The final version of this paper will appear in
Philosophy, Theory, and Practice in Biology (2019) 11:1
doi:10.3998/ptpbio.16039257.0011.001

Evolving Across the Explanatory Gap

Peter Godfrey-Smith
University of Sydney

1. Introduction

One way to express the most persistent part of the mind-body problem is to say that there is an "explanatory gap" between the physical and the mental. The gap is not usually taken to apply to all of the mental, but to subjective experience, the mind's "qualitative" features, or what is now referred to as "phenomenal consciousness." The "gap" formulation is due to Joseph Levine (1983). He acknowledged the appeal of intuitions of separability between physical facts, of any kind we can envisage, and this aspect of our mental lives. Subjective experience seems not to be the sort of thing that *is just* the physical under another guise. The immediate focus of Levine's discussion was a family of arguments due to Saul Kripke (1972). Kripke argued that some influential claims of mind-body identity could not be, as materialists claimed, contingent. If these identities are real then they are necessary. But we can clearly conceive that the mental and physical *could* come apart – an intuition that "identity theorists" conceded. Given that the identity

is necessarily present if present at all, from the fact it is not necessary we can see it is absent. Kripke concluded that physicalism is false.[1] Levine wanted to resist this:

> I find this intuition [of separability of mental and physical] important, not least because of its stubborn resistance to philosophical dissolution. But I don't believe this intuition supports the metaphysical thesis Kripke defends – namely, that psycho-physical identity statements must be false. Rather, I think it supports a closely related epistemological thesis ... namely, that psycho-physical identity statements leave a significant *explanatory gap,* and, as a corollary, that we don't have any way of determining exactly which psycho-physical identity statements are true. One cannot conclude from my version of the argument that materialism is false, which makes my version a weaker attack than Kripke's. Nevertheless, it does, if correct, constitute a problem for materialism, and one that I think better captures the uneasiness many philosophers feel regarding that doctrine. (1983, p. 354)

Levine could not see how the gap could be closed without taking physicalism in the direction of eliminativism. That is one non-dualist response. Another is to revise our view of the physical. This is seen in recent moves made by some philosophers towards panpsychism (Strawson 2006, Chalmers 2015). Unlike Levine, they hold tight to the qualitative as it appears, and reshape our view of the physical.

Here I engage with Levine's challenge and argue that the gap can be closed, using different resources from those cited above. I don't see this paper as completing the task, but as narrowing the gap and reshaping it. A full solution will include both a critical diagnosis of intuitions that shape people's view of the problem, and new pieces of positive theory.[2] Neither suffices on its own. In this paper I use a particular set of concepts to take us some distance on both the critical and positive sides.

The idea I pursue is that the bridging of the gap is achieved by way of the notion of *subjectivity.* Speaking roughly, *subjects* are one kind of evolutionary product. The history of life includes the history of subjectivity, and subjective experience is the experience *of a subject*. The notion of a subject's point of view is also an important

---

[1] For defences of the "identity theory" see Place (1956), Smart (1959), Lewis (1966), Armstrong (1968), and also Feigl (1958), whose view is discussed below. Kripke's discussion cited Armstrong and Lewis.

[2] Some other parts of my view are developed in Godfrey-Smith (2016a, 2016c).

resource on the critical side, especially in defusing arguments against materialism.[3]

When framing his problem, Levine asked for an explanation of why a particular physical arrangement gives rise to particular *qualia*. It is common now to see him as asking a question about consciousness, or "phenomenal consciousness," and the term "subjective experience" is sometimes used interchangeably with these. I think the last of these terms provides the best framing, but my view should not be seen as dependent on this set-up. Levine also presented his gap through a demand for an explanation of why the feel of seeing *red*, for example, goes with *this* brain state. I don't offer explanations of this kind – those explanations are a task for neurobiology. Instead, I try to make progress on the question of why it should be that some physical systems have subjective experience of *some* sort, while many others presumably do not, and I take steps to indicate a plausible general shape for explanations that fulfill Levine's request.

The first part of this paper develops the critical side. A number of ideas seen in existing defenses of materialism can be unified and improved through more explicit attention to subjects and their points of view. Those ideas on the critical side take for granted that there *are* such things as subjects with points of view. The second part of the paper then outlines a theory of the evolution of subjectivity. It takes steps towards treating subjectivity as a biological phenomenon. The work this paper does in bridging the explanatory gap comes from this dual use of the notion of a subject – as a critical tool and a target of explanation.

2. Subjectivity and its Diagnostic Role

The concepts of a subject and subjectivity are probably not part of "folk" psychology, but I take them to be broad and informal concepts, not tied to any particular philosophical theory, that pick out a cluster of features associated with mentality and experience. To be a subject, in this first informal sense, is to be a bearer or possessor of mental states,

---

[3] I make no distinction here between "physicalist" and "materialist" views, preferring the latter term for historical reasons (as Lewis 1994 does). Both terms have problems. A modern version of "materialism" is as much about energy as matter, and "physicalism" suggests a relationship to a field of study rather than a kind of entity. The substantive, as opposed to terminological, side of this question is discussed briefly in section 4.

though the term gestures more towards some mental states than others. It gestures towards the experiential side of the mind, to sensings, seemings, and feelings rather than more "cognitive" states, such as plans, and states linked closely to action, such as intentions.

The "something it's like..." formulation, often used in discussions of qualia, draws on the notion of a subject (Nagel 1974). If there's something it's like to *be* a particular entity, then that entity is a subject of experience. I take this to mean: there's something it *feels like* to be that entity. Like Nagel, I see these features as the right way in to an understanding of the specialness of the mental. Nagel himself is pessimistic about the possibility of a materialist account of the mind, and often expresses the challenge in terms of puzzles around subjectivity: "The main question, how anything in the world can have a subjective point of view, remains unanswered" (1986, p. 30).[4]

This phrasing poses the problem in a way that is not encumbered with baggage of other kinds. The formulations that do bring with them extra baggage are especially those that encourage a reification of sensations as objects of experience (the "act-object" view of perceptual experience). Those formulations encourage a project of trying to fit some special *entities,* to which subjects are related, into a physical inventory. Instead, talk of "qualia" and the like should be seen as talk of *ways* that subjects undergo experience, or as features of psychological states and processes. Views of that kind can be defended independently of a commitment to physicalism (Nagel 1965). They have usually been developed in "adverbial" terms, and in this form they have been criticized for their handling of complex experiences in which many properties are involved (experience of a red square *and* a green circle: Jackson 1976). Tye (1984) formulated a version of adverbialism that evades this problem (the "structured operator" view), though he no longer endorses that account, as he has adopted a representationalist view of perception

---

[4] I don't agree, though, with another attempt to place subjectivity at the center of the problem of materialist explanations of consciousness. Kriegel (2005) argues that explaining the subjective character of experience requires explaining "how a mental state may include within it an awareness of itself" (p. 30) – how such a state can represent a feature of the environment and itself at once. If this was required, it would indeed make the problem more difficult, but it's enough to explain how some systems are experiencing subjects with points of view, while others are not.

that is thought to make adverbial moves unnecessary. Without taking a position on representationalism, I work here within the family of views that includes adverbialism: ascription of properties to a sensation should be seen as part of the specification of a psychological state ascribed to a subject, or perhaps specification of a psychological process or event.[5]

A notion closely related to subjectivity is that of a *point of view* (as in Nagel above). This I also see as a useful bridging concept. But mere invocation of the idea of point of view does not take us far. "To have a point of view" in the most minimal sense is easy. Even a rock might be said to have one – it has a situatedness in its environment, and sensitivity to some events rather than others. This does not make the rock an experiencing subject; there are descriptions of point-of-view phenomena that are inadequate to the question. Below I will discuss, in a biological framework, what gives rise to points of view in a richer sense. First, though, I will take the existence of subjects for granted and look at the role of this concept on the critical and diagnostic side. I'll do this by first thinking about subjects and their points of view in general, and then move to debates about materialism.

Each of us has, at each waking moment, a point of view on things other than ourselves. Our attempts to describe, represent, and understand things are activities we engage in *as* subjects. In these attempts to understand the world we make use of representational tools such as words – and pictures, maps, and other media – and we have points of view *on* those representational tools.

"Point of view" talk, in its paradigm cases, is visual. Given where we are, an object looks a certain way to us. This usage can be extended to other senses and to features of a subject other than location. From there, "point of view" talk becomes a shorthand, often metaphorical, for all sorts of talk of relations between a subject and

---

[5] One justification for proceeding this way is as follows. Suppose that a biological, materialist treatment of subjective experience that had this "state-based" framing was generally adequate – adequate in most cases, even if, contra Tye, queries remained over "many property" cases of the sort used by Jackson. That is, suppose that we had an account of why, in a large class of relatively simple cases, it feels a certain way to be a subject of certain kind. I suggest that this would be *such* a step forward, such a removal of mystery, that it would be reasonable to expect a treatment of the highly structured cases to follow, probably in a way derivative on the state-based account.

subject-matter. "View" also has a broad, semi-metaphorical or dead-metaphorical sense in which it can refer to any claim or belief.

Some claims we make, and some representations, *express* a point of view, while others do not. If someone says, "The room looked full from where I was," their statement makes explicit reference to a point of view, and its truth depends not on the room's actual fullness, but on how it looked from a particular location. Other claims do not express a point of view: "Ants are animals," "Cody is in Wyoming." A claim like that will *reflect* the point of view of the person or persons who put it together – it will reflect what evidence they had, and what they saw as worth saying. But those are etiological facts, not facts about content; no point of view figures in the truth condition.[6]

It's common to talk of a "third person point of view," but there is really no such thing. Points of view are things that subjects, first-persons, have. The phrase "first person point of view" is somewhat redundant – this is the only kind of point of view. The distinction being made with the language of first and third does gesture towards something real. To "represent something from the third-person point of view" is to represent it in way that does not express the point of view of any particular subject. But there is no viewpoint-like entity – "the third person point of view" – that the claim is associated with, in the way that various other claims or representations *do* express a point of view, yours or someone else's.

Subjects can also be *objects* of knowledge and description. Each of us is a subject surrounded by other subjects – other people, animals. When we think about other people, we do so from our point of view as subjects. A paradigm case for these discussions is a situation where one person, *A,* has their own experiences and a first person point of view on things, and person *B* has a (first person) point of view on *A*. Person *B* watches *A* as they do things that suggest that they, too, are a subject. *B* might also look inside *A*'s brain, and might formulate theories of what is going on inside *A* and other subjects. Though people often say "we have a third person point of view" on other subjects and their

---

[6] Familiar claims like this one from Burge – "All representation is necessarily from some perspective or standpoint" (2010, p. 50) – do not distinguish between the inevitable way that representations *reflect* perspectives (in the evidence behind them, in what is seen as worth saying) and the narrower phenomenon in which some representations *express* a perspective.

mental attributes, in fact we have a ordinary first-person point of view on them – as objects of our experience, memories, and descriptions. When we come up with theories of what might be inside subjects and how they do what they do, however, we will often use language that does not *express* a particular point of view. Intersubjectivity is a familiar goal of scientific description.

There's a resulting divide between what a subject experiences and any scientific description of that subject that is not something to be overcome, but to be understood as an inevitable and ontologically innocent product of the role played by points of view. A description of how someone's brain works cannot *encapsulate* an experience had by that subject. (The same would be true of a description of how souls work, if they existed.) The "divide" resulting from the fact that the experience of taking an external observer's stance on a subject is different from the experience of that subject itself does not indicate that something is missing from a physical inventory of the world, does not indicate a need to add something that might bridge the gap.

I'll now start to bring these ideas into contact with debates about materialism, beginning with "knowledge arguments," and then moving to arguments based on the conceivability of separations between mental and physical. In both cases I make use of replies developed by others, and my aim is to supplement and unify them.

Some mental states feel like something to be in them. Further, it is often easy (for a normal adult human) to take a step back and engage in reflection *on* some of those states – easy to attend to them and think about them. This fact affects both the ongoing feel of experience (as our reflection *on* experience feels like something), and various cognitive capacities. Some forms of epistemic engagement with experience depend on the having of, or capacity to have, particular experiences. These include:

- *Remembering* experiences – recreating them with a form of episodic memory, and perhaps then assessing them: "I remember how good I felt when I heard that...,"
- *Comparing* experiences – "I felt better when *X* happened than when *Y* happened,"
- *Imagining* experiences – which might have an important role in remembering,
- *Attending* to them as they happen, thereby making it possible to do better at things like remembering and comparing them.

Some of these activities include making truth-evaluable claims, and some probably do not. Herbert Feigl (1958) used the phrase "knowledge by acquaintance" for forms of knowledge that depend on the having of experiences of a certain kind. That phrase may have been an unfortunate choice, as it suggests old-fashionedness and infallibility. But the idea is fine (and it would be worth coining a new term).[7] There are things we can do, with a cognitive or epistemic character, that depend on our special viewpoint on the experiences we have.

Feigl (1958) considered several scenarios where *theorists about subjects* lack a first-person point of view on some aspect of the mental, but nonetheless try to understand it. Suppose that Martians, who lack emotions, tried to understand human emotions. Or suppose that "all of mankind had been completely blind up to a certain point in history, and then acquired vision." (p. 416). And suppose also (perhaps oddly) that in our blind state we studied how vision would work, if we had it. In that situation we would in principle be able to foretell all the discriminatory and linguistic behavior which would result from vision, but there are some kinds of cognitive engagement with visual experience that we could not have. In particular, we could not "*imagine*" these experiences, and could not "*recognize* (or *label*) them as 'red,' 'green,' etc." (also p. 416).

Is there a kind of *knowledge* the blind scientists lack? Yes, Feigl says, they lack certain kinds of knowledge by acquaintance. Feigl embraces this as part of his physicalism. A world of material subjects, who can have experiences and turn those experiences to epistemic ends, and who formulate theories using external representational media and constrain these theories with intersubjectively available evidence, will be a world in which scientific descriptions of subjects lack certain roles.

Frank Jackson later used thought-experiments of this kind *against* materialism (1982). His imagined "Mary" is a vision scientist with complete knowledge of the physical facts about vision, initially confined to a room where all objects are black and white. Jackson argued that when Mary is released and sees colored objects, she learns

[7] Perry (2001) also endorses this concept as Feigl uses it, but not Feigl's terminology.

"something about the world and our visual experience of it" (p. 130), so the knowledge of vision she had before her release was not complete.

What was true of Mary before she has left the room? I'll start with things that everyone agrees on. First, Mary has not (yet) *had certain things happen* to her. She has not had some particular experiences. So she has no memory traces of a certain kind. Reliving those memories would have a qualitative character, related to the qualitative character of seeing itself.[8] Further, as noted by Feigl, she is not able to cognitively engage with color experiences in the way people outside the room can; she cannot recognize or label them.

To say these things is not yet to say whether she *learns something new* when she leaves the room. In response to that question, the literature since Jackson's paper has developed several different ways of expressing a similar reply. You can say that Mary knows all the facts about color vision, via one form of knowledge, before she leaves her room and later gains access to some of those facts with another kind of knowledge. Instead perhaps, or as well, Mary gains some know-*how*. Like Lewis (1994), I think these replies, and some others, are essentially similar.[9]

Jackson himself has come to endorse a reply of the same kind: "knowing what it is like to sense red can only be something about the new kind of representational state she is in, and the obvious candidates for that 'something about' are her ability to recognize,

---

[8] It can be objected that *false* memory traces of seeing red things would be enough for her to acquire the recognitional abilities (etc.) that Feigl cites. She need not actually have had those experiences. In reply, I say that outside far-fetched scenarios involving unusual neural interventions, though you might be wrong about particular events when you had an experience of red, you have to have had *some* experiences of that sort, in order to have a memory trace of the relevant kind.

The fact that within far-fetched scenarios it *would* apparently be possible to have a memory trace of the right kind despite no visual experiences at all of red objects puts additional pressure on Jackson's argument. If there is a way of creating a memory trace of the right kind from scratch, Mary with her complete factual knowledge would know how to do that, and would not need to leave the room. (This is a close relative of points made in Dennett 2004).

[9] The former is Feigl's preferred formulation: "what one person... *knows* by acquaintance may be identical with what someone else knows by description" (1958, p. 435). This reply need not be presented as a claim about a special family of "phenomenal" concepts.

imagine and remember the state" (2003, p. 271). This appeal to recognition and imagination is close to what Feigl made use of in the context of his thought-experiment in 1958.[10] Jackson says that this view needs to be presented within a representational view of perceptual experience, but the idea that experiences themselves have representational content is not needed to get there, and is a further commitment. If one *is* committed to representationalism about perceptual experience, then the capacities that Feigl and Jackson list (imagining, recognizing, etc.) will be consequences of the new representational states that Mary eventually acquires. If representationalism is denied, Mary can still remember, compare, and imagine experiences in new ways after she leaves the room.

The view that discussions have converged on recognizes three elements in the situation: (1) Subjects *have* experiences; (2) subjects *engage cognitively* with their own experiences (through memory, imagination, and immediate labeling); and (3) subjects treat *other* subjects and their experiences as objects of investigation and description, with the aid of intersubjectively deployable language. The differences between these,

---

[10] Feigl's discussion has been underappreciated for a long time, with Perry's 2001 discussion the main exception. It is worth looking at the crucial passages in Feigl and Jackson side by side:

Jackson 2003: "knowing what it is like to sense red can only be something about the new kind of representational state she is in, and the obvious candidates for that 'something about' are her ability to recognise, imagine and remember the state" (p. 271).

Feigl 1958: "What is it then that we would not or could not know...? I think the answer is obvious. We would not and could not know (then) the color experiences by *acquaintance*; i.e., (1) we would not *have* them; (2) we could not *imagine* them; (3) we could not *recognize* (or *label*) them as “red,” “green,” etc. (p. 416)

In the "Introduction" to Ludlow, Stoljar and Nagasawa (2004), Stoljar and Nagasawa say, based on an interpretation of remarks in his 1967 "Postscript," that Feigl's view was that a Martian super-scientist *could* have complete knowledge of human mental life. In Feigl's more detailed discussion of these thought-experiments in his 1958 paper, he makes it clear that his view is that a Martian might predict all human behavior but would lack knowledge by acquaintance of human mental states, and would lack the recognitional and imaginative abilities (etc.) listed above. Feigl is in the camp of those who reply to knowledge arguments (*avant la lettre*, in Feigl's case) by saying that newly-sighted scientists, or those who leave monochrome rooms, do not learn new facts about the world (they do not, as Lewis said, exclude some possible worlds from being candidates for actuality), but undergo new experiences and become able to exercise new abilities.

especially those between (2) and (3), do not motivate an addition to our ontology. Part of what has caused the problem here could be a vague sense that deploying a scientific description *might* somehow encapsulate the experience of another subject. But to think that is not to fully take on board a materialist view of the world, a view in which material subjects both have experiences and develop theories about them. Expressions and records of theories are further physical objects and processes; the theories can say what's there, but can't give you a point of view other than your own. The only encapsulation-*like* thing they can do is prompt memories and imaginings, which involve what Feigl called knowledge by acquaintance, and depend on the capacity to have experiences of particular kinds.

I now turn to Chalmers' arguments against materialism, which are based on similar issues to those motivating Kripke and Levine. Chalmers says that we can conceive of an exact physical duplicate of a conscious human, where this duplicate lacks consciousness (qualia). From our ability to conceive of this scenario, we can infer that it's "metaphysically possible." But if the mental is truly no more than the physical, then it's not metaphysically possible for this "zombie" duplicate to exist. From the possibility of the zombie, then, we can infer the falsity of materialism.

The reply I draw on uses a view about different kinds of imagination, introduced by Nagel in a footnote (1974) and fleshed out by Chris Hill and Brian McLaughlin in a response to Chalmers (1999). The anti-materialist argument is based on the *appearance of separability* of the mental and physical, the fact that we can "conceive" of one without the other. Nagel suggests that this appearance is the product of the structure of the human imagination, especially the relation between what he calls *perceptual* and *sympathetic* forms of imagination. To imagine something perceptually, we induce in ourselves a mental state that resembles a state we would be in if we were perceiving that thing. When we imagine something sympathetically, we induce in ourselves in a mental state that resembles the thing we are imagining. This second method can only be used to imagine mental states and processes, either our own or someone else's. We can also freely separate and recombine, in our minds, the products of these two kinds of imagining. We then find "that we can imagine any experience occurring without its associated brain state, and vice versa. The relation between them will appear contingent even if it is

necessary, because of the independence of the disparate types of imagination." (Nagel 1974, footnote 11).[11] As Hill and McLaughlin say in their response to Chalmers, this does not imply that Cartesian intuitions are false. The separability of mental and physical might still be real. But "the imaginative sources of Cartesian intuitions create no presumption in favor of the truth of such intuitions" (1999, p. 488).[12]

If the Chalmers problem were genuine, not only physicalist views would face it. Chalmers accepts that many forms of panpsychism encounter problems with conceivability arguments (2015). *Whatever* might be posited as a basis for "macrophenomenal" states (whole-person phenomenally conscious states), we can conceive of that basis present and conscious states absent. Perhaps the only view that is not bothered at all is a version of substance dualism where the non-physical component is irreducibly macrophenomenal – something like Descartes's view. People who take

---

[11] This reply does need filling out. The Chalmers thought experiment is one about physical replicas of people, who lack subjective experience. The presence of the physical is imagined together with the *absence* of consciousness. The imaginative act is special even in the context of the Nagel analysis. There are a few ways it might work. We might perceptually imagine the physical and leave the mental "blank" in sympathetic imagination. Or we might employ an inference, noting first that one can sympathetically imagine *any* experience in conjunction with any physical set-up, and inferring that the absence of experience is also an option. I am indebted to Ryan McElhany for pressing this issue.

This kind of reply to Chalmers' arguments is often now called the "phenomenal concept strategy," but the main idea does not require the existence of a special set of concepts, and my version here makes no claims about concepts.

[12] The "footnote 11" reply to intuitions about the separability of mind and body has been criticized by Tyler Doggett and Daniel Stoljar (2010). Their main point is that the allegedly unreliable form of imaginative combination described by Nagel seems often to lead us to *reliable* modal intuitions and *justified* modal claims. My response is that once we see the processes at work, we should indeed question all intuitions of dissociability that derive solely from this source. (Doggett and Stoljar also discuss several interpretations of the argument. I support what they call the "No Reliable Combination" version).

Why didn't Nagel see this as a response to his own challenges to materialism? I take it that Nagel was offering this analysis as something that could augment a materialist theory that seemed to have some traction on the mind-body problem, but encountered a problem with modal issues. This treatment of the imagination would help with the modal challenges. But Nagel also thought we had no materialist theory giving us traction on the problem (and still thinks this now).

conceivability arguments seriously find themselves looking for a trade-off: views that don't have problems with conceivable separations are the most implausible on other grounds – the closest to folk dualism or a doctrine of souls. Views that are less scientifically problematic have more problems with conceivability.

I'll now make some comments that unify these treatments of Jackson's and Chalmers's arguments. In Mary-like cases, a person (Mary) has a point of view on their own experiences, and a point of view also on other subjects, whom she studies. Those subjects engage with color in a certain way – a way that includes recognition, episodic memory, and so on – while she does not. Then when she joins their ranks, she goes through new mental processes and also gains new cognitive capacities directed on color (recognition, instant discrimination, some forms of memory). Mary, like any material subject, can acquire a new vantage point on some of the world's contents and features, and this will change the kind of knowledge she has of them. This account of Mary's before-and-after sequence does not require that we recognize a set of non-physical facts.

In Chalmers-type cases, we don't have a situation where points of view differ across agents, or within an agent across times, but one where different points of view are reflected in different kinds of imagination. We are invited to consider a situation where there is a putative subject, *A*. By means of sympathetic imagination, we can imagine *A* having a point of view as a subject. We can also imagine the same subject as seen from another point of view – we perceptually imagine *A*'s body, or brain, or the chemicals that make it up. To do this is to imagine having the point of view of another subject, *B*. We then find that we can freely recombine the products of these imaginative acts. We can switch off either the first imagining, and have everything "dark," or switch off the second, and have a disembodied soul. But this does not show a real separability of mental and physical, as it can be explained as an artifact of the relations between different points of view as they figure in different forms of imagination.

No matter what they are made of, subjects will look different from the inside than they do from the outside. Difficulty in describing the character of subjective experience in scientific terms is to be expected, no matter how the metaphysics goes. But the fact that subjectivity and points of view are so diagnostically useful when dealing with challenges to materialism does not tell us more about what they *are*. As Nagel said,

"The... question, how anything in the world can have a subjective point of view, remains unanswered...."

3. The Evolution of Subjectivity

The treatment of subjectivity in the previous section reshapes the explanatory gap but does not bridge it. For that, we need an account of the place that subjectivity has in a physical world. I'll approach that question with a biological, evolutionary approach. I'll argue that subjectivity is an explicable biological phenomenon. It will be contentious whether any account of this kind really does the job. Some will concede that biology can give a theory of how living systems come to have a range of cognitive features, but will insist that for any such system, we can ask the further question of whether there is something it's like to *be* that system, a question that biology has no resources to address. My aim is to discuss subjects and their evolution in a way that makes as much contact as possible with this challenge.

The role I am giving subjectivity as a bridging concept suggests that the target of explanation now is a kind of thing, or system – systems that *are subjects*. That is roughly, but not entirely, right. The situation is not one where *subjects* are definite evolutionary products, in the way *birds* are. Rather, there is collection of biological properties that bear on subjectivity in the sense that has been philosophically problematic, that are targets for evolutionary explanation; these are "subjectivity-relevant" biological features. Biologically, subjectivity has a complement in *agency*. In a biological account, subjectivity and agency are handled together; they are tied together in evolutionary processes. But the side of the mind that is picked out by subjectivity is central to problems in the philosophy of mind – at least as things have been framed in the tradition I am responding to. As I said above, this is the side of the mind that concerns sensing, seeming, and feeling. In the section above on the diagnostic role of subjectivity, the idea of *point of view* had an important role. Not all subjective experience is sensory, but this provides a good path in.[13] One thing we can try to do is explain the evolution of systems

[13] This is discussed in my "Materialism, Subjectivity, and Evolution," (unpublished Jack Smart lecture, Australian National University, 2017).

with genuine points of view. This is not just the explanation of sensing – of "viewing" and its relatives – but explanation of a *locus* of this viewing, a self – the metaphorical "point" in the idea of point of view.

In a minimal and unhelpful sense, entities with "points of view" are everywhere. A rock, as I said earlier, could be said to have something like a point of view. It has a location in relation to other things that affect it, and its responses show some specificity; there are ways it is sensitive and ways it is insensitive to what happens around it. If warmed, it retains heat which dissipates slowly. Consider also a digital video camera (if turned on and running). Here there is not much output, but the system has complex ways of responding to stimuli, modifying its inner state. I'll assume that these are both "foil" cases, non-subjects or pseudo-subjects.

If we think about the rock and its apparent failures as a subject, a first fact that can be noted is that there is no non-arbitrary way to mark out the system itself, and its boundaries. Half the rock meets the same minimal criteria as the whole rock, as does the rock plus some of the air around it. *Any* connected collection of matter (not too huge) has (i) a location and (ii) some sensitivity and response properties – an input-output profile. In the case of more genuine subjects, there is a non-arbitrary distinction between self and other. Subjects are located and bounded, to at least a fair degree. How definite this bounding tends to be is an interesting problem, with special cases and controversies. Problem cases arise especially when subjects partially merge or split, and also from tightly integrated extensions of sensory systems.[14] But those phenomena, which must be part of an eventual story, are different from arbitrariness. I take it that the world contains, at each time, a fairly determinate set of experiencing subjects, with unique spatial locations. These features are important in giving rise to the "point of view" phenomena associated with subjectivity.

[14] I have in mind "extended mind" debates (Clark and Chalmers 1989) and some unusual phenomena involving twins, especially the case of Krista and Tatiana Hogan, discussed in Montero (2016).

In a biological framework, one of the first things that can be explained is the existence of entities that are marked off from their surrounds in relevant ways.[15] This is a consequence of basic features of living systems. Living systems have traditionally been seen as *self-maintaining* systems, and this is indeed one important feature of them. Living systems are out of thermodynamic equilibrium with their surroundings. Traffic exists between such systems and their surrounds, but this to-and-fro must be limited and controlled, so metabolic systems build and maintain boundaries. Metabolic processes continually recreate both a pattern of organization and the system's distinctness from its surrounds. The term "self-producing" is sometimes used for this feature of living systems (Maturana and Varela 1980), and the term is a good one in emphasizing that what living systems do goes well beyond familiar kinds of maintenance and damage control. They are continually synthesizing the molecules that make them up, preserving a pattern of activity through turnover in materials.[16]

These are basic features of cellular life (as opposed to the form of "life" seen in viruses and their relatives). If we take, for example, present-day prokaryotes, such as bacteria and archaea, these objects each maintain a membrane, with channels and receptors that cross it. Some of the activities that bridge the membrane are purely metabolic, while others contribute to control – they guide flexible patterns of response, tracking what is going on around the cell and reacting to it. The result is an arc between sensitivity and activity that at least resembles the arc between perception and action in animals. Cells sense environmental conditions, process what is sensed, and act in response in ways that maintain the cell's characteristic patterns of activity, including these relations between cell and environment.

Unicellular organisms, then, are subject-*like* entities, entities with some of properties relevant to subjectivity. They are non-arbitrarily bounded systems (i) that

---

[15] Van Inwagen (1990) argues that only living organisms are real "material beings" at a macroscopic, non-fundamental level. Other putative objects are too indeterminate with respect to their spatial and temporal boundaries to be real. That is a stronger conclusion than I think is motivated, but it is true that organisms – or more exactly, cells – are special kinds of material beings, as they are self-demarcating in a way not seen in other objects.

[16] This concept, and other more biological themes in this section, are discussed in more detail in Godfrey-Smith (2016b).

maintain and regenerate their organization, (ii) that do so in a way that includes maintenance of the boundary between system and environment, and (iii) whose sensitivity to external events is coupled to responses that contribute to these self-maintaining and self-demarcating patterns of activity.

The origin of cellular life, probably quite early in the history of the earth, rearranged matter into these things. A world that contains proto-subjects in this sense is a world with further relevant features. A proto-subject, or collection of them, can't comprise *everything*. There must be an environment as well. Demarcated and self-maintaing systems of this kind exist by means of energetic transactions with a milieu that must have different, and complementary, features. That milieu must be a source of useable energy, and a sink for higher-entropy outputs.

Animals are one form of multicellular life. In the transition to multicellularity, many cells come to live as a metabolic whole, in some cases also exhibiting complex behavior at the level of those wholes. These behaviors are made possible by a division of labor across parts and interaction between them. Multicellularity itself has evolved over a dozen times, but the animal path is distinctive in the sort of unit that resulted.

Before looking at that path, I will note a general feature of multicellular collectives, important in animals but not only in them. In the case of cells I made much of the boundedness of these objects. In the formation of multicellular animals, a larger unit comes to exist, but these entities have unclear boundaries in one respect: it is often not so clear which cells are in and which are out. It tends to be indefinite where an animal begins and ends, though it is more definite where each cell begins and end (and even here there are exceptions).[17] This is primarily because of symbiotic interactions between the eukaryotic cells in the system and various bacteria and other microbes. Multicellularity is more permissive with respect to cell-level parts than unicellularity is with respect to cell-versus-other. Cells can enter *heterogeneous collaborations* (Pradeu 2011) as well as like-with-like associations. These collaborations appear to be ubiquitous in animals and plants

[17] Cell boundaries become very vague in some parts of plants. Many fungi and some other organisms (including some seaweeds such as *Caulerpa*) have an external boundary but have partly or entirely abandoned cell boundaries within the collective.

and can be more or less tight. As a result, multicellular collectives often do not have definite boundaries at all.

In other respects, the evolution of animal life resulted in highly integrated collectives. For a large category of animals, including ourselves, there is definite individuality, achieved not so much through demarcation of boundaries, but through a coordination of parts that connect a whole system. We are readily countable things; there is a fairly definite number of humans in the world at each time, even though the boundary between each human and environment is more indefinite.[18] There is not, in contrast, a definite number of corals or grasses. Those organisms have a *modular* design, made of small units, like coral polyps, that are organism-like to various degrees in their own right. In familiar plants like oak trees, there is also partial independence of each branch, especially in reproduction. A coral or tree is in some respects more like a social entity than a single organism. Vertebrate animals like us, and also many invertebrates, are *unitary* organisms with respect to their organization and development. There is tighter integration across the whole and less autonomy of parts. (Cancer, often deadly for us, routinely arises in trees but is usually of little consequence.) Rapid and targeted interaction between parts in such organisms is ubiquitous, especially via the elaborate organ of coordination and integration seen in nearly all animals: the nervous system.

Nervous systems probably arose early in animal evolution. The earliest fossil records of animals, from the Ediacaran period (635-540 million years ago), suggest that for some time the sensory and behavioral capacities of animals were minimal, but various lines of evidence indicate that nervous systems had evolved by this stage. The early parts of the Cambrian period, around 540 million years ago, saw the beginning of a regime of more extensive behavioral interaction, including predation. These capacities probably evolved in parallel in several lines, rather than radiating from a single source. This period sees the evolution of image-forming eyes, claws and similar means to manipulate objects, along with legs, spines, and fins.

It is likely that a process of evolutionary feedback drove many changes in this period. A plausible picture of the earlier period, the Ediacaran, features a sea floor

---

[18] See also the case mentioned in footnote 14.

covered in microbial life, and in the more complex animals, a lifestyle of grazing with limited movement. However, these animals then became resources of a new kind themselves, as they were concentrations of nutrients.[19] Those resources may initially have been consumed only after death, by scavenging. But a patchier environment of this kind puts a premium on better senses and directed movement. In time, scavenging gave rise to predation. From this period onwards, a process of positive feedback drove the coevolution of nervous systems, bodies, and behavior.

The sequence of events sketched above may change in the light of further work. But within this general family of views about animal evolution, there are resources that bear on the explanation of how a further range of subjectivity-related features came to exist. One such change concerns the nature of perception. Several developments may be important here, and I will first discuss one that involves the relation between perception and action.

Basic forms of sensitivity to environmental conditions, tied to adaptive responses, are seen not only in animals and plants but in bacteria and other single-celled organisms. Changes arise, though, in the form of sensing seen in animals. As noted by Björn Merker (2005), when an organism combines self-propelled movement with good senses, that organism encounters "liabilities of mobility." The animal's own actions can produce some of the stimuli that it has evolved to respond to, introducing uncertainty in the causes of those stimuli. The organism must distinguish what Von Holst and Mittelstaedt (1950/1873) called *exafference* – exogenously caused stimuli – from *reafference* – sensory changes caused by an organism's own actions. One way to do this is to send a signal of some kind from the action-producing part of the system to enable compensation in perceptual processing (an "efference copy" or "corollary discharge"). Such mechanisms establish a loop within the organism's boundaries that compensates for the externally-routed causal loop due to reafference.

Just a few neurons can solve simple versions of this problem; the task is simple when all that must be done is inhibition of the usual behavioral response to a fixed and well-defined class of stimuli (such as contact with the animal's surface). But as sensors

---

[19] For discussion of these features of the Ediacaran and Cambrian, see Budd and Jensen (2016) and Trestman (2013).

multiply and actions become more elaborate, the problem ramifies, and mechanisms dealing with the problem become more complex and centralized. However it is achieved, there is a particular *shape* to the control systems seen in biological subjects that can both move and sense well, even before we consider learning, reasoning, and so on. In these organisms, there is not a simple feed-forward path from senses to effectors. Instead, because of the inevitable *de facto* influence of action on what is sensed, there is accommodation by the animal of the role of self-caused stimuli. What is achieved, by some means, is the organism's own registration of the self/other divide.

Merker claims that in general these mechanisms are implemented peripherally in invertebrates and in a more centralized, brain-based manner in vertebrates. He claims that subjective experience comes only with centralization. Both those further claims can be questioned.[20] Merker's development of this theme also emphasizes the "liability" side of reafference. But controlled motion – of both the parts of a body and the whole – brings sensory opportunities as well as ambiguities (Keijzer 2015). Actions probe the environment in a way that yields useful information, but this too requires that the animal have sense of what it has done, as well as what has happened. The path being initiated, again, is one featuring a transition in the general form of the links between sensitivity to the world and behavioral response.

In the absence of this sort of structure, there is a one-way flow from input to output (seen also in the case of the rock). When we think about perception, that feedforward shape is one in which the subject itself might seem to "go missing." But once animals start to accommodate and utilize reafference, the character of sensing changes. The animal is now not only open to the world, but open to the world *as* the world, as distinct from self. This feature has probably evolved several times in mobile animals.

---

[20] Merker links his view to the tradition that understands subjective experience in terms of the use of an "inner model" of reality. The same considerations involving reafference are seen in "enactive" views of perception and consciousness; here they accompany quite different views about the centrality of the brain and views at least partially opposed to representationalism. See Hurley (2001), O'Regan and Nöe (2001), and Thompson (2007). Some recent work has questioned Merker's claim about the simplicity of control systems that deal with complex motion and reafference in invertebrates, especially for the case of insects: see Barron and Klein (2016).

This action-involving shift in sensing is not the only one relevant to subjectivity. Burge (2010) argues that the arrival of "perceptual constancy" phenomena mark a shift from mere sensing to genuine perception, and, further, that this marks the advent of genuine *perspective* – point of view in a strong sense.[21] The constancy mechanisms Burge makes use of here do not, in most cases, require a specific role for action. I won't try to determine in this discussion the relative importance, or the evolutionary ordering, of constancy phenomena and reafference compensation; both, and perhaps other developments in sensing, are relevant to subjectivity.[22] Instead I will add to the discussion a different family of features.

While many discussions of the evolution of consciousness focus on perception, another tradition views a particular kind of *feeling* as a plausible candidate for a primitive, basic form of subjective experience. What is important, on this view, is not outward-directedness and perceptual engagement with the world, but the valenced registration of events – the distinction between good and bad, welcome and unwelcome. The folk notion of "feeling" identifies this phenomenon quite well; these are transitory internal events with a motivational role.[23]

Evaluation, in some form, is ubiquitous in living things. This applies both to internal goings-on – the registration of imbalances and deficiencies – and tracking the relevance of external events with the (extero)senses. There seems no reason to think that evaluations must always be *felt*. Bacteria and other unicellular organisms show patterns of approach and withdrawal. In us, a lot of homeostatic feedback is not experienced. How might evaluation give rise to genuine feeling? As with the perceptual capacities discussed above, several transitions may be relevant. As in the earlier case, I'll discuss one in some detail and then indicate a possible role for others.

---

[21] "[A]ction guided by perception derives from a perspective in a way that action in response to mere sensory registration does not" (Burge 2010, p. 337).

[22] See also Feinberg and Mallatt (2016), who argue that spatially organized mapping of sensory input has this role.

[23] Views that start the evolutionary story here, in different ways, are seen in authors such as Ginsburg and Jablonka (2007), Derek Denton (Denton el al. 2009), and Damasio and Carvalho (2013).

Many animals, but not all, can learn that an action has positive or aversive consequences – I will refer to this as *instrumental* learning. What is needed to achieve this is a capacity to open-endedly associate specific behaviors (or behaviors performed in particular circumstances) with an evaluation of their effects. An internal system must register welcome and unwelcome events and use this registration to reshape its behavior.

Independently of evolutionary questions, a link between this kind of learning and subjective experience has become influential in recent discussions of pain. Animals can exhibit *nociceptive* behaviors, adaptive responses to damage, in circumstances that suggest, given the neurological setting, that no feeling is present. This motivates work on distinctions between more rudimentary and more complex damage-related responses, where the complex ones include learning.[24] The function of felt pain, in us and other animals, may be largely to act as part of a means for rewiring behavior; its point is not merely to change behavior in a momentary way, but to bring the distinction between welcome and unwelcome events to bear on the guidance of future actions.

What relation does this set of factors have to those involving perception and a sense of self, discussed above? There seems to be at least some scope for dissociation of the two. The first set of traits gives rise to an outward-directed form of subjectivity; it explains why mobile animals have points of *view* in a sense that does not apply to other kinds of life. The second trait, which involves valuation, has weaker connections to mobility, and does not require the perception of objects as objects. Via Brembs (2017), though, we can note a point of intersection with first evolutionary path. Instrumental learning requires that an animal know what it has *done*. It might not need to know much about what is happening around it, but does need to track its own actions and, often, the context of their production. So a tracking of the self-other divide may be integral to this second set of traits associated with subjectivity, as well as the first.

At present it seems likely that instrumental learning has several independent origins, as it appears to be scattered across major animal groups. A review by Clint Perry et al. (2013) lists it as confirmed in many but by no means all mobile animals. It has been seen in insects, crustaceans, cephalopods, and gastropod molluscs, but not reported in

[24] See Allen (2004), Elwood (2011), Adamo (2016).

spiders, myriapods such as centipedes, or echinoderms. Some of these gaps may reflect the difficulty of showing that an animal *can't* do something, and the 2013 review is presented as a review of what has been shown to be present, without claims about what has been shown absent. So there are empirical uncertainties here, but there appears to be some dissociation of traits associated with complex perception and those associated with evaluation.[25]

If complex perception and evaluation are separable, this raises the possibility that there are two kinds of phenomena that we vaguely group as "subjective experience," both of which are present in us but which are distinct in principle and sometimes found separately. If we ask, introspectively, about conspicuous features of human experience that may have early forms, it might be intuitive that one side of the phenomenon involves tracking external objects and events *as* external – achieving a point of view on things – while another involves distinctions between good and bad, a distinction that might be present in phenomenal washes that have no definite referral to organism or to environment. Such an attempt at pulling-apart of the two features has to reckon with the points made by Brembs and Merker about the role of a sense of self-versus-other on both sides – this might be a common denominator in the evolution of subjectivity. But there may at least be some differences in "style" across animals in these respects. Some spiders, for example, are mobile and show complex perceptual capacities, but they "score low" with respect to evidence for motivating feelings. Perhaps they are sophisticated trackers of the world but motivationally robotic. Among the animals that are reported to show instrumental learning by Perry et al. (2013), the least mobile are gastropods – slugs and snails. So far, though, they show only quite simple forms of instrumental learning.

[25] Some difficulties here concern the identification of instrumental learning, which shades off into other phenomena involving reward systems that operate on shorter time-scales – systems responsible for simpler approach and withdrawal behaviors, for example. Classical conditioning when acting in combination with reward systems that guide moment-to-moment behaviors can give the appearence of instumental conditioning, as in some "conditioned place preference" behaviors. Even platyhelminth flatworms, which have very simple nervous systems, can exhibit learning of this kind. A number of authors believe the familiar distinction between classical instrumental conditioning is problematic, either in principle or at least in relation to standard methods. These findings and their significance are discussed in more detail in Barron, Søvik, and Cornish (2010).

When discussing the sensory side, I said that several different developments may be relevant – reafference compensation, perceptual constancies. On the evaluative side, I discussed instrumental learning as one way of putting valuation to work, but there are also other ways. These do not involve long-term change (learning) but sophistication in moment-to-moment choices. A good example is evaluative trade-offs, such as choices that weigh one aversive possibility against another (see Elwood 2012 for an example in crabs).

A range of questions are unresolved about the distribution and evolution of these subjectivity-relevant properties, even aside from questions about their relation to the "explanatory gap." The last thing I'll do in this discussion of animal evolution is sketch a possible evolutionary path that ties the two sets of features – sensory and evaluative – together.

Suppose that we are in an early Cambrian context; animals are becoming more mobile and must start to track events in real time. This results in a changing of the shape of perception to a self/other modulated form. With the rise of mobility also comes a richer repertoire of behaviors, and more possibility for behavioral novelty – for *doing* new things, and doing them in new circumstances. The ability to do new things puts a premium on instrumental learning. New behaviors can be produced in different circumstances, to good or ill effect. This suggests an ordering of the two traits discussed: complex perception first, motivating feelings second. The test cases that probe such relationships, as indicated above, include spiders and some other land-based arthropods (complex perception, simpler evaluation) and gastropods (simpler perception and motion, perhaps richer evaluation).

Many details remain unclear, but these ideas give us a picture of some candidate transitions from rudimentary to richer forms of subjectivity. Summarizing some steps in abbreviated form, early stages feature living systems that show some of the form or "outline" of subjectivity – self-demarcation, traffic with an environment, and capacities for control that include treating some conditions as indicators of others. These control processes embody, in tacit form, a distinction between welcome and unwelcome occurrences, between desirable and undesirable stimuli. Within the animal lineage, as a result of demands and opportunities associated with the animal mode of body

organization and lifestyle, forms of sensing arise that include tracking of the self/other divide. A divide that is universally *present* in cellular life becomes *salient to* the organism itself. Instrumental learning puts the new repertoires of behavior found in mobile animals to better use. This step, perhaps along with other moves towards sophisticated evaluative capacities, require the internal registration of valence. In both cases, what was present *de facto* in biological subjects – self/other, good/bad – becomes salient to the animal itself.

An account of the origin of subjectivity may also include additional factors. Hypotheses in this area often put much weight on the integration of processing (Dehaene 2014). Integration of this kind is said to give rise to a coherent "model of the world." That may indeed be important, in concert with the earlier-evolving features discussed here. The main points I want to make in this paper are coarse-grained. Consider just animals that are the products of all the evolutionary processes outlined above. These, I suggest, have become *subjects* in a strong sense of the term. They are also *material* subjects; their subjecthood is due to their material properties and embedding in the world. The way that they are a corralling of matter and energy makes them centers of receptivity and agency in a way that gives them a definite point of view.

Several further questions emerge from this discussion. First, as discussed above, to what extent do the various subjectivity-relevant features form a correlated package? To what extent are the forms of subjectivity seen in different animals deeply different rather than having a common core? Second, from a biological point of view all the various subjectivity-relevant features exist in partial and intermediate forms. If our treatment of subjective experience is guided by the biology, that suggests a graded treatment of this feature itself – of the fact of it *feeling like something to be* a particular organism. Philosophically, this outcome seems awkward and hard to think about, and if this is where we end up, we will want to look for new ways to describe the graded character of the situation on the mental side. We might say the goings-on in various organisms are more experience-like, more experiential. The position that would result from accepting a gradualist view here need not be one in which the gradient runs down to encompass bacteria and the like. It might do so, but the account sketched above has been one in which some distinctively *animal* features – features of the relation between animal

sensing and action – have special importance. Then the unicellular proto-subjects discussed early in this section would provide a frame or foundation for later stages that engender subjective *experience*.

Especially with respect to its details, the evolutionary sketch in this section should be seen as a *how-possibly* explanation. But it's a little more than that; the aim has been to walk through a series of steps such that once we reach one stage, we can see the next ones waiting. The result could then be called a *how-possibly-necessarily* explanation. Here the first modality – "possibly" – is epistemic. The second modality is not a strong sense of necessity, but something that contrasts with mere accident – a kind of robustness or predictability. It's a story about how things might have trodden a predictable path, on a planet like ours, a path by which subjectivity reliably arose. A story like this begins to make the origin of subjects *intelligible*, as Nagel and Levine asked.

4. The Gap

How much of the explanatory gap has been closed? Not all, in several respects. First, Levine asked for explanations of why the feel of seeing red goes with *this* brain state, and I've not addressed those questions at all. Instead, I've tried for progress on the question of why one physical system has subjective experience while another does not. The aim, in picturesque terms, is to develop a "sideways-on" view of animals, which, as it becomes more detailed, not only gives us a view of the animal *from* the side, but brings us *alongside* them. A theory of this kind does not enable you to jump in and *be* them, but enables you to see that there is *something* it's like to be them, and in some cases to make mappings between their experiences and our own.

I said at the beginning of this paper that a complete solution to the problem will include a mix of critical, diagnostic moves with new pieces of theory. More fully, I think a solution will have four pieces.

The first piece is a diagnosis and getting-over of misleading intuitions. Attention to the cognitive situation of the philosopher thinking about the mind-body problem shows several sources of illusion. I discussed some of these in section 2. Second, the appearance of an unbridgeable gap comes in part from thinking in crude ways about the physical. We are probably misled by intuitive conceptions of matter and of physical processes,

conceptions which remain too close to a mechanistic picture. In some recent work, a rethinking of the physical has led a revival of panpsychism. That is a step too far, but I agree with some of the critical discussions of physicalism that accompany this work. The category of "the physical," as it functions as a resource in the explanation of the mental, is not very well-defined. Given the reasonable suspicion that physics itself may change in its view of what is basic, there is a good chance that "the physical" will look like a quite different sort of explainer in 100 years, just as "the physical" looked different in 1900 before Einstein and quantum mechanics, and in 1800 before Maxwell.[26] Recognizing this is not a move towards panpsychism.

The third piece is a more specific critique of intuitive conceptions of matter, one that bears on the character of living systems. The mind has a basis in a particular kind of physical activity (nanoscale, aqueous, stochastic), which our experience with macroscopic solid objects, especially machines, equips us poorly to think about.[27] The fourth piece is an account of the nature and origins of subjectivity, as attempted in the previous section.

In the light of what I've laid out, I'll return once more to general challenges to explanations of mentality in materialist terms. Can we play a movie in our heads in which we imagine an evolved animal subject, with its nervous system making possible all the things I discussed in the previous section, but no qualia? Yes, we can play the movie (I can, anyway). But that does not mean much. The appearance of separability is an uninformative product of two forms of imagining. Can a description of the make-up of these things, using the intersubjective language of scientific theory, encapsulate what it feels like to be one of these subjects? No, and materialism does not require that it do so.

Someone might say that all that I've been describing here – and all that's envisaged in a fuller account – is a set of "cognitive" features, which are in principle distinct from anything "qualitative." That is not an argument, however. The term "cognitive" is this context is a broad one for features related to perception, information-

---

[26] For criticism of familiar thinking about "the physical" as an explainer, see Crane and Mellor (1990). The attitude sketched at the end of this paragraph is close to what Stoljar calls "Nagelian monism, and defends, in his (2005).

[27] This theme is discussed in detail in Godfrey-Smith (2016c).

processing, and behavioral control. It's a mistake to think that anything we learn on the cognitive side *must* leave untouched questions about the qualitative; it might go that way, but it might not. The "cognitive" in this broad sense may include many features related to subjectivity and point of view. The "qualitative" isn't the converse of the "cognitive" in the sense of an unexplained residue, a set of features we can know in advance to be inexplicable in those terms.

In a related move, a critic of an earlier version of this paper objected that the sort of biological story given here helps us understand some things, but not the *phenomenal*. My project, though, is to explain phenomenality in terms of subjectivity. The phenomenal is the intrinsic character of experience in a subject. The aim is to show that some physical systems are subjects, in a way such that there is something it's like to be them. *What* it feels like is a consequence of the details of their structure and make-up, what kind of subject they are. As discussed above, it's not possible to describe their experience in a way that verbally encapsulates it. *Having* the experience, and having cognitive capacities dependent on having the experience, requires *having* the features that give that subject's experience its character – being an animal of a certain kind in a certain situation.

A living system *is* some way (trivially), and, given how it is, it goes through certain states when events occur that affect it, especially events that affect it in ways relevant to its self-maintaining and goal-directed activities. It is the business of a living system to register such events and determine what to do in response. In a great many cases, this responding to events is comparatively simple – unfocused and unintegrated. But on some paths through the evolution of animal life, there is the building of a center of agency of a different kind, one whose capacities includes the tracking of external events *as* external and of unwelcome events *as* unwelcome. It's the business of such animals to have a point of view and to put it to use. There's not only a way things go *in* such a system, but a way things go *for* it.

My approach here has been biological in a general sense, but in the previous section I also spent a lot of time on the history. What is the role of the evolutionary story itself? It might be objected that history cannot, in principle, be of great relevance here. What matters is the synchronic side; what matters is how a set of physical processes going on *right now* might be sufficient for subjective experience right now. Subjects

would have, or not have, their experience-related features if they came into existence by a route other than biological evolution.[28]

A historical narrative might even be worse than irrelevant, due to further quirks of human psychology. The flipside of the philosopher's explanatory gap is the psychologist's *illusion of explanatory depth*. This term was introduced by Rozenblit and Keil (2002) for a phenomenon in which ordinary people think they understand how things work (how a zipper works, or a helicopter) much better than they actually do. Their confidence and feeling of ease far outstrips their actual knowledge. Tania Lombrozo has offered a account of how the illusion arises that is relevant here.[29] She suggests that the phenomenon may be especially marked in cases where devices have clear functions (such as a zipper), and the illusion arises in part from the psychological effects of a sense of functional understanding. Our ability to apply a teleological description to a complex process may generate an over-estimation of one's mechanistic understanding of that process; it generates the sense that because one knows *why,* one also knows *how*.

Imported into the present discussion, that idea would be the basis for a criticism of the following form: an evolutionary story gives us an account of why it *would* make sense for a physical system with subjective experience to evolve, *if* physical processes were sufficient to give rise to experience. But the explanatory gap is the idea that we don't know how *any* physical set-up could suffice. Because the evolutionary story gives us an understanding of something in the vicinity, it gives us a sense of more understanding than we really have.[30]

In reply, I agree that there is no substitute for the synchronic story, the story about how biological processes present at a time are sufficient for subjective experience at that

---

[28] Thanks to Michael Fitzpatrick for pressing this point.

[29] As Lombrozo says, Rozenblit and Keil (2002) do touch on this explanation, but it is not their emphasis. Lombrozo links this idea to findings showing that children treat functional information as an inappropriate stand-in for mechanistic information; for example, they offer answers to "why" questions when asked a "how" question (Abrams, Southerland, and Cummins 2001).

[30] This is like the opposite of a point about epiphenomenalism that goes back to T.H. Huxley's steam-whistle (1874). Huxley and others suspect that consciousness has no function, and see this as an evolutionary puzzle. I say: subjectivity and hence subjective experience do have an evolutionary rationale, *if* they can be features of a physical system. But that does not show they *are* physical.

time. But the evolutionary account does play a role in relation to the gap. Part of this relates to the critical side of the project. Recall the second and third of the four pieces listed above, the need to set aside intuitions about the physical that artificially magnify the problem, giving the appearance of antipathy between mind and matter. Nagel (2012) and others talk about "dead matter," ponderous and dull, when describing this terrain, emphasizing the sheer differentness of mental and physical. In such a context, the evolutionary story gives us a sense of why it's *natural to matter* to do these things, why matter can organize itself in a subjectivity-producing way.

The evolutionary story plays the fullest version of this role in concert with an account of the origins of life, a topic I've not tried to tackle. It is helpful, in seeing how the pieces fit together, to imagine a best-case scenario, even though it is some way off. Suppose we had an account of the origin of life, as well as the history of cells, animals, and so on – an account of how physical and chemical processes, at work with the elements and forms of energy present on earth, gave rise to the pockets of order that are living systems. (Koonin and Martin 2005 sketch a scenario in which the boundedness that is characteristic of cellular life is initially provided by external physical scaffolding, replaced later by self-built membranes, in a process occurring at ocean vents.) The result, cellular life, consists of systems that respond to what happens, keeping themselves going. The evolution of multicellular organisms enables divisions of labor that give the mechanisms of sensing and acting new capacities. An image-forming eye, for example, must be laid out in space in a manner that (almost) requires multicellularity.[31] As animals become large and mobile, tracking and responding to events in real time, they become subjects of a richer kind. The synchronic story still has to play its role, but the gulf is reduced.

Before closing, I'll briefly discuss two issues that were not addressed above. First, the view I presented described transitions within biology, between different kinds of living systems. It's common now to think that being a living organism is optional with respect to mentality; a system could have experience without being alive – perhaps while being an AI system realized in an ordinary computer. I'll address this issue briefly, as the

---

[31] For the qualifications in "almost," see Nilsson and Colley (2016).

aim of this paper is more to describe one road to subjectivity than to argue against other roads.

The question returns us to the second of the contrast cases, systems that presumably lack subjective experience, mentioned earlier: the video camera. A camera is an artifact that, in one sense, is *all* point of view. We build it to do a kind of quasi-visual registration, to be an artificial extension of one side of the subject-related things that we do. That quasi-visual registration is accompanied, in the camera, by none of the usual context of vision. It has none of the features discussed in the previous section that make patterns in incoming light significant to a system.[32] The camera is very sensitive to light coming in through its lens, but just as light will affect the camera's sensor, changes in temperature will affect various of its parts, and the camera doesn't care about either. It doesn't use light, or anything else, to maintain itself and preserve its thermodynamically improbable structure.

What about artificial systems more generally? Could there be an experiencing subject, more complex than a camera, realized in ordinary computer hardware? This question can be taken in two ways. One way is tantamount to asking about artificial life; could there be a living system that was not made of the usual biological materials? I don't know, but this is not a problematic possibility for me. The other way of posing the question is to ask whether all the psychologically relevant properties of a living agent could be realized in an artificial system which was not alive. The traits discussed in the previous section involving perception and learning are not out of reach of computational modeling, and could exist in at least some form in a robot that had nothing like a metabolism. If those features (and their relatives) make a big difference in the biological cases, why don't they do the trick here, too?

To deny that they do, it has to be argued that the context in which these traits appear in the robot is so different as to cast doubt on their role. That is easier with the second family of traits I discussed, those associated with the internalization of valence. Instrumental learning (and its relatives) can be modeled computationally, but while a programmer can stipulate that such-and-such is a "reward" signal in the system, it is often

---

[32] Autofocus systems in cameras do not, as far as I can tell, make use of a reafference principle, though one can imagine a system that did.

not clear why this is more than stipulation. In a living system engaged in continual self-production, value to the system has a basis other than stipulation by an outsider.

What if a form of instrumental learning was part of a program by which a robot acted in a way that prevented its being damaged or unable to function? That is indeed a start, but this will be either a vastly simpler and "thinner" affair than the ubiquitous and distributed self-production seen in a living system, or it will be on the way to realizing an artificial metabolism.

A similar set of claims can be made for the traits involving sensing and self/other registration. Sensing is input, or influence across a boundary. But for something to be sensing at all, it needs to be embedded in a particular context and have particular kinds of downstream effects within that system. There also has to be the right sort of unit in place, defining a boundary over which influence can come *in*.

Artifacts can be marked out from their surrounds in various ways, with metal casings and the like. And in a robot, some of what happens downstream of the sensory periphery can be relevantly similar to what happens in a nervous system. But it is similar in a very partial way, not only with respect to make-up, but with respect to what is being *done*. Decades of computer-influenced discussion in philosophy have led to the acceptance of a set of assumptions about which features are relevant and which are irrelevant when comparing living systems to computational artifacts. Processes of inference, and the like, are taken to be relevant, while the self-defining processes that are characteristic of living things are usually seen as irrelevant. But while these processes may be irrelevant to modeling some kinds of cognitive processes, they could be highly relevant to other aspects of the mental, including subjectivity. The reafference compensation seen in living things is occurring in a context where a certain kind of unit is in place, a unit that engages in genuine sensing. The artificial systems usually imagined in this context are very different.

Returning to the question: could an artificial system have subjective experience? I say: perhaps, but such a system would be very different from what is usually envisaged – different not just in material composition, but in what it does.

Second, in this paper I've drawn several times on Thomas Nagel's treatments of challenges faced by materialism. Given this, it's relevant that Nagel himself has a further

objection, expressed in his 1965 paper "Physicalism" and reasserted in *The View from Nowhere* (1986).

> The feeling that physicalism leaves out of account the essential subjectivity of psychological states is the feeling that nowhere in the description of the state of a human body could there be room for a physical equivalent of the fact that *I* (or any self), and not just that body, am the subject of those states. (1965, p. 354)

This is distinct from the challenge of Nagel's 1974 "bat" paper, which is not concerned with facts about the perspective of a particular *I.* Here the concern is that even when a complete account of the world's make-up is imagined, "the thought that TN is *me*" seems to have further content. If this problem is real, it acquires special salience when a treatment of the mind-body problem is organized, as mine is, around the ideas of subjectivity and point of view. Nagel did note in his 1965 paper that though this was "the source" of his uneasiness about physicalism, "it is no more an argument against physicalism than against most other theories of mind, including dualism" (p. 354). The question's critical force is then limited. But in any case, I think there is no problem; I take an entirely deflationary route.

In a world of material subjects, each, if they can speak, will speak of themself with terms like "I." They can all ask the question: Why am I this physical system? Or: Why is that physical system *me*? This sort of formulation does point towards a substantive question: Why does *this* physical structure (which I have) go with *these* experiences (which I have). That is another way of asking the question asked by Levine. But once that question is answered, there is no further question to be answered by some additional piece of metaphysics or science; all that remains are phenomena involving indexicality and point of view. Each of us can ask questions about one of the world's subjects (oneself) using a form of reference that can't be applied to anything else. These include questions about the association between subjective experience and physical properties, but the answers to those questions are recycled versions of answers to questions that have no indexical elements. Given that these are your experiences, you must be this physical system (with a particular structure and placement in the world),

rather than another. So I do not think this last objection of Nagel's gives us reason to worry.

When someone gives a defence of materialism, it is often seen as an expression of hard-headedness; "This brute, hard-edged stuff is all there is." That is not my motivation. If there is a temperamental orientation at work here, it has the nature of monism. But more fundamental is the idea that subjects, subjectivity, and experience are made up, constructed, *put together* out of matter that also does other sorts of things. Matter that has no inherently mental nature *comes to be organized* in a subject-realizing way. Subjectivity is organizational, and subjectivity is the core of the mind.

The aim of this discussion is to point us past a familiar standoff. A materialist gives a story about how animals come to perceive, act, remember, and so on, and the critic says: "OK, but why should it feel like anything to be such an organism?" There will come a point where this seems a misguided question – a question asked because the critic is not seeing the nose on their face, or the wood for the trees, or the university for the colleges. At that point, we know the gap is closed. We are not at that stage yet. But assume a further development of the ideas above, or something like them. Then imagine the exchange: "Now we have a organism which is self-maintaining, and it uses its senses to guide actions, does so in a way that includes continual registration of the distinction between self and other, and goes into states that function as internal rewards and punishments, used to reshape its behaviors in an ongoing way.... And you *also* want me to tell you why it should feel like something to be that organism?"

Acknowledgments: Thanks to Michael Fitzpatrick, Ryan McElhany, Leonard Katz, Henry Shevlin, Thomas Nagel, Tyler Wilson, and a colloquium audience at Stanford University for discussions and correspondence. A first version of this material was the basis for a seminar at the CUNY Graduate Center in 2015. Those ideas were revised and used in the 2017 Seybert Lectures at the University of Pennsylvania ("The Origin of Subjects"). I am grateful to everyone present at those presentations of the ideas, and to

the Seybert Commission. Sarah Robins, as referee, made very challenging and helpful comments on the penultimate version.

References

Abrams, E., S. Southerland, and C. Cummins (2001). "The How's and Why's of Biological Change: How Learners Neglect Physical Mechanisms in their Search for Meaning," *International Journal of Science Education* 23: 1271-1281.

Armstrong, D.M. (1968). *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Adamo, S. (2016). "Do Insects Feel Pain? A Question at the Intersection of Animal Behaviour, Philosophy and Robotics," *Animal Behavior* 18: 75–79.

Allen, C. (2004). "Animal Pain," *Noûs* 38: 617-643.

Arnellos, A. and A. Moreno (2015). "Multicellular Agency: An Organizational View," *Biology and Philosophy* 30 (2015): 333–357.

Barron, A. and C. Klein (2016). "What Insects Can Tell Us About the Origins of Consciousness," *Proceedings of the National Academy of Sciences* 113: 4900-4908.

Barron, A., E. Søvik and J. L. Cornish (2010). "The Roles of Dopamine and Related Compounds in Reward-Seeking Behavior Across Animal Phyla," *Frontiers in Behavioral Neuroscience* 4: 163. doi: 10.3389/fnbeh.2010.00163.

Brembs, B (2017). "Operant Behavior in Model Systems." *Reference Module in Neuroscience and Biobehavioral Psychology*, 2017. http://dx.doi.org/10.1016/B978-0-12-809324-5.21032-8.

Budd, G and Jensen, S. (2015). "The Origin of the Animals and a "Savannah" Hypothesis for Early Bilaterian Evolution," *Biological Reviews.* doi: 10.1111/brv.12239

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D. (2003). "Consciousness and its Place in Nature," in S. Stich and F. Warfield, (eds.), *Blackwell Guide to the Philosophy of Mind*. Oxford: Blackwell.

Chalmers, D. (2015). "Panpsychism and Panprotopsychism," in T. Alter and Y. Nagasawa, (eds.), *Consciousness in the Physical World: Perspectives on Russellian Monism.* Oxford: Oxford University Press, pp. 246-276.

Clark, A. and D. Chalmers (1989). "The Extended Mind," *Analysis* 58: 7-19.

Crane, T. and H. Mellor (1990). "There is No Question of Physicalism," *Mind* 99: 185–206.

Damasio, A. and G. Carvalho (2013), "The Nature of Feelings: Evolutionary and Neurobiological Origins," *Nature Reviews Neuroscience* 14: 143-152.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Penguin Random House.

Dennett, D.C. (2004) "'Epiphenomenal' Qualia?. In Y. Nagasawa, P. Ludlow & D. Stoljar (eds.), *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument.* Cambridge MA: Bradford/MIT, pp. 127-136.

Denton, D., M. J. McKinley, M. Farrell, and G. F. Egan (2009). "The Role of Primordial Emotions in the Evolutionary Origin of Consciousness," *Consciousness and Cognition* 18: 500–514.

Doggett, T. and D. Stoljar, (2010)."Does Nagel's Footnote Eleven Solve the Mind-Body Problem?" *Philosophical Issues* 20: 125–143.

Elwood, R. (2011). "Pain and Suffering in Invertebrates?" *ILAR Journal* 52: 175-184.

Elwood, R. (2012). "Evidence for Pain in Decapod Crustaceans." A*nimal Welfare* 21: 23-27

Feigl, H. (1958). "The 'Mental' and the 'Physical,'" In *Minnesota Studies in the Philosophy of Science, Volume 2: Concepts, Theories, and the Mind-Body Problem*, H. Feigl, M. Scriven, and G. Maxwell (eds.). Minneapolis: University of Minnesota Press, pp. 370-497

Feigl, H. (1967). *"The 'Mental' and the 'Physical': The Essay and A Postscript.* Minneapolis: University of Minnesota Press

Feinberg, T. and J. Mallatt (2016). *The Ancient Origins of Consciousness*. Cambridge MA: MIT Press.

Ginsburg, S. and E. Jablonka (2007). "The Transition to Experiencing: II. The Evolution of Associative Learning Based on Feelings," *Biological Theory* 2: 231–243.

Godfrey-Smith, P. (2016a). "Animal Evolution and the Origins of Experience," In D. Livingstone Smith (ed.), *How Biology Shapes Philosophy: New Foundations for Naturalism.* Cambridge: Cambridge University Press, pp. 51-71.

Godfrey-Smith, P. (2016b). "Individuality, Subjectivity, and Minimal Cognition," *Biology and Philosophy* 31: 775-796.

Godfrey-Smith, P. (2016c). "Mind, Matter, and Metabolism," *Journal of Philosophy* 113 (2016): 481-506.

Hill, C. S. and B. McLaughlin (1999). "There Are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy," *Philosophy and Phenomenological Research* 59: 445-454.

Hurley, S. (2001). "Perception and Action: Alternative Views." *Synthese* 129: 3–40.

Huxley, T. H. (1874). “On the Hypothesis that Animals are Automata, and its History,” *The Fortnightly Review* 16: 555–580.

Jackson, F. (1982). "Epiphenomenal Qualia," *The Philosophical Quarterly* 32: 127-136.

Jackson, F. (2003). "Mind and Illusion," in *Royal Institute of Philosophy Supplement* 53: 251-271.

Keijzer, F. (2015). "Moving and Sensing without Input and Output: Early Nervous Systems and the Origins of the Animal Sensorimotor Organization." *Biology and Philosophy* 30: 311–331.

Koonin, E. and W. Martin (2005). "On the Origin of Genomes and Cells Within Inorganic Compartments," *TRENDS in Genetics* 21: 647-54.

Kriegel, U. (2005). "Naturalizing Subjective Character," *Philosophy and Phenomenological Research* 71: 23-56

Kripke, S. A. (1972). *Naming and Necessity*. Cambridge MA: Harvard University Press.

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

Ludlow, P., Y. Nagasawa and D. Stoljar (eds.) (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument.* Cambridge MA: MIT Press.

Lewis, D.K. (1966). "An Argument for the Identity Theory," *Journal of Philosophy* 63: 17–25.

Lewis, D.K. (1994). "Reduction of Mind," In Samuel Guttenplan (ed.), *A Companion to Philosophy of Mind*. Oxford: Blackwell Publishers, pp. 412–431.

Lombrozo, T. (2006). *Understanding Explanation: Studies in Teleology, Simplicity, and Causal Knowledge.* PhD Dissertation, Psychology Department, Harvard University.

Maturana, H. and F. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living.* Boston Studies in the Philosophy of Science, vol. 42. Dordrecht: Reidel.

Merker, B. (2005). "The Liabilities of Mobility: A Selection Pressure for the Transition to Consciousness in Animal Evolution," *Consciousness and Cognition* 14: 89–114.

Montero, B. (2016). "What Combination Problem?" In G. Brüntrup and L. Jaskolla (eds.), *Panpsychism: Contemporary Perspectives*. Oxford: Oxford University Press.

Nagel, T. (1965). "Physicalism," *The Philosophical Review* 74: 339-356.

Nagel, T. (1974). "What is it Like to be a Bat?" *Philosophical Review* 83: 435-450.

Nagel, T. (1986). *The View from Nowhere.* Oxford: Oxford University Press.

Nagel, T. (2012). *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. Oxford: Oxford University Press.

Nilsson, D.-E. and N. Colley (2016) "Comparative Vision: Can Bacteria Really See?" *Current Biology* 26: R355–R376

O'Malley, M. (2014). *Philosophy of Microbiology*. Cambridge: Cambridge University Press.

O'Regan, K. and A. Noë (2001). "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences* 24: 939-73.

Perry, C., A. Barron, A., and K. Cheng (2013). "Invertebrate Learning and Cognition: Relating Phenomena to Neural Substrate," *WIREs Cognitive Science 2013*. doi: 10.1002/wcs.1248

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. (The 1999 Jean Nicod Lectures). Cambridge MA: MIT Press.

Place, U.T. (1956). "Is Consciousness a Brain Process?" *British Journal of Psychology* 47: 44–50.

Pradeu, T. (2011). "A Mixed Self: The Role of Symbiosis in Development," *Biological Theory* 6: 80-88.

Rozenblit, L., and F. Keil (2002). "The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth," *Cognitive Science* 26: 521-562.

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

Stoljar, D. (2015). "Russellian Monism or Nagelian Monism?" In T. Alter and Y. Nagasawa (eds.), *Consciousness in the Physical World: Perspectives on Russellian Monism*. Oxford: Oxford University Press, pp. 324-345/

Strawson, G. (2006). "Realistic Monism - Why Physicalism Entails Panpsychism," *Journal of Consciousness Studies* 13: 3-31.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Belknap Press.

Trestman, M. (2013). "The Cambrian Explosion and the Origins of Embodied Cognition," *Biological Theory* 8: 80–92.

Van Inwagen, P. (1990). *Material Beings*. Ithaca: Cornell University Press.

Von Holst, E. and H. Mittelstaedt (1950/1973). "The Reafference Principle (Interaction Between the Central Nervous System and the Periphery)," in *Behavioral Physiology of Animals and Man*: *The Collected Papers of Erich von Holst,* Vol. 1. Translated by Robert Martin. Coral Gables: University of Miami, 1973 (first published in *Die Naturwissenschften* 37 (1950): 464-476).