Qualia explained away

A Commentary on Daniel C. Dennett

David H. Baßler

In his paper "Why and how does consciousness seem the way it seems?", Daniel Dennett argues that philosophers and scientists should abandon Ned Block's distinction between access consciousness and phenomenal consciousness. First he lays out why the assumption of phenomenal consciousness as a second medium is not a reasonable idea. In a second step he shows why beings like us must be convinced that there are qualia, that is, why we have the strong temptation to believe in their existence. This commentary is exclusively concerned with this second part of the target paper. In particular, I offer a more detailed picture, guided by five questions that are not addressed by Dennett. My proposal, however, still resides within the framework of Dennett's philosophy in general. In particular I use the notion of intentional systems of different orders to fill in some details. I tell the counterfactual story of some first-order intentional systems evolving to become believers in qualia as building blocks of their world.

Keywords
Dispositions | Intentional systems | Predictive processing | Qualia | Zombic hunch

Commentator

David H. Baßler
davidhbassler@gmail.com
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Daniel C. Dennett
daniel.dennett@tufts.edu
Tufts University
Medford, MA, U.S.A.

1 Introduction

The first of Rapoport's Rules[1] for composing a critical commentary states that one should present the target view in the most charitable way possible (Dennett 2013a). Although I generally agree with many of Daniel Dennett's views, especially his argument against the existence of qualia (constituting the first part of the target paper), the diagnosis that there is the *zombic hunch*,[2] along with his strategy for explaining why it exists, the connection between qualia and predicted dispositions, was hard to grasp. Dennett presents the idea that when we talk about qualia, what we really refer to are our dispositions in earlier works (e.g., Dennett 1991). But the connection to predictive pro-

1 Dennett named these rules after social psychologist and game theorist Anatol Rapaport. They are not to be confused with another "Rapoport's Rule", named after Eduardo H. Rapaport (cf. Stevens 1989). Here is the full list of Dennett's Rapaport's Rules:
1. "You should attempt to re-express your target's position so clearly, vividly, and fairly that your target says, 'Thanks, I wish I'd thought of putting it that way.'"
2. "You should list any points of agreement (especially if they are not matters of general or widespread agreement)."
3. "You should mention anything you have learned from your target."
4. "Only then are you permitted to say so much as a word of rebuttal or criticism."
(Dennett 2013a, p. 33)

2 A philosophical zombie has nothing to do with any other sort of zombie. It behaves in *every* way like a normal person. The only difference is, that it lacks phenomenal experiences (though *ex hypothesi* it believes that is has phenomenal experiences). The zombic hunch is the intuition that a philosophical zombie would be different from us.

cessing is new (see also Dennett 2013b). There still seem to be some stepping stones missing, which I hope to fill in with my reconstruction. My goal is to provide a complete story that sticks as close to Dennett's argument as possible. This paper is not supposed to be a "rebuttal" or "criticism", but an "attempt to re-express [Dennett]'s position" (see footnote 1).

The structure of this commentary is as follows: in the first section I shall give a short outline of Dennett's explanation of why we have the zombic hunch. Since this involves the predictive processing framework, I shall give a very short introduction to this first. Following this, I present a short list of five questions that have not, in my opinion, yet been sufficiently addressed. In the second section I present an interpretation, or perhaps an extension, of Dennett's answers to these questions, by relying on the concept of an intentional system and using a strategy involving telling the counterfactual story of the evolution of some agents who end up believing in qualia (although *ex hypothesi* there are none). In the third section I shall analyze which features qualia should have, according to the beliefs of these agents, and show that there is at least a significant overlap with features many consider qualia to have.

I want to give a short justification for the unorthodox way of accounting for *beliefs* about x instead of for x's existence itself. This is a general strategy found in other areas of Dennett's work. For example, he has asked, "Why should we think there is intentionality although there is none?" (Dennett 1971), "Why should we believe there is a god although there is none?" (Dennett 2006), and "Why should we think there is a problem with determinism and free will although there is none?" (Dennett 1984, 2004). Dennett's philosophy can in parts be seen as a therapeutic approach to "philosopher's syndrome"—"mistaking failures of imagination for insights into necessity" (e.g., Dennett 1991, p. 401; Dennett 1998a, p. 366)—by making it easier to see why we are convinced of the existence of something, even when there are good reasons to believe that it doesn't exist.

I want to draw attention to Hume's *Of Miracles* (Hume 1995, X), where he states that the likelihood of a testimony about miracles being wrong is always greater than the likelihood of the miracle itself. This serves as a nice analogy for the case at hand: we might think of our own mind as a good "witness", but we already know too much about its shortcomings. So we should be suspicious when it cries out for a revolution in science or metaphysics, because this cry rests on the belief that something is missing, when no data but this very belief itself makes the demand necessary. Instead we should examine what else could have led our minds to form this conviction.

2 Dennett's proposal

In "Why and how does consciousness seem the way it seems?" Dennett gives an argument for why philosophers and scientists should abandon Ned Block's distinction between access consciousness and phenomenal consciousness, zombies, and qualia altogether. The argument is twofold: first Dennett lays down his argument for why the assumption of phenomenal consciousness as a second medium whose states are conscious experiences or qualia is "scientifically insupportable and deeply misleading" (Dennett this collection, section 2). It is insupportable because there is simply no need to posit such entities to explain any of our behavior, so for reasons of parsimony they should not be a part of scientific theories (see also Dennett 1991, p. 134). The assumption is deeply misleading because it makes us look for the wrong things, namely, the objects our judgments are about, rather than the causes of these judgments, which are nothing like these objects.

In a second step Dennett shows why creatures like us must be convinced that there are qualia, that is, why we have such a strong temptation to believe in their existence, *even though* there are no good reasons for this (Dennett this collection, section 2 and 3; other places where Dennett acknowledges this conviction, the zombic hunch, are Dennett 1999; Dennett

2005, Ch. 1; Dennett 2013a, p. 283). The following sections are exclusively concerned with this part of the target paper.

After completing the second step, Dennett explains why we ascribe qualia their characteristic properties—simplicity and ineffability (Dennett this collection, section 4 & 5). Although I also say something about this point (see section 4), Section 6 is an intuition pump (cf. e.g., Dennett 2013a) that will help the reader to apply Dennett's alternative view to the experience of colors.

Before I present a short outline of Dennett's second step, I want to briefly describe the predictive processing framework. This is necessary since both Dennett's argument as well as my reconstruction make use of this framework. I shall not go into details of hierarchical predictive processing (PP) accounts here, since at least three papers in this collection (Clark, Hohwy, and Seth), as well as the associated commentaries (Madary, Harkness, and Wiese), are concerned with this topic and also offer ample references for introductory as well as further reading. I will instead give a very short description of the points that are most relevant to Dennett's argument and recommend the above-mentioned papers and the references given there to the interested reader.

2.1 Predictive processing

In the PP framework, the brain refines an internal generative stochastic model of the world by continuously comparing sensory input (extero- as well as interoceptive) with predictions continuously created by the model. The overall model is spread across a hierarchy of layers, where the sensory layer is the lowest and each layer tries to predict (that is, to suppress) the activation pattern of the layer beneath it. The whole top-down activation pattern might be interpreted as a global hypothesis about the hidden causes of ongoing sensory stimulation. The difference between predicted and actual activation (*prediction error*) is what gets propagated up the hierarchy and leads to changes in the hypothesis. To be exact, this is only one possibility. Another is that this leads to an action that changes the input in such a way that the prediction is vindicated (*active* inference, see e.g., Friston et al. 2011). However, although this aspect of PP —that it provides one formally-unified approach to perception and action—is a strength of the framework, it is not important here, given the context of this commentary. These changes are supposed to follow Bayes' Theorem, which is why one might speak of Bayesian prediction (cf. e.g., Hohwy 2013).

The higher the layer in the hierarchy the more abstract the contents and the longer the time-scales or the predictive horizon. One example of a very abstract content is "only one object can exist in the same place at the same time" (Hohwy et al. 2008, p. 691, quoted after Clark 2013, p. 5).

One point to keep in mind is that, according to Hohwy (2014), this framework implies a clear-cut distinction between the mind and the world. That is, there is an *evidentiary boundary* between "where the prediction error minimization occurs" and "hidden causes [of the sensory stimulation pattern] on the other side" (Hohwy 2014, p. 7). I will come back to this point later in this commentary.

2.2 The outline of Dennett's argument

1. Our own dispositions, expectations, etc. are part of the generative self-model instantiated by our brains. "We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will expect next" (Dennett this collection, p. 5)
2. When our brains do their job (described in (1)) correctly, i.e., there are no prediction-error signals, we misidentify dispositions of the organism with properties of another object. For instance, instead of attributing the disposition to cuddle a baby correctly to the organism having the disposition, our brain attributes "cuteness" to the baby.[3] Color qualia

3 "Think of the cuteness of babies. It is not, of course, an 'intrinsic' property of babies, though it seems to be. […][W]e expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world with which we are in-

and other types of qualia also belong to this category.[4]
3. This means, under a personal level description, that we believe that there are properties *independent of the observer*, such as the cuteness of babies, the sweetness of apples, or the blueness of the sky, etc.
4. This is why it is so hard for us to doubt that qualia exist in the real world.

The crucial points seem to be (1) and (2). Before I lay out my interpretation I want to highlight some points that are not addressed in Dennett (this collection), but which are crucial if we are to have a complete picture. In the section *Our Bayesian brains*, I present a reconstruction that addresses these issues.

2.3 Five questions

1. **Why do we need to monitor our dispositions?** As noted in Dennett (2010), self-monitoring, in the sense of monitoring of our dispositions, values, etc., isn't needed unless one needs to communicate and to hide and share specific information about oneself at will. In his paper, Dennett does not address this issue, yet presupposes that "among the things in our *Umwelt* that matter to our well-being are ourselves". This is obvious if one reads "ourselves" as the motions of our bodies, but not so obvious if one includes things like "what we will think next, and what we will expect next", as Dennett does (Dennett this collection, p. 5). The next question is concerned with this latter form of self-monitoring:
2. **How is self-monitoring accomplished?** Hohwy (2014) refers to an evidential boundary in the predictive processing framework (see the section 2.1): there is a clear distinction between the mind/brain and the world (of which the body without the brain is a part), whose causal structure is yet to be revealed. Our expectations are part of our mind, which, if talk of the boundary is correct, does not have direct access to its own states *as* its own states—the mind is a black box to itself. So the prediction of its expectations needs to be indirect (just like the predictions of the causes of the sensory stimulation in general), and therefore the question arises how the self-monitoring of the mind is achieved according to Dennett. There is a further concern with self-monitoring, which one might call the "acquisition constraint" (cf. e.g., Metzinger 2003, p. 344):
3. **How did this self-monitoring evolve in a *gradual* fashion?** Large parts of Breaking the Spell are dedicated to making understandable how "belief in belief" could have evolved over the centuries, beginning long before the appearance of any religion. Dennett's goal here is quite similar: the explanation aims to make understandable how we came to believe in qualia, etc. But a step-by-step explanation is missing. I consider this form of the acquisition-constraint one of the most crucial for any satisfying explanation of this sort: each single step has to be understandable as one likely to have happened. One reason for this is that it would support a more fine-grained and mechanistic understanding; another is that it would satisfy the gradualism-constraint of Darwinism, which says that minds (just like anything else) "must have come into existence gradually, by steps that are barely discernible *even in retrospect*" (Dennett 1995, p. 200, emphasis in original).

teracting has the properties we expected it to have" (Dennett this collection, p. 5).

4 The intuition pump of Mr. Clapgras in Dennett's section 6 is there to make the point that colors can be seen as dispositional properties of the organism rather than as properties of perceptual objects, in the same way as cuteness. Whether one is convinced by this or not, the intuitive problem seems to be the same: science tells us there are no properties like cuteness or color, while the zombic hunch tells us that this cannot be true. A more detailed discussion can be found in Dennett (1991, p. 375). I will not go into this here, but for the sake of argument I shall assume that this admittedly counter-intuitive categorization is acceptable. The reader's willingness to accept it might be helped by the following point given by Nicholas Humphrey, which reminds us that although at first thought colors do not *seem* to have action-provoking effects (like cuteness or funniness), after second thought one might think differently:
"As I look around the room I'm working in, man-made colour shouts back at me from every surface: books, cushions, a rug on the floor, a coffee-cup, a box of staples—bright blues, reds, yellows, greens. There is as much colour here as in any tropical forest. Yet while almost every colour in the forest would be meaningful, here in my study almost nothing is. Colour anarchy has taken over." (Humphrey 1983, p. 149; quoted in Dennett 1991, p. 384).

Once we know why and how our brains accomplish the task of monitoring our dispositions and how they came to do so, one might still wonder why (as claimed in point 2, page 3) exactly these abstract properties of the organism would be misidentified as concrete properties of other things:

4. **Why do we misidentify our dispositions?** One of Dennett's central claims is that we misidentify our own dispositions, which leads to belief in qualia.[5] Although misidentification seems to be ubiquitous (see superstition, religion, magic tricks, the rubber hand illusion—Botvinick & Cohen 1998; and even full body illusions—Blanke & Metzinger 2009) it nonetheless requires a special explanation in each case: is this a shortcoming of a system that has no disadvantages, or is it even something that benefits the system in some way (cf. McKay & Dennett 2009)? Keeping this last possibility in mind one might ask:
5. **Why are we so attached to the idea of qualia?** There seems to be something more that leads people to believe in qualia. There is the intuition that without qualia we would be very different—we would be "mere machines", we could not *enjoy* things like a good meal or the smell of the air after it rains (a discussion of this characteristic of beliefs-about-qualia can be found in Dennett 1991, p. 383). Some might go further and say that our whole morality rests on the existence of qualia of pain and suffering (this worry is dealt with in Dennett 1991, p. 449). However, what I am concerned with here is not whether it is true that qualia are the basis of our morality, but why we should think them to be so. From the argument presented by Dennett it is not clear why we are so attached to the idea of qualia. It is not obvious why we do not react as disinterestedly to their denial as we did to the revelation that there is no ether.[6] But, as a matter of fact, we react differently: this is not like when any other entity, posited for theoretical reasons, is shown to not exist; it is as if without qualia we couldn't possibly be *us*.

3 An interpretation

3.1 Intentional Systems Theory

An important part of what follows is Intentional Systems Theory (IST). What is crucial here is that according to IST, all there is to being an agent in the sense of having beliefs and desires upon which to act is to be describable via a certain strategy: the *intentional stance*. The intentional stance is a "theory-neutral way of capturing the cognitive competences of different organisms (or other agents) without committing the investigator to overspecific hypotheses about the internal structures that underlie the competences" (Dennett 2009, p. 344). If one predicts the behavior of an object via the intentional stance, one presupposes that it is optimally designed to achieve certain goals. If there are divergences from the optimal path, one can, in a lot of cases, correct for this by introducing abstract entities or false beliefs. Since there are presumably no 100%-optimally-behaving creatures in the world, every intentional profile (a set of beliefs and desires), generated via adoption of the intentional stance, contains a subset of false beliefs.[7] It seems that humans have a "generative capacity [to find the patterns revealed by taking the intentional stance] that is to some degree innate in normal people" (Dennett 2009, p. 342). I will come back to this point and its connection to PP in the next section.

Let us assume for the sake of argument that IST gives a correct explanation of what it is to be an agent (in the sense of someone who has beliefs and desires and acts according to

5 What qualia are [...] are just those complexes of dispositions. When you say 'This is my quale,' what you are singling out, or referring to, whether you realize it or not, is your idiosyncratic complex of dispositions. You seem to be referring to a private, ineffable something-or-other in your mind's eye, a private shadeshade of homogeneous pink, but this is just how it seems to you, not how it is. (Dennett 1991, p. 389).

6 This property of the beliefs is acknowledged in Dennett (2005), p. 22, fn 18: "[The Zombic Hunch] is visceral in the sense of being almost entirely arational, insensitive to argument or the lack thereof".

7 See Dennett (1987) for an elaborate discussion of the intentional stance and its implications, Dennett 1998b for the ontological status of beliefs and desires, Bechtel (1985) for another interesting interpretation, and Yu & Fuller (1986) for a discussion of the benefits of treating beliefs and desires as abstracta.

them), and that PP allows us to see how an agent can be implemented on the "algorithmic level"(see Dennett's discussion in Dennett 1987, p. 74, where he refers to the IST as a "competence model"). Whenever I say that an agent believes, wants, desires, etc. something I mean it in exactly the sense found in IST.

Intentional systems can be further categorized by looking at the content of their beliefs, e.g., a second-order intentional system is an intentional system that has beliefs and/or desires about beliefs and/or desires, that is, it is itself able to take an intentional stance towards objects (Dennett 1987, p. 243). A first-order intentional system has (or can be described as having) beliefs and desires; a second-order intentional system can ascribe beliefs to others and itself. If something is a second-order intentional system it harbors beliefs such as "Peggy believes that there's cheese in the fridge". But taking the intentional stance towards an object is an ability that comes in *degrees*. I now want to describe what one might call an intentional system of *1.5th order*, an intermediate between first- and second-order intentional systems. This is a system that is not able to ascribe full-fledged desires and beliefs with arbitrary contents to others or itself. We, as intentional systems of high order, have no difficulty in ascribing beliefs and desires with very arbitrary contents, such as "She wants to ride a unicorn and believes that following Pegasus is a good way to achieve that goal". But the content of beliefs and desires that such an intentional system of 1.5th order can ascribe should be constrained in the following way:

1. An intentional system of 1.5th order is able to ascribe desires only in a very particular and concrete manner, i.e., actions that the object in question wants to perform with certain particular existing objects, that the system itself knows about (e.g., the desire to eat the carrot over there), but not goals directed at nonexistent objects, described by sentences like "he wants to build a house", or objects the ascriber itself does not know about.
2. It is only able to ascribe beliefs to others that it holds itself. That means it is able to take the basic intentional stance with the default assumption that the target object in question believes whatever is true (if we assume the ascriber's beliefs are in fact all true), but lacks the ability to correct the ascriptions if it leads to wrong predictions for the behavior of the target. A real-world example can be found in Marticorena et al. (2011): rhesus macaques in a false belief task can correctly predict what a person will do, given that the person knows where the object is hidden and they have seen the person getting to know this. They can also tell when a person doesn't have the right knowledge, but they cannot use this information to make a prediction about where the person will look.

The implementation of such an intermediate between first- and second-order intentional systems can be easily imagined following predictive coding principles, as I will soon show. Following this, I argue that this sets down the basic fundaments for systems evolving from this position to be believers in qualia, etc.

The reason for introducing this idea is that I want to show how, given predictive processing principles and a certain selection pressure, a 1.5th-order-intentional-system might develop from a first-order-intentional-system. In a next step, I will argue that under an altered selection pressure such a system might become a full-fledged nth-order-intentional-system, where n is greater or equal to two. Systems evolving in such a way, as I will describe, are bound to believe in the existence of something like qualia. In some sense this is only a just-so story, but the assumed selection pressures are very plausible, and the empirically-correct answer might not be too far away from this.

3.2 Our Bayesian brains[8]

To see how the pieces fit together imagine the situation of some first-order intentional systems, agents, which are the first of their kind. They act according to their beliefs and desires. They do so because the generative models im-

8 This section takes strong inspiration from Wilfrid Sellars' section "Our Rylean Ancestors" in Sellars (1963, p. 178).

plemented in their brains generate a sufficient number of correct predictions about their environment for them to survive and procreate. They do a fairly good job of avoiding harms and finding food and mates. Since they are first-order intentional systems, the behavior of their conspecifics amounts to unexplained noise to them, because they are unable to predict the patterns of most of their behavior (which is what makes them *merely* first-order intentional systems), though they might well predict their behavior as physical objects, e.g., where someone will land if she falls off a cliff, for instance.

When resources are scarce, this leads to competition between these agents and it becomes an advantage to be able to predict the behavior of one's conspecifics. This behavior is by definition pretty complex (they are intentional systems), but one can get some mileage out of positing the following regularity: some objects in the world have properties that lead to predictable behavior in agents, e.g., if there is an apple tree this will lead to the agents approaching it, if they are sufficiently near, etc., whereas if there is a predator, they will run from it, etc. Their model of the world is populated by properties of items that allow the (arguably rough) predictions of *agent behavior*. One might indeed say that the desires of the agents are *projected* [9] onto the world.[10] Those who acquire this ability are now 1.5th order intentional systems (see above; monkeys and chimpanzees might turn out to be such, see Roskies this collection).[11] However, findings in this area are controversial. See Lurz 2010), since they can predict the behavior of others, given that their behavior is indeed explainable via reference to actually-existing objects, such as apples or potential sexual partners. In addition to these properties, there is a new category of objects in "their world": beings that react to these properties in certain ways.[12]

In a next step we might suppose that a system of communication or signaling evolves (the details are not important), turning our intentional systems of 1.5th order into communicative agents. As communicative beings they have an interest in hiding and revealing their beliefs according to the trustworthiness of others and their motives (cf. Dennett 2010). That is, any of those beings needs to have access to what it itself will do next, so that they can hide or share this information, depending on information about the other. One might think of hiding the information about one's desire to steal some food, and so on.

This is a situation where applying the predictive strategy that was formerly only used to explain the behavior of others to *oneself* becomes an advantage for each of the agents.[13] Agents like this believe in the existence of a special kind of special kind of properties, i.e., they predict their *own* behavior on the basis of generative models that posit such properties: they believe that they approach apples *because* they are *sweet*, cuddle babies *because* they are *cute*, laugh about jokes *because* they are *funny*. Applying the strategy to their own behavior puts them in the same category (according to the generative model) as the others: they are unified objects that react to cer-

9 What I mean by "project" is that instead of positing an inner representation whose content is "I (the system in question) want to eat that apple" and whose function is a desire, along with correct beliefs about the current situation, what is posited is an eat-provocative property of the apple itself. Both theoretical strategies allow for the prediction of the same behavior. The crucial difference is that attributing new properties to objects that are already part of the model is a simpler way of extending the model than positing a complex system of internal states to each agent. Thus it is also more likely to happen. It's definitely much simpler than extending the model to incorporate all the entities that explain the behavior on a functional level (i.e., all the neurons, hormones etc.). It is successful to the same extent the intentional stance is successful, that is, in an arguably noisy way, but still successful enough to gain an advantage (since *ex hypothesi* all the conspecifics are intentional systems).

10 This is very close to Gibson's affordances (e.g., Gibson 1986) in that "values and meanings are external to the perceiver" (p. 127) and in a couple of other respects (ibid.). It is, however, different in that the postulated properties serve to predict the behavior of *others* and not to guide the behavior of the organism itself. For the relation between Gibsonian affordances and predictive processing see e.g., Friston et al. (2012).

11 "[R]ecent work on non-human primate theory of mind suggests that monkeys and chimpanzees have a theory of mind that represents goal states and distinguishes between knowledge and ignorance of other agents (the presence and absence of contentful mental representations), even if it fails to account for misrepresentation." (Roskies this collection, p. 12).

12 The selection of goals and other cognitive capabilities, etc., is all placed outside of the target object (see footnote 9). It will approach the object that has the highest attraction value, given that there is no object with a higher repulsion value, i.e., there is no internal selection process represented *as* internal selection. What makes other agents special objects, in this model, is that they react to properties that no other things react to, not that they have an internal life that is somehow special.

13 Notice that according to PP, there is no shortcut to be taken: the mind is a black box to itself—it has to infer its own properties just as any others.

tain properties, not a bunch of cells trying to live among one another.[14]

The agent-models of these beings might improve by integrating the fact that sometimes it is useful to posit non-existing entities or omit existing entities in order to predict the behavior of a given conspecific (think of subjects in the false belief-task looking in the wrong box). By this the concept of (false) beliefs arises. One can imagine how they further evolve into full-fledged second and higher-order intentional systems, in an arms-race for predicting their fellows.[15]

A further step: they develop sciences like we did and will come to have a scientific image of the world, which contains no special simple properties of objects that cause "agents" to behave in certain ways. They come to the conclusion that the brain does its job without taking notice of properties like cuteness or redness, "instead relying" on computations, which take place in the medium of spike trains and nothing but spike trains (cf. target, section 1). Their everyday predictions of others and most importantly of themselves still rely on the posited properties. And some might wonder whether there isn't something missing from the scientific image.

According to the scientific image, they, as biological organisms, react to photons, waves of air, etc., but these are not the contents of their own internal models employed in solving the continuous task of predicting themselves. The simplest things they react to seem to be colors and shapes, (perceived) sounds, etc. The reaction towards babies is explained via facial proportions and the like, but this is far from what their generative models "say", which is "the reaction to babies is caused by their cuteness".

They begin to build robots, which react to babies like they do. They say things like, "all this robot reacts to are the patterns in the baby's face, the proportions one can measure; but although it reacts like we do, it does not do so because of the baby's cuteness". Of course only non-philosophers might say that science misses a property of the baby, but philosophers still see that there is *something* missing, and since cuteness is not a property of the outside world, they conclude that it must be a property of the agents themselves.

This seems to me to be the current situation. We have the zombic hunch because it seems to us that there is something missing and it seems so because our generative models are built upon the assumption that there are properties of things out there in the world to which systems like us react in certain ways. We never consider others like us to be zombies because they are agents like us or better: we are systems like them. We dismiss robots because we know they can only react to measurable properties, which do not *seem* to us to be the direct cause of *our* behavior.

4 An analysis

Is it true that properties such as cuteness do not correspond to anything? In a sense it is false to deny that any such correspondence exists: such properties do correspond to the cuddle-provocativeness of a baby, the eating-provocativeness of an apple, etc., *as a cause of the behavior of agents*. They are "lovely" properties (Dennett 1991, p. 379), and there is a way to measure them: we can use ourselves as detectors. But the reason we, intuitively, do not accept a robot as a subject like ourselves is because we know how the robot does it: we know that it calculates, maybe even in a PP-manner —we know that it does not react directly to the properties that seem to exist and that seem to count. Neither do we, or the beings described above. But their own prediction of themselves treats such complex properties as simple, because there is nothing to be gained by being more precise than is necessary for *sufficiently* accurate prediction.[16]

This is my reconstruction of Dennett's claim that the mind projects its dispositions

14 This is where one might speak of the origin of a self-model (Metzinger 2003) in some sense, where there is not only a model of the body (built up by proprioceptive inputs) but also a model of the self as having (primitive) goals, at least in any given moment.

15 Maybe language plays an important part in this further development as an external scaffold (cf. Clark 1996; Dennett 1994). One fact supporting this view is that monkeys do not seem to be able to understand the concept of false belief (and therefore the concept of belief) (cf. Marticorena et al. 2011, but also Lurz 2010 for an overview of this debate).

16 This is also true of affordances (see e.g., Gibson 1986, p. 141).

onto the world via Bayesian prediction. I want to draw attention to some of the features ascribed to those properties that this story predicts:

1. **These properties are "given directly" to a person**
   The overall generative model depicts the whole organism as a unified object that reacts *directly* to the posited properties in the world. Any system that represents itself in such a way is bound to believe that there are properties of the world given directly to the object, which it takes to be itself. In subpersonal terms this object and these properties, as well as their relation to each other, are postulated entities that explain the sensory input. For instance, the fact that others talk about the system as someone with beliefs and desires (which is rooted in the same principle) can be explained by predicting itself in the same way.
2. **These properties are irreducible to physical, mechanical phenomena.**
   Since the generative model does not depict these properties as built up from simpler ones, but simply posits them to predict lower-level patterns, these properties don't seem (to the system) to be reducible to other properties.
3. **These properties are atomic, i.e., unstructured.**
   There are as many posited properties as there are distinct dispositions to be tracked. This also explains why one can learn to find structure in formerly unstructured qualia (cf. Dennett 1991, p. 49) once new discriminative behavior is learned.
4. **These properties are important to our lives/beings as humans/persons**
   This felt importance is obvious, given the putative role they play in the explanation provided by the generative model. These properties seem to be the causes of all our behavior: if one did not feel the painfulness of a pain, one would not scream; if one did not sense the funniness of a joke, one would not laugh, etc. Since the model is still needed for interacting with others, despite theoretical advances in the sciences this felt importance of qualia to our lives is very difficult to overcome.
5. **These properties are known to every living human being; it is not possible to sincerely deny their existence**
   This is due to the fact that our brains predict the behavior of others via a model that posits direct interaction between "agents" and first-order, non-relational object properties—the entities that are then named "qualia".

This list has considerable overlap with lists of features ascribed to qualia (e.g., Metzinger 2003, p. 68; Tye 2013), lending support to the thesis that we don't need a revolution in science to accommodate qualia, but rather a change in perspective: we might look at the creatures described above and see that "[t]hey are us" (Dennett 2000, p. 353).

5 Conclusion

I have given an interpretation of Dennett's theory of why there seems to be something more to consciousness than science can explain. My aim was to thereby address crucial questions, while sticking as closely to Dennett's philosophy as possible. The answer is a just-so story that shows how (plausible) selection pressures lead to beings that cannot help but believe that they are *more* than just "moist robots" (Dennett 2013a, p. 49)—because some important entities seem to be missing from the scientific description.

This story answers the questions why and how beings like us monitor their dispositions, and how this ability could have evolved. It also offers an answer as to why we don't recognize them as representations of our dispositions and why qualia are unlike other theoretical entities in that they are important for what we consider ourselves to be. The notion of an intermediate between first- and second-order intentional systems was introduced as a new conceptual instrument for satisfying the acquisition constraint and to lay the fundaments for the belief in mind-independent simple properties that dir-

ectly cause the behavior of agents. This in turn is the basis for the belief in qualia as intrinsic properties of experience.

This story might not provide an "insight into necessity" (cf. Dennett 1991, p. 401), but I am happy if it contributes to showing and clarifying a possibility: although it may *seem* that our best hypothesis for accounting for our belief in qualia is that they actually exist, this hypothesis might still be explained away.

Acknowledgements

I want to thank Thomas Metzinger and Jennifer Windt for the unique opportunity to participate in this project. I am also very grateful for the helpful remarks they and two anonymous reviewers gave to an earlier version of this paper.

References

Bechtel, W. (1985). Realism, instrumentalism, and the Intentional Stance. *Cognitive Science*, *9* (4), 473-497. 10.1207/s15516709cog0904_5

Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13* (1), 7-13. 10.1016/j.tics.2008.10.003

Botvinick, M. & Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature*, *391* (756). 10.1038/35784

Clark, A. (1996). Linguistic anchors in the sea of thoughts. *Pragmatics & Cognition*, *4* (1), 93-103. 10.1075/pc.4.1.09cla

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

——— (2015). Embodied prediction. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, *68*, 87-106.

——— (1984). *Elbow room. The varieties of free will worth wanting.* Oxford, UK: Clarendon Press.

——— (1987). *The Intentional Stance.* Cambridge, MA: MIT Press.

——— (1991). *Consciousness explained.* New York, NY: Back Bay Books/Little, Brown and Company.

——— (1994). The Role of Language in Intelligence. In J. Khalfa (Ed.) *What is Intelligence? The Darwin College Lectures.* Cambridge, UK: Cambridge University Press. 10.1075/pc.4.1.09cla

——— (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York, NY: Simon Schuster Paperbacks.

——— (1998a). Self-portrait. *Brainchildren: Essays on designing minds* (pp. 355-366). Cambridge, MA: MIT Press.

——— (1998b). Real patterns. *Brainchildren: Essays on designing minds* (pp. 95-120). Cambridge, MA: MIT Press.

——— (1999). The zombic hunch: Extinction of an intuition. *Royal Institute of Philosophy Millennial Lecture*

——— (2000). With a little help from my friends. In D. Ross, A. Brooks & D. Thompson (Eds.) *Dennett's Philosophy: A Comprehensive Assessment* (pp. 327-388). Cambridge, MA: MIT Press.

——— (2004). *Freedom Evolves.* London, UK: Penguin Books.

——— (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness.* Cambridge, MA: MIT Press.

——— (2006). *Breaking the spell. Religion as a natural phenomenon.* New York, NY: Penguin.

——— (2009). Intentional Systems Theory. In B. P. McLaughlin, A. Beckermann & S. Walter (Eds.) *The Oxford handbook of philosophy of mind* (pp. 339-349). Oxford, UK: Oxford Handbooks Online.

——— (2010). The evolution of why. In B. Weiss & J. Wanderer (Eds.) *Reading Brandom: On making it explicit* (pp. 48-62). New York, NY: Routledge.

——— (2013a). *Intuition pumps and other tools for thinking.* New York, NY: W. W. Norton & Co..

——— (2013b). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences*, *36* (3), 29-30. 10.1017/S0140525X12002208

——— (2015). Why and how does consciounsess seem the way it seems? In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104* (1-2), 137-160. 10.1007/s00422-011-0424-z

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8* (1), e1002327-e1002327. 10.1371/journal.pcbi.1002327

Gibson, J. J. (1986). *The ecological approach to visual perception.* Hillsdale, NJ: Erlbaum.

Harkness, D. (2015). From explanatory ambition to explanatory power-A commentary on Jakob Hohwy. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Hohwy, J. (2013). *The predictive mind.* Oxford, UK: Oxford Univeristy Press.

——— (2014). The self-evidencing brain. *Noûs*, *online.* 10.1111/nous.12062

——— (2015). The neural organ explains the mind. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Hohwy, J., Ropstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108* (3), 687-701. 10.1016/j.cognition.2008.05.010

Hume, D. (1995). *An inquiry concerning human understanding.* London, UK: Pearson.

Humphrey, N. (1983). *Consciousness regained.* Oxford, UK: Oxford University Press.

Lurz, R. W. (2010). Belief attribution in animals: On how to move forward conceptually and empirically. *Review of Philosophy and Psychology*, *2* (1), 19-59. 10.1007/s13164-010-0042-z

Madary, M. (2015). Extending the explanandum for predictive processing-A commentary on Andy Clark. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Marticorena, D. C.W., Ruiz, A. M., Mukerji, C., Goddu, A. & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, *14* (6), 1467-7687. 10.1111/j.1467-7687.2011.01085.x

McKay, R. T. & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*, 493-561. 10.1017 / S0140525X09990975

Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

Roskies, A. (2015). Davidson on believers: Can non-linguistic creatures have propositional attitudes? In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Sellars, W. (1963). *Science, perception and reality.* London, UK: Routledge & Kegan Paul Ltd..

Seth, A. (2015). The cybernetic bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Stevens, G. C. (1989). The latitudinal gradients in geographical range: How so many species co-exist in the tropics. *American Naturalist*, *133* (2), 240-256.

Tye, M. (2013). Qualia. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*

Wiese, W. (2015). Perceptual presence in the Kuhnian-Popperian Bayesian brain—A commentary on Anil Seth. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Yu, P. & Fuller, G. (1986). A critique of Dennett. *Synthese*, *66* (3), 453-476. 10.1007/BF00414062