible.

\(^{25}\) Smoking may yield other rewards, namely *social rewards*. For example, some smokers claim that smoking helps them to meet people or feel less shy, helps them to look *cool*, and so forth. These sorts of things may (or may not) be real benefits of smoking, but here I am interested only in the *intrinsic* reward that smoking delivers.
2.2 The Function of Explanation

We noted above that it is evolutionarily beneficial to produce explanations. This may seem rather obvious, but it is important to think about why generating explanations is beneficial. In general, cognitive systems are adaptive because they afford access to certain useful properties of the world. For example, vision affords access to shape, size, and other spatial properties, to color properties, to textural properties, and so forth. Detecting these properties is clearly useful for survival. Knowing the size and location of a predator helps the organism to avoid the predator. Knowing the color of a potential food item helps the organism to know whether the food is nutritious or poisonous. And so forth. So we should wonder, to what sorts of properties does explanatory cognition afford access?

One way of thinking about explanation-producing cognitive mechanisms is as causality detectors that afford cognitive access to the deep causal structure of the world. Again, the analogy with visual perception is illustrative. The theory-generation system is like a perceptual system, in that its function is to detect worldly properties. Individual theories are like individual visual perceptions, in that they represent the world as having certain properties. The difference between the two systems is that they detect different properties. Whereas the visual system detects spatial properties, the theory-generation system detects causal properties:

> uncovering causes is the central feature of the theory-formation system from an evolutionary point of view, and other features of theories derive from it, just as uncovering the spatial character of moving objects seems to be central to the visual system.

While assessing Gopnik’s claim is beyond the scope of this paper, it is nonetheless a plausible claim. Perhaps there are other worldly properties that deserve consideration as targets for the theory-formation system: formal or structural properties, categorical properties, and so forth. For our purposes here, the issue of which properties our explanation-producing cognitive mechanisms

---

27 Gopnik (1998: 104)
track is not crucial. What is crucial is that any account on which the function of explanation is to track worldly properties is committed to a sort of realism about explanation. The causal view illustrates this point nicely, so we will continue with the causal view in mind.

2.3 Explanatory Realism

The key claim of explanatory realism is that explanations are assessable for veridicality. Although this is an intuitively plausible thesis to begin with, the analogy with perception amplifies the intuition that explanations are assessable for veridicality:

The visual system takes retinal input and transforms it into representations of objects moving in space. While these representations of objects are not perfectly veridical, they at least approach a greater level of veridicality than the retinal input itself, and presumably this fact explains the evolution of perception. Similarly, we can think of the theory formation system as a system that takes the input of the perceptual system and transforms it into representations of the underlying causal structure of the world. The theory formation system is designed to go beyond perceptual representations just as perceptual representations go beyond sensory input. The representations that result, representations formulated in terms of abstract theoretical entities and laws, more closely approach the actual causal structure of the world than the representations of perception... To apply a theory to some pattern of evidence is to assign the evidence a particular causal representation. Theories change in the face of evidence in order to give better causal representation.28

To say that explanations are assessable for veridicality is primarily to eschew a hard-core anti-realist theory of explanation (explanatory instrumentalism, for example). That is, the explanatory realist rejects the view that explanations are assessable only in virtue of utility of one sort or another. Instead, explanations can be evaluated with respect to their fit with the world.

The contrast with instrumentalism and other hard-core anti-realist doctrines might suggest that explanatory realism is itself a rather extreme doctrine. But explanatory realism need not be an extreme view (though it is certainly compatible with hard-core forms of realism). The explanatory realist is not committed to a correspondence theory of truth, for example. Nor is the explanatory realist committed to the denial of explanatory pluralism. There might be many equally good (i.e.

28Gopnik (1998: 103)
equally veridical) explanations that conflict with one another, or that are stated in terms that are not inter-translatable.

The core claim of explanatory realism is merely that explanations are assessable for veridicality. All that is required for veridicality assessment is that (i) explanations track some real properties of the world, and (ii) explanations can do a better or worse job of tracking those properties. Questions about the tracking relation, which properties are tracked, and what constitutes better or worse tracking are all left to be spelled out by the theorist. So explanatory realism is somewhat flexible: one is free to adopt either a hard-core realism, or a more squishy sort of realism.

Explanatory realism isn’t just a working assumption made by Gopnik and other scientists investigating the psychology of explanation. We can find independent motivation for the view in the philosophy of science literature. For example, Kim argues that good explanations track real connections in the world\(^{29}\). Hempel’s deductive-nomological account of explanation is a paradigmatic realist account. Trout goes so far as the claim that “Most of the widely discussed accounts of explanation have been objectivist [i.e. realist].”\(^{30}\) Of course, not everybody agrees that explanatory realism is on the right track. As with most any account in most any domain of philosophy, explanatory realism is controversial. I merely want to point out that both psychologists and philosophers find it plausible on independent grounds. For the remainder of the paper, I will provisionally suppose a broadly realist account of explanation.

### 2.4 Explanatory Illusions

What is the relationship between a good explanation and the *aha!* feeling? On a realist conception of explanation, an explanation is good just in case it represents the world in the right way. But experiencing the *aha!* feeling does not require that an explanation represents the world in the right way. So the natural thing for the realist to say is that the *aha!* feeling is merely *contingently*

---

\(^{29}\)Kim (1988)

\(^{30}\)Trout (2002: 217), my emphasis.
related to good explanations. The contingent relation opens up the possibility of the *aha!* feeling occurring in the absence of a good explanation, and vice versa. Gopnik emphasizes this point:

> [T]he phenomenology of explanation is, in the canonical case, connected with the operation of a distinctive cognitive system, the theory-formation system. Moreover, that theory-formation system evolved because in general and over the long run, and especially in childhood, it gives us a more veridical picture of the causal structure of the world around us. In a particular case, however, explanatory phenomenology may be divorced from the [successful] operation of this cognitive system.\(^{31}\)

To say that “explanatory phenomenology may be divorced from the successful operation of the cognitive system” is to say that the *aha!* feeling and good explanation are *doubly dissociable*. Good explanations can occur without any *aha!* feeling. For example, experts who deploy a particular explanation repeatedly and frequently may no longer experience the *aha!* feeling.\(^{32}\) Certain areas of quantum mechanics are arguably cases in which good theories that fail to produce the *aha!* feeling. Conversely, the *aha!* feeling can occur in the absence of a good explanation (more on this shortly), or in the absence of any explanation at all. For example, certain types of mystical experiences involve a powerful *aha!* feeling in the absence of any particular explanation.\(^{33}\) Call these sorts of experiences *explanatory illusions*.

Another theorist who holds that the *aha!* feeling and good explanation are doubly dissociable in particular cases is Trout. But where Gopnik remains neutral on the general reliability of the *aha!* feeling as an indicator of good explanation, Trout goes further and argues that the *aha!* feeling is in general a *poor* guide to good explanation:

> [T]he psychological sense of understanding [what I’ve been calling ‘the *aha!* feeling’] is just a kind of confidence, abetted by hindsight, of intellectual satisfaction that a question has been adequately answered. Thus this sense of satisfaction is confidence that one enjoys an accurate description of the underlying causal factors sufficient to bring about the phenomenon we are examining. But confidence is, notoriously, not an indicator of truth.\(^{34}\)

\(^{31}\)Gopnik (1998: 112)

\(^{32}\)Or perhaps they experience some *aha!* feeling, but the intensity of the feeling is diminished.

\(^{33}\)Gopnik (1998: 112)

\(^{34}\)Trout (2002: 213)
Trout points to historical examples in order to motivate his thesis (for example, a geocentric astronomical model produced the *aha!* feeling in Ptolemy). However, he also musters considerable evidence from cognitive psychology. In particular, Trout thinks that the *hindsight bias* and *over-confidence bias* support the claim that explanatory phenomenology is not a reliable indicator of good explanation. We should also append Keil and Wilson’s *illusion of explanatory depth* to our list of known explanatory illusions, though it is not discussed by Trout. All of these explanatory illusions involve *overestimating* the quality of our explanations, so in order to keep things moving briskly I will not discuss all three illusions here. I think that the hindsight bias evidence is the simplest and most telling of the three, so I will discuss it at the expense of the other two illusions:

Fischoff, Slovic and Lichtenstein (1977) asked adults subjects to indicate the most frequent cause of death in the U.S., and to estimate their confidence that their choice was correct (in terms of ‘odds’). It turns out that, when subjects set the odds of their answer’s correctness at 100:1, they were correct only 73 percent of the time. Remarkably, even when they were so certain as to set the odds between 10,000:1 and 1,000,000:1, they were correct only between 85 percent and 85 percent of the time.36

Essentially, the evidence from the overconfidence bias suggests that we are often more confident in our explanations than we ought to be (hence the name of the effect). To put it a bit differently, we often think that our explanations are *better* than they really are. So in such cases, the *aha!* feeling is illusory in that it is exaggerated relative to the quality of the explanation. This kind of illusion is roughly analogous to size illusions in visual perception (the moon illusion, for example). Although the overconfidence bias and other known explanatory illusions suggest that we tend to *overestimate* the quality of our explanations, it could also turn out that under certain circumstances we tend to *underestimate* the quality of our explanations. It is important to keep a watchful eye out for this sort of illusion.

36Trout (2002: 226)
3 The Explanatory Gap as an Explanatory Illusion

How is all this related to the explanatory gap? Roughly, the idea is that a great many philosophers take the *aha!* feeling to be an indicator of a good explanation— or even as a necessary condition on a good explanation. But good explanations appear to be only contingently connected to the *aha!* feeling. We can have explanatory phenomenology without a good explanation (hindsight bias, for example), and we could potentially have a good explanation without the *aha!* feeling (quantum mechanics, arguably). Perhaps the explanatory gap is just an extreme case of a good explanation without any *aha!* phenomenology. If this sort of situation obtained, a good explanation of consciousness would be a case where the function of explanation is fulfilled without triggering the normal reward mechanism. So if explanation is like orgasm, then a materialist explanation of consciousness would be like creating a child *in vitro*. This result would clearly be to the advantage of the materialist, as it would effectively ‘explain away’ the explanatory gap as a product of our explanatory psychology.

It’s worth noting that strictly speaking, this sort of strategy for dealing with the explanatory gap is an instance of the “phenomenal concept strategy” as defined by Chalmers (above). Though it is a “phenomenal concept strategy” in letter, the approach has a rather different spirit. The strategy under consideration does not directly appeal to any special phenomenal concepts. Further, it is clearly an empirically-driven strategy, whereas most standard approaches to the phenomenal concept strategy are largely a priori. Another nice feature of this strategy is that it takes the slogan that “the explanatory gap is a cognitive illusion” quite seriously. Some phenomenal concept theorists wield this slogan without making good on the analogy to perceptual illusions.\(^{37}\) On the current strategy, by contrast, the explanatory gap is literally an illusion: an explanatory illusion.

The astute reader will have noticed that there is a rather glaring lacuna in the strategy, however. All known explanatory illusions are cases where the *aha!* feeling occurs in the absence of a good explanation.

\(^{37}\)For example, Tye (1999).
explanation. So far, there aren’t any known cases where the converse occurs— but this is just the sort of case that would be *required* in order to explain away the explanatory gap. At this point, the advocate of the explanatory illusion should point out that the research project is still in its early stages, and that more data is forthcoming. This may sound like a rather wimpy response, and really it *is* wimpy. Even so, we can consider the above argument for the explanatory illusion strategy as a non-wimpy argument for the following conditional:

*If we can find a clear case of a good explanation without the *aha!* feeling, then* we can leverage that sort of case to explain away the explanatory gap.

While this is clearly not a complete solution for the materialist, it *does* serve to open up a new avenue of inquiry. Perhaps there is a distinctive feature of the psychology of explanation that can help the materialist to deal with the explanatory gap. And even though the explanatory illusion strategist doesn’t yet have a complete account in hand, there is something to be said about how to approach the search for the desired sort of case: we might investigate the psychology of *reductive explanation*.

### 3.1 Reductive Explanation In General Doesn’t Yield the *Aha!* Feeling

Philosophers of mind like to talk about “reductive explanations.” Reductive explanations are explanations that involve identity statements connecting categories (or kinds) of a relatively high level of explanation to categories (or kinds) of a relatively lower level of explanation. For example, “water is H₂O” and “lightning is a certain sort of electrical discharge” are reductive explanations of water and lightning, respectively.³⁸ Relatedly, paradigmatic reductive explanations involve something like essences (e.g. H₂O is the essence of water), in such a way that the causal profile of the essence explains the causal profile of the target phenomenon. Since materialists want to say that the essence of consciousness is ultimately physical, they will need to deploy a reductive explanation.

³⁸Or at least, they are *gestures* in the direction of such explanations.
Some philosophers have argued that reductive explanations require a certain sort of epistemic connection: namely, a priori entailment relations between the essence and the target of explanation. For example, David Chalmers and Frank Jackson argue for this sort of view.\footnote{Chalmers & Jackson (2001)} Of course, this view is controversial: Ned Block and Robert Stalnaker argue that there can be reductive explanations without any related a priori entailment.\footnote{Block & Stalnaker (1999)} Indeed, they argue that in general, reductive explanation does not require a priori entailment. Block and Stalnaker claim, for example, that there is no a priori entailment from the complete physical story about the world to the conclusion that water is \(H_2O\). Instead, the inference is ampliative (or inductive). Does reductive explanation really require a priori entailment? I won’t attempt to adjudicate this hairy dispute here, but I have a few suggestions that I hope will be useful.

First, I think that it is useful to point out a reason that some philosophers might be inclined to think that reductive explanation does require a priori entailment. Reflecting on the phenomenal character of explanation is helpful in this regard. When a subject grasps an a priori entailment, it is plausible to think that the subject normally experiences an aha! feeling. Consider, for example, grasping the truth that \(2 + 2 = 4\). One can grasp this proposition in a single mental act, and in doing so one experiences a rather strong aha! feeling. There is similar phenomenology for other a priori truths: grasping a logic proof, or grasping a truth in virtue of meaning, for example. I want to suggest that because a priori entailment is typically accompanied by the aha! feeling, and the aha! feeling is thought by many to be a necessary condition for a good explanation, philosophers like Chalmers and Jackson may be tempted to conclude that a priori entailment is necessary for a good explanation. That is, the link between a priori entailment and reductive explanation is, plausibly, the aha! feeling. But notice that this is essentially the question at issue in the arguments from the preceding sections. The suggestion, then, is that the a priori entailment debate is a red herring.

\footnote{Chalmers & Jackson (2001)} \footnote{Block & Stalnaker (1999)}
Second, I think it’s helpful to point out that the question of whether reductive explanation is typically accompanied by the *aha!* feeling is something that we ought to be able to test empirically. We could start by giving subjects paradigm cases of reductive explanation, and observing the result. It may turn out that in general, reductive explanations do not yield *aha!* phenomenology. If so, the link between good explanation and explanatory phenomenology will have been broken. Consequently, I think that this is an area of research that materialists should be very eager to explore.

4 Future Research

At this stage of the game, it would be premature to draw conclusions. The ‘new strategy’ is really just the *skeleton* of a strategy, which will accrue *meat* as more data comes in. The obstacle to the approach is formidable, but perhaps not insurmountable. By learning more about the cognitive mechanisms that support explanatory phenomenology, we may learn why the explanatory gap arises; and if we’re lucky, this will make the gap seem less mysterious. The only way to find out is to learn more about the relevant mechanisms.

Even though the approach under consideration is still in its ‘skeletal’ stage, we should take it as a cue regarding future research directions. Researchers interested in the nature of conscious now have a new domain of empirical research to pursue (or to *monitor*, for armchair theorists who simply read about the stuff). Thus far, nobody has found the *key* insight that would demystify consciousness, so it is well worth keeping an eye out in as many domains as possible. In the past, theorists have looked to fields like neuroscience (binding), artificial intelligence (strong AI), and physics (quantum theory) for new insights regarding the nature of consciousness. Now we can (and should) look to research involving the psychology of explanation as well.
5 References


