On the psychological origins of dualism:

Dual-process cognition and the explanatory gap*

Brian Fiala, Adam Arico, and Shaun Nichols

University of Arizona

Abstract:

Consciousness often presents itself as a problem for materialists because no matter which physical explanation we consider, there seems to remain something about conscious experience that hasn't been fully explained. This gives rise to an apparent *explanatory gap*. The explanatory gulf between the physical and the conscious is reflected in the broader population, in which dualistic intuitions abound. Drawing on recent empirical evidence, this essay presents a dual-process cognitive model of consciousness attribution. This dual-process model, we suggest, provides an important part of the explanation for why dualism is so attractive and the explanatory gap so vexing.

---

1. The Explanatory Gap

Perhaps the most broad and unassuming philosophical question about consciousness is "What is the relationship between consciousness and the physical world?" It is *prima facie* difficult to see how the pains, itches, and tickles of phenomenal consciousness could fit into a world populated exclusively by particles, fields, forces, and other denizens of fundamental physics. But this appears to be just what physicalism requires. How could a thinking, experiencing mind be a purely physical thing?

One approach to this problem emphasizes our *epistemic* situation with respect to consciousness, and especially the distinctively *explanatory* situation. Epistemic approaches focus on whether we can acquire knowledge, justified belief, or an adequate explanation regarding the nature of consciousness. Thomas Huxley famously gestures at this aspect of the problem of consciousness:

> But what consciousness is, we know not; and how it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed his lamp. (Huxley 1866, 193)

* Acknowledgements: We have several people to thank for discussion and comments on the manuscript: Sara Bernstein, Mark Collard, Jonathan Cohen, Chris Hitchcock, Terry Horgan, Bryce Huebner, Chris Kahn, Josh Knobe, Uriah Kriegel, Edouard Machery, Ron Mallon, J. Brendan Ritchie, Philip Robbins, Ted Slingerland, and Josh Weisberg.

Huxley's suggestion is that no account can be given of the relationship between consciousness and the brain, where an "account" amounts to something like 'an adequate scientific explanation'. Huxley's skepticism about the prospects for a physicalist account of consciousness drives him to the view called "epiphenomenalist dualism," according to which consciousness is not itself a physical phenomenon and has no causal impact on any physical phenomena, but is nonetheless (and rather mysteriously) *caused by* physical phenomena (1874/2002).

More recently, Levine has introduced an updated version of this problem, which he dubs "the explanatory gap." According to Levine, "psycho-physical identity statements leave a significant *explanatory gap*" (1983). That is, theories that attempt to explain consciousness by identifying it with some physical property or process will inevitably seem to leave out something important about consciousness. Specifically, what's supposed to be left out is the felt quality of *what it's like* to undergo a conscious experience such as seeing the color red. Since it appears inevitable that purely physical theories will 'leave something out,' Levine suggests that there is a serious worry that such theories will inevitably fail to adequately *explain* consciousness. Levine elaborates on this suggestion by claiming that there is an air of "felt contingency" about the relationship between consciousness and processes in the brain (and indeed, between consciousness and *any* physical process). That is, there seems to be something contingent or arbitrary about any purported connections between physical processes and conscious states. But good explanations are not arbitrary.[1] As a result, it is hard to see how a theory invoking 'mere' brain activity could be a complete explanation of consciousness. Levine concludes, along with Huxley, that the explanatory gap is a serious obstacle for physicalism.

One prominent argumentative strategy at this juncture is to draw on this apparent epistemic obstacle as support for conclusions about the nature of consciousness. For example, one might take the explanatory gap(s) discussed by Huxley and Levine as indicative of a corresponding duality in nature. If no physical theory can fully *explain* consciousness, it seems doubtful that consciousness *is* something physical. For if something is not fully physically explicable then it is not a completely physical phenomenon. Therefore, consciousness must not be a physical phenomenon. This formulation of the argumentative strategy is overly simple, but it serves to illustrate the strategy of arguing from epistemic premises to conclusions about the nature of consciousness.

While the explanatory gap is central to contemporary philosophy of mind, it is plausible that the gap gives philosophical expression to a much more pervasive phenomenon – even people without any philosophical training find it bizarre and counterintuitive to think that consciousness is nothing over and above certain processes in the brain. Paul Bloom takes this to be part of a universal inclination towards *folk dualism*. According to Bloom, people are "dualists who have two ways of looking at the

[1] Admittedly, this gloss on the issue of modality and reductive explanation crushes many subtleties. We apologize for this injustice. For reasons that will soon become clear, our primary focus in this paper is on the psychological aspects of the problem of consciousness, rather than on the modal aspects.

world: in terms of bodies and in terms of souls. A direct consequence of this dualism is the idea that bodies and souls are separate" (2004, 191). Folk dualism is associated with a range of beliefs including beliefs about free will and personal identity. The rift between consciousness and the physical world is taken to be one central element of folk dualism:

> People universally think of human consciousness as separate from the physical realm. Just about everyone believes, for instance, that when our bodies die, we will survive – perhaps rising to heaven, entering another body, or coming to occupy some spirit world. (Bloom 2006, 211).[2]

What makes brain-consciousness dualism so seductive in both philosophy and ordinary life? Why does the explanatory gap carry so much intuitive weight? We suspect that the answers to these questions have much in common. The common answer we have in mind is *psychological* in nature, and so we turn to the psychological underpinnings of the attribution of consciousness.

2. The Psychology of Attributing Conscious States

Claims about cognitive architecture figure centrally in our explanation of the capacity to attribute conscious states, and how this capacity figures in dualistic patterns of thought. Specifically we claim that *dual-process* cognitive architecture plays a key role in the psychology underlying explanatory gap intuitions and folk dualism. We thus begin with a brief introduction to dual-process models.

2.1 Dual-Process Models

In recent years dual-process theories have been proposed for all sorts of cognitive phenomena, including moral judgment (Haidt, 2001), decision-making (Stanovich and West, 2000; Stanovich2004), probabilistic reasoning (Sloman, 1996), and social cognition (Chaiken and Trope, 1999). A crude version of dual-process theory holds that mental systems fall into two classes. In one class, *System 1,* we find processes that are quick, automatic, unconscious, associative, heuristic-based, computationally simple, evolutionarily old, domain-specific and non-inferential. In the other class, *System 2,* we find processes that are relatively slow, controlled, introspectively accessible, rule-based, analytic, computationally demanding, inferential, domain-general, and voluntary.

Since processes from these two classes of systems operate so differently, it shouldn't be surprising that a *System 1* process and a *System 2* process sometimes

[2] Experiments by Richert & Harris (2006 & 2008) indicate that, pace Bloom (2004), people do not explicitly identify the mind with the soul. Beliefs about the soul are rather messy, but people do largely think that the soul is independent of the body (Richert & Harris 2008). And Richert & Harris seem to agree that people take consciousness to be tied to the soul: "the concepts of identity, consciousness and soul are deeply intertwined in human cognition about other humans" (Richert & Harris 2008, 99-100).

produce conflicting outputs with respect to the same cognitive task or subject matter. For instance, consider the following argument:

> All unemployed people are poor.
> Rockefeller is not unemployed.
> Conclusion: Rockefeller is not poor.

On reading this argument, many people judge incorrectly that the argument is valid. According to dual-process theory, that is because people's belief in the conclusion biases a *system 1* reasoning process to the incorrect answer. However, most people can be brought to appreciate that the argument is not valid, and this is because we also have a *system 2* reasoning process that has the resources to evaluate the argument in a consciously controlled, reasoned fashion (see, e.g., Evans 2007). Of course, *System 1* and *System 2* can (and quite often do) arrive at the same verdict. For instance, changing the first premise of the above argument to "Only unemployed people are poor" allows both systems to converge on the judgment that the argument is valid.

Although the dual-process paradigm provides a tidy picture of the mind, it is unlikely that all mental processes will divide sharply and cleanly into the two categories, such that either a process has all the characteristic features of *System 1* or all of the characteristic features of *System* 2. It would not be surprising, for instance, to find processes that are fast and computationally simple but not associationistic (cf. Fodor 1983). So if we find that a process has one characteristic *system 1* feature, it would be hasty to infer that the process has the rest of the characteristic *system 1* features. Nonetheless, the dual-process approach is useful so long as one is clear about the particular characteristics of a given psychological process (cf. Samuels, forthcoming; Stanovich and West, 2000; Stanovich, 2004).

We think the distinction between processes that are automatic and processes that are consciously controlled really does capture an important difference between cognitive systems in many domains, including the domain of conscious-state attributions. We suggest (i) that there are two cognitive pathways by which we typically arrive at judgments that something has conscious states, and (ii) that these pathways correspond to a *System 1*/*System 2* distinction. On the one hand, we propose a "low-road" mechanism for conscious-state attributions that has several characteristic *System 1* features: it is fast, domain-specific (i.e., it operates on a restricted range of inputs) and automatic (the mechanism is not under conscious control).[3] On the other hand, there are judgments about conscious states that we reach through rational deliberation, theory application, or conscious reasoning; call this pathway to attributions of conscious states "the high road."

[3] We remain neutral on whether the mechanism has other features of System 1, like being associationistic, evolutionarily old, and computationally simple.

2.2 Conscious attribution: The low road

In an earlier paper (Arico et al. forthcoming), we sketch a model – the AGENCY model – of one path by which we come to attribute conscious states.[4] In our dual-process approach, the AGENCY model describes the "low road" to conscious-state attribution. According to this model, we are disposed to have a gut feeling that an entity has conscious states if and only if we categorize that entity as an AGENT, and typically we are inclined to categorize an entity as an AGENT only when we represent the entity as having certain features. These features will be relatively simple, surface level features, which are members of a restricted set of potential inputs to the low road process. Previous research has identified three features that reliably produce AGENT categorization: that the entity appears to have eyes; that it appears to behave in a contingently interactive manner; and that it displays distinctive (non-inertial) motion trajectories.

We developed the AGENCY model as a natural extension of work in developmental and social psychology. In their landmark study, Fritz Heider and Marianne Simmel (1944) showed participants an animation of geometric shapes (triangles, squares, circles) moving in non-inertial trajectories. When participants were asked to describe what was happening on the screen, they tended to utilize mental state terms—such as "chasing", "wanting", and "trying"—in their descriptions of the animation. This suggests that certain types of distinctive motion trigger us to attribute mentality to an entity, even when the entity is a mere geometric figure.

More recently, developmental psychologists Johnson, Slaughter, and Carey (1998) presented 12-month-olds with various novel objects, one of which was a "fuzzy brown object". Johnson, et al. found that when the fuzzy brown object included eyes, infants displayed significantly more gaze-following behavior than when the fuzzy brown object did not include eyes. They also found that infants displayed the same gaze-following behavior when the fuzzy brown object, controlled remotely, moved around and made noise in apparent response to the infant's behavior. Johnson and colleagues explain these effects by suggesting that when an entity has eyes or exhibits contingent interaction, infants (and adults) will categorize the entity as an *agent*. Once an entity is categorized as an *agent,* this generates the disposition to attribute mental states to the entity, which manifests in a variety of ways, including gaze following, imitation, and anticipating goal-directed behavior. Figure 1 depicts the model of mental state attribution suggested by these studies.[5]

---

[4] This model was developed in the wake of recent work on the folk psychology of consciousness (Gray et al. 2007, Knobe & Prinz 2008, Sytsma & Machery 2009). As with the other work in the area, our model focuses on attributions of conscious states to others. But it's possible that a quite different mechanism is required to explain attributions of conscious states to oneself.

[5] Similar results have been found by Shimizu & Johnson, who built upon a finding by Amanda Woodward. Woodward (1998) had infants watch a human arm reach for an object, and then moved the object to a different location; when the arm reached, again, to the original location of the object, rather than to the object in its new location, infants

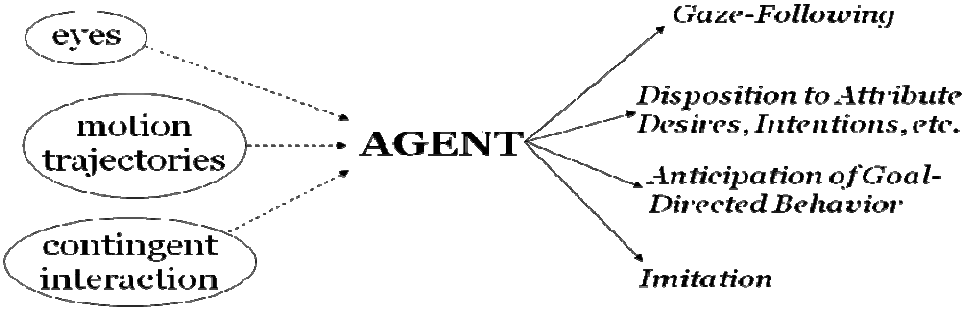

Figure 1: Model of AGENT detection (a la Johnson 2003)

We propose that this cognitive process also explains many of our everyday attributions of consciousness. In addition to imitation, gaze-following, and reasoning about beliefs and desires, we suggest that agent-categorization also plays a central role in disposing people to regard entities as capable of having conscious experiences.

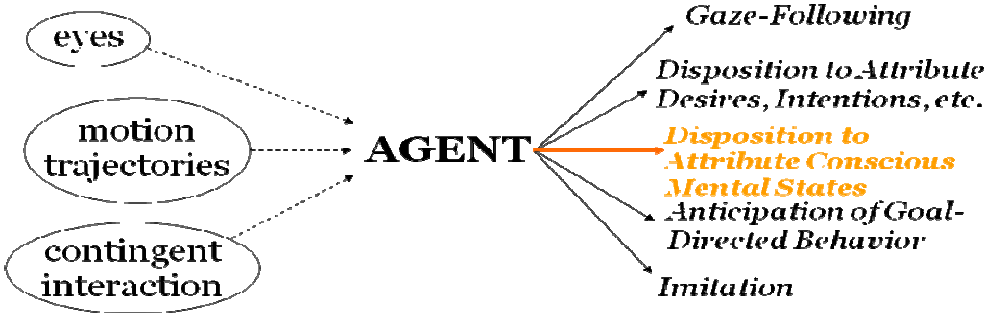

Figure 2: The AGENCY model

This model has empirically-testable predictions. For instance, if we assume that the AGENCY model depicts the primary low road mechanism for attributing conscious states to others, we should expect to find that people will not be immediately intuitively compelled to attribute conscious states to entities that typically lack the triggers for categorizing an object as an AGENT. More specifically, the model predicts that people will not have any immediate intuitive inclination to attribute conscious states to things such as trees, clouds, cars and rivers, since all of them lack eyes, contingently interactive behavior, and the distinctive kinds of motion trajectories investigated by Heider and

---

showed surprise. Shimizu & Johnson (2004) found that babies also showed surprise when a "blob" behaved similarly to the human arm in Woodward's experiments, but only when the "blob" had contingently interacted with a confederate. That these babies showed the same surprise for the contingently interactive "blob" as for a human arm suggests that they were utilizing psychological reasoning in making predictions for both. A broader survey of these kinds of results is presented in Johnson (2003).

Simmel.[6] In addition, the model predicts that people will be automatically inclined to attribute conscious states to the kinds of entities that have the superficial cues.

This is precisely what we found in a reaction time experiment (Arico et al. forthcoming). One characteristic of dual process models, including ours, is that the low road system is supposed to be very fast; by contrast the high road system, which draws on a broader information base, is comparatively slow. In a reaction time paradigm under speeded conditions, low-road processing should occur quickly and automatically, with high-road processes lagging behind. Given this standard interpretation of response times, the AGENCY model predicts slower reaction times when participants *deny* conscious state attributions to objects that are typically classified as AGENTS (as compared broadly to non-AGENTS). The idea is that if someone were to overtly respond that entities categorized as AGENTS don't feel pain (e.g. because they lack appropriate neural structures), this would require overcoming the hypothesized low-road disposition to attribute conscious states to those entities, which would take some extra time. To test our model, we presented subjects with a sequence of Object/Attribution pairs (e.g., *ant/feels pain*), and the subjects were asked to respond as quickly as possible (Yes or No) whether the object had the attribute. Attributions included properties like "Feels Happy" and "Feels Pain", while objects included various mammals, birds, insects, plants, vehicles, and natural objects. We recorded both participants' overt judgments and the time taken to answer each item. We found that participants quickly affirmed attributions of consciousness for those objects that typically possess the relevant features (mammals, birds, insects), while they responded negatively in response to attributions of consciousness to things that typically lack those features (vehicles, inanimate moving objects like clouds and rivers). More importantly, the reaction time results confirmed the predictions entailed by the AGENCY model of low-road consciousness attributions. Participants responded significantly more slowly when they *denied* conscious states to objects that *do* have the superficial AGENCY cues, namely, insects. This result is neatly explained by our hypothesis that insects automatically activate the low road to consciousness attribution; in order to deny that insects have conscious states, subjects had to "override" the low-road output, which explains why reaction times are slower in such cases.[7]

These experiments provide support for the AGENCY model of low-road consciousness attribution. However, there are numerous ways that this low-road process might be triggered, and many of the details of this process are largely underdetermined by the existing data. Nonetheless, the data corroborate our proposal

---

[6] Of course, anthropomorphized cartoon versions of such objects may well induce an immediate inclination to attribute conscious states. On our account, the natural explanation is that cartoons induce consciousness-attribution precisely because they have the right kinds of triggering featural cues.

[7] Why do people override the low road at all? Why not accept the gut feeling that a spider (for example) can feel pain? People might override because of known facts about arachnid neuroanatomy: for example, that spiders lack nociceptors. Another possibility is that people override because of socially acquired 'scripts' about spiders: for example, "*Of course* spiders don't feel pain!"

that low-road attributions of conscious states are generated by an AGENT mechanism that is triggered by a restricted range of inputs. (See Figure 3).

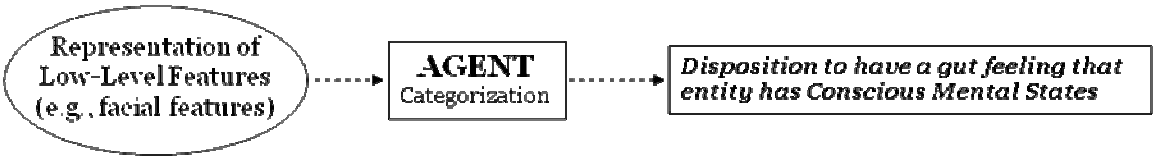

Figure 3: The AGENCY model of the low-road path to attributions of conscious states.

2.3 Conscious attribution: The high road

Thus far we have focused the discussion on one pathway for attributing conscious states. The low-road mechanism is not, however, the only pathway for attributing conscious states. One might instead rely on deliberation and inferential reasoning to conclude that an entity satisfies the criteria laid out by some scientific theory and, thus, judge that it (probably) has conscious states. For instance, if a trusted source tells Uriah that a certain organism has a sophisticated neural system (including nociceptors and dorsolateral prefrontal cortex), and if Uriah relies on a rudimentary theory of pain processing, then he might infer that the organism can probably feel pain. Or if a reliable informant tells Terry that some entity is conscious, Terry might conclude from that testimony that the entity is, in fact, a conscious thing. What matters is that one can come to think that an entity has conscious states via a pathway that has features typically associated with System 2: processing that is domain-general, voluntary and introspectively accessible. The process is domain general in that the inputs are not restricted – evidence can potentially come from anywhere. The process is voluntary because we can control when reasoning starts and stops. And it is introspectively accessible because the steps of the inferential process are typically available to consciousness. Let us examine each of these three features in a bit more detail.

Like most reasoning, high-road attributions of consciousness can potentially draw on an immense supply of information for evidence regarding an entity's having conscious states. Potential resources include the individual's current perceptual state, background beliefs, memories, and testimony from trusted sources. From there, the high-road reasoning process is only constrained by whatever rules of inference the individual has internalized.[8]

High-road attributions of consciousness are voluntary actions in the same sense that many conclusions reached via deliberate inferences are voluntary. Such conclusions

[8] Of course, limitations of working memory, time constraints, and motivational factors will also have some impact on the process.

are voluntary because we're able to choose to initiate and sustain the process of reasoning about another entity's consciousness (for example). The high-road process is engaged when one deliberately contemplates whether another entity is conscious, which is something that one can decide to do, to continue to do, or to stop doing. The high road attributions are not the result of an automatically triggered process that necessarily runs to completion upon activation; they are typically the result of someone deliberately thinking about whether something is conscious.

Because the high road proceeds through deliberate reasoning, the inferential steps from the initial assumption(s) to the conclusion are typically introspectively accessible to the individual. That is, at any point one can take notice of the line of reasoning that the high-road is processing, and know what inferences are being drawn. Where the low road is hidden from such introspective access, the high road is 'transparent' to introspection.

We offer one last example to illustrate high-road consciousness attributions. Consider Mill's argument that other humans have sensations and other mental states. Mill writes,

> [B]y what considerations am I led to believe… that the walking and speaking figures which I see and hear, have sensations and thoughts…? I conclude that other human beings have feelings like me, because, first, they have bodies like me… and because, secondly, they exhibit the acts, and other outward signs, which in my own case I know by experience to be caused by feelings. (Mill 1865)

This is, of course, just one example of how high-road reasoning might proceed, but it suffices to highlight many of the hallmark features of such reasoning. Mill draws on his observations about the similarity of bodies, as well as beliefs about behavior and its link to experience. At any point, Mill could able to stop or revise the line of reasoning he is pursuing. And not only is he aware of how the reasoning proceeds, but he is able to verbalize it for his reader. The high road to consciousness attribution is represented in figure 4:

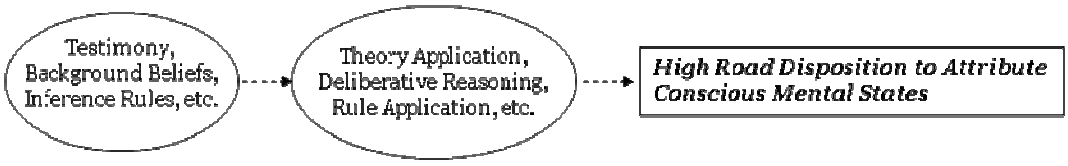

Figure 4: The high-road path to attributions of conscious states.

We have emphasized the distinctness of the two processes, but of course the two systems both deliver outputs concerning conscious states. And often the outputs will converge. Mill argues on the basis of analogy that others have sensations. This is a paradigmatically high road affair. But when Mill observed other humans, his low road was, no doubt, also activated. His high-road argument and his low-road reactions converged on the conclusion that other humans have sensations.

Although the two processes will often converge, this won't always be the case. We have already seen one illustration of this. While insects trigger the low road to attributing conscious states, many people explicitly reject the claim that insects can feel pain on the basis of facts about the limitations of insect neural systems. To take a different sort of example, as philosophers worrying about the problem of other minds, one might well come to doubt the philosophical arguments that others are conscious. This would be a case in which the high road to attributing conscious states to others does not yield the conclusion that others are conscious. However, even the skeptic about other minds will still have low-road reactions when humans swirl about him.

3. Dual-Processing and Explanatory Gap Intuitions

How exactly might our dual-process model explain the intuitive force of the explanatory gap? As sketched above, we maintain that third-person mind attribution involves two distinct cognitive processes whose outputs may either comport with or fail to comport with one another. When looking at other people, both of these systems typically produce affirmative outputs: the low road is activated by (at least) one of the other person's surface features, producing a low-level, intuitive attribution of consciousness. At the same time, we can use the high road to reason deliberately about another entity's being conscious (for instance, as John did in §2.3). Since the two systems generate the same answer in typical cases, there is typically no resistance to the idea that other people are conscious. However, when we consider the mass of grey matter that composes the human brain (and on which the majority of physicalist reductions of consciousness will focus), the result is altogether different.

Consider Jenny, who is in the grip of physicalism about consciousness. Using high road reasoning, she could apply the hypothesis that consciousness is identical to a certain kind of brain process, in which case Jenny's high road would produce the output that specific brain processes or brain regions *are* conscious experiences.[9] For example, Jenny might believe that consciousness is identical to populations of neurons firing in synchrony at a rate between 40Hz and 60Hz; on this basis she could infer (using the high road) that specific brain regions that are firing synchronously are conscious experiences. (Crick & Koch, 1990). If Jenny knew that Jimmy's brain had regions that were firing synchronously between 40-60Hz, she could infer (using the high road) that Jimmy's brain states are conscious experiences. But since this description of Jimmy's brain does not advert to any of the featural cues that trigger AGENCY categorization, Jenny's low road is not activated, and thus remains silent on whether the synchronously firing neurons are conscious.[10]

---

[9] We use the example of a "type-identity" theory of consciousness for ease of exposition. A similar point could be made using "token-identity" (or functionalist) theories, or other sorts of physicalist theories.

[10] Of course, if Jenny were to view a *picture* of Jimmy (or Jimmy himself), her low road would be activated by the presence of the relevant featural cues, and she would be disposed to attribute conscious states to Jimmy. But saying that Jimmy (the person) activates Jenny's low road is very different from saying that *Jimmy's brain* activates Jenny's low road.

This example, we think, helps to illuminate why physicalist explanations of consciousness leave us feeling as if something has been left out: our low-level, *low road* process remains silent where it would normally provide intuitive confirmation of our *high road* output.[11] In place of the harmony between systems that we typically experience when looking at other people (or any other mammal, for that matter), discussions of neurons, neurotransmitters, and so on create a disparity between the two systems, which in turn produces a feeling that the characterization is somehow incomplete.[12] This, we think, is an important part of the explanation for why dualism is so attractive and the explanatory gap is so vexing.[13]

---

[11] Of course, it happens quite often that high-road representations are not accompanied by any corresponding low-road representations. For example, I might use the high road to reason to the conclusion that $e = mc^2$, but there would be no corresponding low-road representations of energy, mass, or the speed of light. (Thanks to Josh Weisberg for the example). Does our theory predict a kind of gap in this case? No. Our theory only predicts these intuitions for cases in which the underlying cognitive architecture is configured for dual processing. In such cases, both high-road and low-road representations play a role in controlling behavior and inference. In cases that only involve *system 2* processing, system 2 is free to control inference and behavior unfettered. Thus it is only in cases involving dual processing that dissonance between *system 1* and *system 2* can arise. Thus, the case of consciousness is distinct from cases of pure *system 2* reasoning, because (we claim) it does involve dual processing.

[12] Our view here is anticipated in important ways by Philip Robbins and Tony Jack, who write: "The intuition of a gap concerning phenomenality [i.e., consciousness] stems at least in part from the fact that our brains are configured in such a way as to naturally promote a dualist view of consciousness." (2006, 75) However, Robbins & Jack's explanation of the explanatory gap keys on *moral capacities*. They write, "At the heart of our account of why consciousness seems to defy physical explanation is the idea that thinking about consciousness is essentially linked to, indeed partly constituted by, the capacity for moral cognition" (2006, 75). On our view, while moral cognition might be associated with conscious attribution, the order of explanation would be reversed. The AGENCY system is primitive and not directly a moral capacity. Yet, we suggest, the AGENCY mechanism provides the primitive basis for consciousness attribution.

Robbins & Jack discuss an objection to their view that is relevant here. The objection is that their view entails that people who lack basic moral capacities, like psychopaths, should fail to feel the explanatory gap (76-78). Robbins & Jack discuss various ways in which their theory can address the objection, but it's important to note here that our theory makes no such predictions. We expect that a creature might lack moral capacities while retaining the AGENCY mechanism and the associated attributions of conscious states.

[13] We intend for this explanation to apply specifically to intuitions about the *explanatory gap*, as opposed to other puzzling cases involving consciousness. This is worth mentioning because it is quite common for philosophers to advance *unified* explanations of the explanatory gap, zombie scenarios, the knowledge argument, and so forth. Our ambitions in this paper don't extend that far. We will be well satisfied if we manage to illuminate the source of the explanatory gap.

We are suggesting that disparate outputs from the two consciousness-attribution processes produce a sense that something isn't right. This proposal resembles some accounts of what is happening in Capgras syndrome, a psychological disorder in which patients think that a loved one has been replaced with a superficially similar duplicate. Davies & Coltheart (2000) describe Capgras as follows:

> Patients who suffer from the Capgras delusion believe that someone close to them, often a close relative, usually their spouse in the first instance, has been replaced by an imposter who looks just like the replaced person… Capgras patients sometimes elaborate their delusion to the extent of invoking some pieces of technology, perhaps robots or, in a biotechnological age, clones (p. 10).

The Capgras delusion is noteworthy both for its bizarre quality and for its relatively circumscribed nature. Typically, the delusion does not "spread out" through the afflicted subject's network of belief at large. For example, Capgras patients tend not to be especially interested in the whereabouts or well-being of their spouse, despite apparently holding the belief that their spouse has gone missing. But the delusion persists nonetheless.

This unusual syndrome demands explanation. On Stone and Young's prominent account, the Capgras delusion arises from an unusual yet persistent subjective experience, in which the purported imposter "looks right," yet does not "feel right" (Stone and Young 1997).[14] The unusual and persistent experience is supposed to give rise to the unusual and persistent delusion. It is hypothesized that in the typical case, our recognition of faces is supported by at least two distinct cognitive processes. One process identifies the morphology of the face and outputs a morphological representation, and another process outputs an affective response (e.g. a feeling of familiarity). In Capgras patients, the morphological process is intact and outputs normal morphological representations, but the process supporting the affective response is damaged and does not output any feeling of familiarity. Thus, Capgras patients who undergo the relevant experiences sometimes say things like, "She looks just like my wife, but I don't feel any love for her." It is easy to see how peculiar experiences like this could play a role in generating the delusion, even if they do not fully explain the syndrome. The point is that on the Stone & Young account, the Capgras delusion results, at least in part, from a breakdown in processing that involves a mismatch between the outputs of distinct processes. Normally, the morphological and affective mechanisms provide matching outputs. But in the pathological case, the output of the morphological process is not corroborated by any output from the affective process. Delusion results.

On our view, the explanatory gap works much like the foregoing account of the Capgras delusion. People's natural inclination to judge that broadly physicalistic accounts of consciousness "leave something out" depends on a cognitive architecture involving two distinct processes. In typical cases of consciousness attribution, the two processes produce harmonious outputs, and lead to unsurprising attributions. But in the

[14] Though it is controversial whether this sort of experience provides a *complete* explanation of the Capgras delusion, it is somewhat less controversial that experiences of this sort play a key role in the delusion.

case of the explanatory gap, we claim, one of the relevant cognitive processes fails to produce any output, thus leading to the disharmonious sense that the neural description is fundamentally incomplete as an explanation of consciousness.[15]

4. Objections & replies

Our proposal, while new, has already met with a number of objections. In this section, we deal with what we take to be the most important of objections we've encountered thus far.

4.1. Objection: What About Intentionality?

One natural objection is that if our proposed model predicts an explanatory gap for consciousness, then it must also predict an explanatory gap for "about-ness" or *intentionality*. In our view, the activation of AGENT leads to attributions of conscious states like pain, and also to intentional states like desires. Because attributions of conscious states and intentional states are supported by the same mechanisms, we should expect an explanatory gap for intentionality. Our model predicts that completely physicalistic explanations of intentionality will fail to trigger AGENT and consequently fail to elicit the normal pattern of gut reactions and intentionality-attributions, for reasons analogous to the case of consciousness. But, the objection continues, this prediction is problematic because while there is an explanatory gap for consciousness, there is none for intentionality. Our model predicts a gap where there is no gap. While consciousness is mysterious and problematic from the standpoint of physicalism, intentionality is relatively easy to locate in the physical world. Or so the objection goes.

For present purposes, we will simply grant the objector the claim that our model predicts that there should be an explanatory gap for some attributions of intentional states. However, it doesn't follow that *all* attributions of apparently intentional states will give rise to an explanatory gap. People routinely attribute apparently intentional states, such as memory and knowledge, to computers (cf. Robbins & Jack 2006, 78-79). For instance, it's perfectly natural to say that a chess program knows that the queen will be lost if it moves the pawn. More simply, it is familiar to say that that hard disks and flash drives have *memory*. These attributions do not come with any air of explanatory mystery. It's possible that we sometimes apply such computationally domesticated intentional attributions to humans as well. Nonetheless, this hardly excludes the possibility that some intentional attributions do indeed invite an explanatory gap. In fact, in one of the earliest apparent expressions of the explanatory gap, Leibniz seems to articulate an explanatory gap that folds the intentional and the conscious together:

---

[15] A key difference between the Capgras delusion and the explanatory gap involves the nature of the underlying processes. Our model appeals to standard dual-process architecture to explain the gap, whereas in Capgras, neither the morphological system nor the affective system is akin to System 2. But that doesn't diminish the thrust of the analogy. The critical point is that independent systems are involved, and they produce disparate outputs about the target domain where harmonious outputs are the norm.

> If we imagine that there is a machine whose structure makes it *think, sense, and have perceptions*, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters into a mill. Assuming that, when inspecting its interior, we will only find parts that push one another, and we will never find anything to explain a perception. (Leibniz, 1714/1989, sec. 17, emphasis added)

Nor is this view merely a curiosity of the 18th century. A number of prominent contemporary philosophers have quite explicitly defended an explanatory gap for intentional states (Cummins 2000, Horgan 2009, McGinn 1988, Rey 2009). Since it is very much a live philosophical question whether there is an explanatory gap for intentionality, we think the intentionality objection is far from decisive.[16]

4.2. Objection: The proposal mislocates the gap, part 1: Phenomenal concepts

It might be objected that our account doesn't illuminate the explanatory gap because the gap is really driven by the difference between the first-person properties that are involved in conscious experience, and the third-person properties adverted to by scientific theories of conscious experience.[17] To explain this objection we first need to review quickly how the apparent alternative goes. A property dualist might maintain that even if mental processes (or events, or things) are identical to physical processes (or events, or things), there still seems to be a distinctive class of mental properties that objective science cannot explain. Specifically, the subjective and qualitative properties of conscious experience seem to resist scientific explanation and reduction to the physical. Relatedly, physicalists (who reject the existence of inexplicable and irreducible subjective properties) may propose something similar at the level of concepts. Such physicalists hypothesize that we possess certain concepts – phenomenal concepts – that systematically fail to accord with the concepts deployed in objective physical science.[18] There is, of course, considerable disagreement about the precise nature of phenomenal concepts, and hence about the precise way in which phenomenal concepts fail to accord with physical concepts. Some theorists maintain that phenomenal concepts are *recognitional* concepts (Loar 1990/1997; Tye 2003); others maintain that they are *quotational* concepts (Block 2006; Papineau 2006); still others maintain that they are

---

[16] Many have thought that *consciousness* is the feature left out of reductive accounts of belief (e.g. Kriegel 2003, Searle 1991). This is, of course, consistent with the AGENCY model since that model proposes that identifying an entity as an AGENT will incline us to attribute both beliefs and conscious states.

[17] In his (1959), J.J.C. Smart attributes this objection to Max Black. Ned Block explicates and responds to this objection in his (2006).

[18] On many accounts of phenomenal concepts, the failure is supposed to be that no conclusion conceived under exclusively phenomenal concepts can be inferred *a priori* from any set of premises conceived under exclusively non-phenomenal concepts. The precise nature of the failure (for example, the reason why the relevant *a priori* inferences are supposed to fail) will depend upon the precise nature of phenomenal concepts.

indexical concepts (Ismael 1999, Perry 2001). Despite the disagreement about what it is to be a phenomenal concept, all these theorists adopt the basic strategy that the explanatory gap is a direct result of the discord between our phenomenal concepts and our physical concepts. Property dualists frame their explanation of the gap in terms of phenomenal properties (rather than concepts), but they can nonetheless agree with physicalists that explanatory gap arguments are intuitively compelling because they exploit a principled difference between phenomenal concepts and objective concepts. This difference results in a kind of 'conceptual gap,' which is supposed to be characteristic of the explanatory gap.

Now we can state the objection. Our model attempts to explain the gap without explicitly adverting to "phenomenal concepts". But since phenomenal concepts are the real source of the explanatory gap, the AGENCY model does not do any real explanatory work. That is, advocates of the phenomenal concept strategy might object that their theory explains the gap, so our theory is impotent.

There are different ways that the objection might be developed. We focus on what we take to be the most instructive version. Assume that there are phenomenal concepts and also that attributions facilitated by the AGENCY mechanism often involve phenomenal concepts. For example, it may be that the AGENCY mechanism normally triggers the phenomenal-concept PAIN *en route* to a pain attribution. In that case, we have no particular quarrel with the claim that phenomenal concepts play an important role in generating the explanatory gap. Our proposed model can in principle be combined with various accounts of phenomenal concepts, and the two sorts of accounts could potentially be seen as complementary. On this understanding, our model can be seen as spelling out certain conditions under which phenomenal concepts will or will not be deployed, without saying much about the phenomenal concepts themselves. Construed in this way, our model would enrich our understanding of the explanatory gap by enriching our understanding of some conditions for the activation of phenomenal concepts.

Alternatively, the AGENCY model itself may be understood as functionally characterizing some phenomenal concept(s). The model specifies the functional dynamics of a distinctive cognitive system that often leads to attributions of phenomenal states. So the model could be seen as explaining why a distinctive concept of consciousness plays a very different functional role than the concepts of consciousness deployed in objective science. On this understanding, the model would yield a distinctive account of phenomenal concepts.

Although our account is thus consistent with phenomenal concept approaches, we don't want to commit ourselves to any theses about phenomenal concepts or their role in generating the explanatory gap. For all we've said here, it remains possible that phenomenal concepts do not play any significant role in underwriting the plausibility of explanatory gap arguments.[19] As a result, even if the phenomenal concept strategy fails

[19] For example, phenomenal concepts do not figure in the accounts of materialists such as Dennett (1991) and Rey (1995). But such accounts seem to be broadly compatible with our AGENCY model.

(despite its present popularity), the AGENCY model can still contribute to a psychological explanation of the intuitive force of the explanatory gap. Thus, the AGENCY model is consistent with the phenomenal concept strategy, and it might be developed as a version of the strategy. But the AGENCY model is not hostage to the strategy.

4.3. Objection: The proposal mislocates the gap, part 2: The first person perspective

A related objection is that the source of the gap involves a difference between self-attributions and other-attributions of consciousness. The idea is that I appreciate the qualitative aspect of my pain *in my own case*, and no scientific description can provide a satisfying explanation of *my* pain experience. So, the problem gets off the ground because of something about self-attributions specifically. Since our proposal focuses primarily on other-attributions, it completely misses the problem of the explanatory gap.

Of course we agree that the explanatory gap can be made salient from the first-person perspective by focusing on one's own experiences. However, it would be somewhat myopic to think that *the* gap essentially involves first-person (or self-attributive) cases. For an explanatory gap presents itself even when we restrict our focus to third-person attributions (i.e. other-attributions). People find it quite intuitive to attribute consciousness to many third parties, including dogs, birds, and ants. Setting aside philosophers in their skeptical moods, people rarely look at horses, cats, or humans and think "How could *that thing* be conscious?" On the contrary, it is virtually automatic that we judge such organisms to have conscious states. However, just as when we reflect on our own conscious states, a "gappy" intuition surfaces when we turn to *specific kinds* of third-person characterizations of consciousness, namely scientific descriptions. People are happy to credit consciousness to cats, but it is counterintuitive that cat-consciousness is ultimately nothing more than populations of neurons firing synchronously at 40-60Hz. That is where our proposal enters the picture. We claim that the gap arises in part because such scientific descriptions do not trigger the low road to consciousness attribution.

Of course, we find a parallel situation when we focus solely on self-attributions of consciousness. When I compare my own conscious experience with scientific descriptions of my own brain, the neural features do not seem to fully explain my conscious experience; and they certainly don't seem to *be* my conscious experience. This intuition is generated (we suggest) because the neural description activates the high road but not the low road. By contrast, we don't get a 'gappy' intuition when viewing our own image in a mirror. We don't think, "Sure *I'm* conscious, but how can *that thing* in the mirror be conscious?" This, we submit, is because the mirror image *does* suffice to activate the low road to consciousness attribution. So the difference between self-attributions and other-attributions cannot by itself explain our 'gappy' intuitions about consciousness. Instead, the explanatory gap emerges at least in part from the contrast between cases in which *there is* intuitive support from the low road, and cases in which there *is not* intuitive support from the low road.

4.4. Objection: The Proposal is Overly General

We have argued that part of the explanation for the explanatory gap is that our gut-level feelings that an entity has conscious states are driven by a low-road process that is insensitive to the kinds of features that we find elaborated in neuro-functional descriptions of the brain. If that's right, then we should expect to find something similar to the explanatory gap in other domains, because dual-process architecture is supposed to be implicated in many domains. But, the objection goes, these expectations go unsatisfied because the explanatory gap phenomenon is restricted to the domain of conscious experience.

One response to this objection is that for all we've said here, consciousness might be the only philosophically important domain in which an explanatory gap obtains. It's certainly possible that the cognitive systems underlying other philosophically important domains do not employ the kind of dual-process architecture that we think drives explanatory gap intuitions. It's also possible that such systems *do* have a dual-process architecture, yet never produce 'gappy' intuitions because dual-process architecture is not sufficient for generating an explanatory gap. After all, in some cases, the two systems might produce harmonious outputs, rather than the disharmony we find in certain attributions of consciousness. So even if our dual-process account is right for the explanatory gap for consciousness, it might turn out to be singular.

That said, we rather suspect that something like the explanatory gap phenomenon does show up in other cases where we try to reductively analyze intuitive notions. Take causation, for instance. There is good reason to think that we have a low-road process that generates gut-level intuitions about causation. Infancy research suggests that babies are especially sensitive to cues like *contact* (Leslie & Keeble 1987). Seeing a stationary object launch after being contacted by another object generates a powerful and intuitive sense of causation. Work on adults brings this out vividly. In a classic experiment, Schlottman & Shanks (1992) showed adult subjects computer-generated visual scenes with two brightly colored squares, A and B. The participants were told that the computer might be programmed so that movement would only occur if there was a color change; participants were told to figure out whether this pattern held. In the critical condition, every movement was indeed preceded immediately by a color change. On half of the scenes, there was no 'contact' between A and B, but B would change color and then move; on the other half of the scenes, there was contact between A and B just before B changed color and then moved. Importantly, the covariation evidence indicates that color change was necessary for motion. And indeed, the participants' explicit judgments reflected an appreciation of this. But these explicit judgments had no discernable effect on their answers to the questions about *perceived* causation in launching events, viz., "does it really seem as if the first object caused the second one to move? Or does it look more as if the second object moved on its own, independent of the first object's approach" (Schlottman & Shanks 1992, 335). Only when there was contact did people tend to say that it "really seemed" as if the first object caused the second object to move. This gut-level sense of causation seems to be driven by a low-road system that is insensitive to covariation information.

When we turn to reductive philosophical explanations of causation, many such accounts seem intuitively incomplete and unsatisfying. For example, Lewis's counterfactual account has absorbed criticism along these lines (Lewis 1973; cf. Menzies 1996, Schaffer 2001). Very crudely, the account claims that C causes E if and only if E wouldn't have happened if C hadn't happened. It is not just that such accounts are counterintuitive, but that they are counterintuitive in a specific way: counterfactual accounts seem to leave the 'oomph' out of causation.[20] Whereas physicalist theories of consciousness seem to be missing *what it's like* to be conscious, counterfactual theories of causation seem to be missing causal *oomph*. It is, we think, an intriguing and promising research question whether this intuitive shortcoming might be illuminated by the considerations we have marshaled here for the explanatory gap. That is, it might be that part of the reason that many reductive explanations of causation are intuitively unsatisfying is the failure of such explanations to trigger the low-road processes that generate the gut-level sense that A caused B.

In light of this work on causation, we take the proposed objection to raise a genuinely interesting possibility for future research. Rather than think of this as an objection to our proposal, we take it to be an invitation to investigate whether we can use the dual-process framework to explain the intuitive shortcomings of reductive analyses in other philosophical domains.

5. Implications

We have argued for a partial explanation of the fact that we find physicalist explanations of consciousness deeply counterintuitive: deliberate reasoning about neural and other physical activity does not activate the cognitive system that generates the gut-level feeling that an entity is conscious. As a result, thinking about neural tissue does not trigger an intuitive sense that the tissue is conscious. If this much is correct, then what are the implications for philosophy? These are treacherous intellectual waters, but we will set sail (whether bravely or foolishly) and sketch one way that our account might be used to elaborate an important strand of a physicalist defense against dualist arguments.

As we discussed at the beginning of the paper, one important impetus to dualism is the fact that it simply seems bizarre to think that conscious experience is nothing over and above brain activity. The fact that physicalism is counterintuitive, we have suggested, also plays an important role in driving the explanatory gap arguments in philosophy. A standard way of deflating the philosophical import of the counterintuitive aspects of physicalism (including the explanatory gap) is to point out that the view seems counterintuitive in virtue of contingent psychological facts about us. Thus, the fact that we find it difficult to wrap our heads around the idea that conscious states are neural states is not a decisive reason for drawing the metaphysical conclusion that conscious states really are not physical states.

[20] Along similar lines, Lewis uses the term "biff" to describe an intrinsic, non-counterfactual relation in the vicinity of the causal relation (2004).

Our present proposal might play a significant role in filling out such an argument by offering a more detailed empirical account of the psychological mechanisms that drive our intuitive resistance to physicalism. To determine how much philosophical weight we should give to our intuitive resistance to physicalism, we would do well to know a good deal about the psychological basis for that resistance. Our proposal is that the resistance is caused partly by the fact that the low-road mechanism will not render a confirmatory gut-feeling to our considered reasons for thinking that conscious states are brain states. A further question is whether we should take that low-road system to carry any epistemic weight, and if so how much weight. Answering this question involves confronting difficult epistemic issues, and we won't presume to do them justice here. But at a minimum, we think there is reason to take a skeptical stance toward the low road's epistemic credentials.

One suggestion is that we should discount the low-road system simply because it is relatively crude and inflexible. By contrast, our reasoned judgments about consciousness are highly flexible and general, and might be thought to be more trustworthy than the low-road mechanism because they take more information into account.[21] This kind of consideration is clearly not decisive, however, because it's plausible that we are often justified in trusting the outputs of relatively crude and inflexible cognitive systems (low-level vision, for example).

Another possibility is that this particular low-road mechanism is untrustworthy, even if there is little reason to doubt the outputs of low-road mechanisms in general. It is highly plausible that a low-road mechanism for detecting other minds (and other conscious minds) would be subject to a high rate of false positives. Considerations from signal-detection theory and evolutionary psychology support this claim. Consider, for example, the high cost of a false negative. Failing to detect another conscious agent could have potentially disastrous consequences: a rival human or (worse) a hungry predator could easily get the jump on the poor sap whose low-road mechanism outputs a false negative. Since an easy way of producing fewer false negatives is to produce more false positives, this is what we should expect the mechanism to do. And indeed, it seems plausible that the low-road mechanism does in fact produce many false positives. The Heider-Simmel illusion seems to provide an obvious case in which our intuitive attributions of mentality are misguided: animated cartoons and movies provide a range of similarly clear examples. In these kinds of cases, it is extremely plausible to think that the low-road mechanism has produced inaccurate outputs.

But what about false negatives? False negatives are more directly relevant to the explanatory gap, because (we claim) the gap is a case in which the low-road mechanism is silent. It's worth noting that even mechanisms with a high rate of false positives may sometimes output false negatives. For example, we might expect a "snake detector" mechanism to have a high rate of false positives, for reasons similar to those given above. But such a mechanism may occasionally fail to detect a snake: the snake might be camouflaged, or irregularly shaped, or seen from a non-standard vantage point. In such

[21] Haidt (2001) and Greene (2003, 2008) reason along these lines, for the conclusions that our reasoned moral judgments are more trustworthy than our intuitive moral judgments.

cases the snake-detector would remain silent. Could our proposed low-road mechanism for consciousness-attribution be similar to the snake-detector in this respect? It is difficult to say, because in the case of snakes we can appeal to an independent and relatively uncontroversial standard about which things count as snakes. But in the case of consciousness there is no such independent standard, since there are core philosophical and scientific disputes about the nature and scope of consciousness. So it seems doubtful whether this kind of consideration could yield a decisive reason for saying that the low-road mechanism is untrustworthy in the relevant cases.

Nonetheless, we think there is reason to handle the low-road mechanism's outputs (or lack thereof) with extreme care. While the low road is routinely triggered by biological organisms, it is rarely or never triggered by the *brains* of those organisms: and according to some of our best theories, the brain is the part of the organism most crucially responsible for its mind. That is, we have reason to suspect that the low-road mechanism is insensitive to some of the features that are most important for mindedness.

One of the salient features of the low-road AGENCY mechanism is that it is responsive to *organisms,* and not to particular bits inside of organisms. There is an obvious explanation for this. The low road is, in some fashion, an adaptation to our environment. It might be a domain-specific mechanism that was shaped by evolutionary pressures. Or it might be a developmental adaptation that children achieve through countless interactions with their environment. We take no stand on that issue here. Regardless of which kind of adaptation it is, the AGENCY mechanism was shaped by the environment to which we (or our evolutionary ancestors) were exposed. As a result, it is unsurprising that the mechanism responds to organisms but not to suborganismic bits. We (and our ancestors) interacted most often with entire organisms, not neurons in a petri dish. Once we see the role of the environment in shaping the mechanism, this should lead us to suspect that the low-road mechanism is a relatively shallow and inflexible informant for a theory of consciousness. The mechanism is sensitive only to gross organismic features, but we need not suppose that this is because consciousness *only* attaches to gross organisms. Rather, the reason the low road mechanism is sensitive to such a restricted set of features is because whole organisms are the parts of the environment that are responsible for shaping the mechanism. Suborganismic features like neuronal firing patterns *never had a chance* to shape the mechanism, because they are hidden away behind skin and bone. So even if these features *are* crucially important for consciousness, we should still expect our low-road mechanism to be insensitive to this fact. As a result, when considering explanations of consciousness, there is reason to doubt that we can assign much evidential weight to the fact that the low-road isn't activated by suborganismic features. The fact that the low-road is silent cannot be taken as significant evidence that consciousness is something other than suborganismic a feature.

Even on the supposition that the proposed low-road mechanism is not to be trusted in the relevant cases, we do not claim to have provided a *complete* psychological or epistemological account of the explanatory gap. For example, more must be said about the psychology and epistemology of attributions of particular kinds of conscious states (e.g. reddish visual experience *versus* blueish visual experience). At least one

author suspects that cognitive systems aimed specifically at processing explanations will play an important role in an account of the explanatory gap.[22] And a range of other cognitive mechanisms may also be involved (our capacities for imagination and visualization, for example). Nonetheless, we think that our present proposal makes a significant contribution to an account of the explanatory gap. Part of the reason we feel the gap, and part of the reason we are seduced by dualism, is that scientific explanations do not resonate with a basic cognitive system that generates the intuition that something is conscious.

[22] Fiala (2009) argues along these lines.

References

Arico, A., Fiala, B., Goldberg, R. & Nichols, S. (forthcoming). "The Folk Psychology of Consciousness."

Block, N. (2006). "Max Black's Objection to Mind-Body Identity." In *Phenomenal Concepts and Phenomenal Knowledge*, T. Alter and S. Walter eds. Oxford University Press: 249-306.

Bloom, P. (2004). *Descartes' Baby*. New York: Basic Books.

Bloom, P. (2006). "My Brain Made Me Do It." *Journal of Culture and Cognition*, 6, 209-214.

Chalmers, D. (1995). "Facing Up to the Hard Problem of Consciousness." *Journal of Consciousness Studies* 2(3): 200-219.

Chalmers, D. (2003). "Consciousness and its Place in Nature." In *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers ed. Oxford University Press: 247-272.

Chalmers, D. (2006). "Phenomenal Concepts and the Explanatory Gap." In *Phenomenal Concepts and Phenomenal Knowledge*, T. Alter and S. Walter eds. Oxford University Press: 167-194.

Churchland, P. (1988). *Matter and Consciousness*. MIT Press.

Crick, F. & C. Koch. (1990). "Toward a Neurobiological Theory of Consciousness." *Seminars in the Neurosciences*, 2: 263-275.

Cummins, R. (2000). "'How does It Work?' vs. 'What are the Laws?' Two Conceptions of Psychological Explanation." In F. Keil and R. Wilson (eds.), *Explanation and Cognition*, MIT Press, 117-145.

Dennett, D. 1991. *Consciousness Explained*. Little-Brown.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. Hove: Psychology Press

Fiala, B. (2009). "The Phenomenology of Explanation and the Explanation of Phenomenology." MS, University of Arizona.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Graham, G. (1998). *Philosophy of Mind: An Introduction*. Oxford: Blackwell.

Gray, H., K. Gray, and D. Wegner (2007). "Dimensions of Mind Perception." *Science*, vol. 315, p. 619.

Greene, J. (2008). "The Secret Joke of Kant's Soul." In W. Sinnott-Armstrong, *Moral Psychology*. MIT Press.

Greene, J. (2003). "From Neural 'Is' to Moral 'Ought': What are the Implications of a Neuroscientific Moral Psychology?" *Nature Reviews Neuroscience*, 4: 389-400.

Haidt, J . (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108, 814-834.

Heider, F. and Simmel, M. (1944). "An Experimental Study of Apparent Behavior." *American Journal of Psychology* 57: 243-259.

Horgan, T. (2009). "Materialism, Minimal Emergentism, and the Hard Problem of Consciousness." In G. Bealer and R. Koons (eds), *The Waning of Materialism*. Oxford University Press.

Huxley, T.H. (1866). *Lessons in Elementary Physiology* 8. London: MacMillan.

Huxley, T.H. (1874/2002). "On the Hypothesis that Animals are Automata, and Its History (Excerpt)." In D. Chalmers ed., *Philosophy of Mind: Classical and Contemporary Readings*, Oxford University Press. Excerpted from *Fortnightly Review* 16: 555-580.

Ismael, J. (1999). "Science and the Phenomenal." *Philosophy of Science* 66:351-69.

Jackson, F. (2000). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

Johnson, S., Slaughter, V. and Carey, S. (1998). "Whose Gaze will infants follow? Features that Elicit Gaze-following in 12-month-olds. *Developmental Science* 1: 233-238.

Johnson, S. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London,* B, 358, 549–59

Kim, J. (2000). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.

Kim, J. (2007). *Physicalism, or Something Near Enough*. Princeton University Press.

Knobe, J. & J. Prinz (2008). "Intuitions about Consciousness: Experimental Studies." *Phenomenology and the Cognitive Sciences* 7(1): 67-83.

Kriegel, U. (2003). "Is Intentionality Dependent Upon Consciousness?" *Philosophical Studies* 116: 271-307.

Kripke, S. (1972). "Naming and Necessity," in *Semantics of Natural Language,* D. Davidson, G. Harman, Dordrecht & Holland eds. Reidel: 253-355.

Leibniz, G.W. (1714/1989). *The Monadology*. In *G. W. Leibniz: Philosophical Essays*, R. Ariew and D. Garber, eds. and trans. Indianapolis: Hackett Publishing Company.

Leslie, A. and Keeble, S. (1987). "Do Six-month-old Infants Perceive Causality?" *Cognition* 25, 265–288.

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

Lewis, D. (1973). "Causation" *Journal of Philosophy* 70:556-67.

Lewis, D. (2004). "Void and Object." In *Causation and Counterfactuals*, J. Collins, N. Hall & L.A. Paul eds. MIT Press: 277-290.

Loar, B. (1990/1997). "Phenomenal States." In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, J. Tomberlain ed. Ridgeview: 81-108. Revised version in *The Nature of Consciousness*, N. Block and G. Guzeldere eds. MIT Press: 597-616.

McGinn, C. (1988). "Consciousness and Content." *Proceedings of the British Academy* 74:219-39.

Mill, J.S. (1865). *An Examination of Sir William Hamilton's Philosophy*. London: Longmans.

Nagel, T. (1972). "What Is It Like to Be a Bat?" *The Philosophical Review* 83(4): 435-450.

Papineau, D. 2006. "Phenomenal Concepts and the Materialist Constraint." In T. Alter and S. Watler (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.

Rey, G. 1995. "Toward a Projectivist Account of Conscious Experience." In T. Metzger (ed.), *Conscious Experience*. Ferdinand Schoningh.

Rey, G. 2009. "Philosophy of Mind." *WIRES: Cognitive Science*.

Richert, R. and Harris, P. (2006). "The Ghost in My Body: Children's Developing Concept of the Soul." *Journal of Cognition and Culture*, 6, 409 – 427.

Richert, R. and Harris, P. (2008). "Dualism Revisited: Body vs. Mind vs. Soul." *Journal of Cognition and Culture*, 8, 99 – 115 .

Robbins, P. and Jack, A. (2006). "The Phenomenal Stance." *Philosophical Studies*,127.

Samuels, R. (forthcoming) "The Magical Number Two, Plus or Minus: Comments on Dual Systems" In Evans, J. & Frankish, K. (eds.) In *Two Minds: Dual Processes and Beyond*, OUP.

Schlottmann, A., & Shanks, D. (1992). "Evidence for a distance between judged and perceived causality." *Quarterly Journal of Experimental Psychology*, 44A, 321 - 342.

Searle, J. (1991). "Consciousness, Unconsciousness and Intentionality." *Philosophical Issues* 1: 45-66.

Shimizu, Y. and Johnson, S. (2004). "Infants' Attribution of a Goal to a Morphologically Unfamiliar Agent." *Developmental Science* 7: 425-30.

Smart, J.J.C. 1959. "Sensations and Brain Processes." *Philosophical Review*, 68: 141-156.

Stanovich, K.E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. The University of Chicago Press.

Stanovich, K. E. and West, R. F. (2000). "Individual Differences in Reasoning: Implications for the Rationality Debate?" *Behavioral and Brain Sciences*, 23(5).

Stone, T., and Young, A.W. (1997). "Delusions and Brain Injury: The Philosophy and Psychology of Belief." *Mind and Language* 12: 327-364.

Sytsma, J and Machery, E. (2009). "How to Study Folk Intuitions about Consciousness." *Philosophical Psychology* 22(1): 21-35.

Tye, M. 2003. "A Theory of Phenomenal Concepts." In A. O'Hear (ed.), *Minds and Persons*. Cambridge University Press.

Woodward, A. (1998). "Infants Selectively Encode the Goal Object of an Actor's Reach." *Cognition* 69: 1-34.