Review of Philosophy and Psychology (2022) 13:427–453
https://doi.org/10.1007/s13164-021-00537-6

Illusionism about Phenomenal Consciousness: Explaining the Illusion

Daniel Shabasson [1]

Accepted: 4 March 2021 / Published online: 1 April 2021

Abstract

According to illusionism, phenomenal consciousness is an introspective illusion. The *illusion problem* (Frankish 2016) is to explain the cause of the illusion, or why we are powerfully disposed to judge—erroneously—that we are phenomenally conscious. I propose a theory to solve the illusion problem. I argue that on the basis of three hypotheses about the mind—which I call *introspective opacity*, *the infallibility intuition*, and *the justification constraint*—we can explain our disposition, on introspection, to draw erroneous unconscious inferences about our sensory states. Being subject to the illusion of phenomenal consciousness consists in having this disposition. I explain our 'problem intuitions' about consciousness (Chalmers 2018)—that our sensory states bear phenomenal properties that are qualitatively *like something* with which we are directly acquainted that is ineffable, atomic, intrinsic (non-relational), private, and non-physical. I also address the *illusion meta-problem* (Kammerer 2019a), which is to explain why illusionism seems especially counterintuitive.

1 Introduction

Illusionism avoids *the hard problem of consciousness* by positing that phenomenal consciousness is an introspective illusion. However, illusionism is highly counterintuitive, for it powerfully seems to us that we are phenomenally conscious—that it is qualitatively 'like something' to see red. To make illusionism plausible, its proponents must solve *the illusion problem* (Frankish 2016: 12), which is to explain the cause of the illusion. Based on three hypotheses about the mind, I propose a solution to the illusion problem, henceforth 'the inference theory'.

My first hypothesis (§2) is *introspective opacity*, according to which a subject lacks introspective access to the process in her brain that forms her introspective judgments classifying her sensory states. **My second hypothesis** (§3), *the infallibility intuition*, is a subject's disposition to judge her introspective judgments classifying her sensory

✉ Daniel Shabasson
DanShab@live.com

[1] Independent Scholar, Larchmont, New York, USA

states to be *infallible*. **My third hypothesis** (§4) is *the justification constraint*: if a subject judges that *p*, she necessarily believes, either explicitly or tacitly, that she is justified in judging that *p*. Thus, the absurdity of: *p, but I have no reason to believe that p*.

In §5, I set out the inference theory based on these hypotheses, arguing that being subject to the illusion of phenomenal consciousness consists in an introspecting subject's disposition to draw *erroneous unconscious inferences* about her sensory states.

In §6, I argue that the inference theory avoids *the phenomenal representation problem* (Frankish 2016: 18–20, 2018: slides 2 & 16), a worry for illusionist theories of consciousness.

In §7, I argue that the inference theory explains our 'problem intuitions' about consciousness (Chalmers 2018: 11). These include the intuitions that our sensory states bear phenomenal properties, that through introspection we are *directly acquainted* with their essences, or 'what they are like', and that what they are like is *ineffable*, *atomic, intrinsic* (*non-relational*), *private*, and *non-physical*.

In §8, I argue that the inference theory solves Kammerer's *illusion meta-problem* (Kammerer 2019a), which is to explain why illusionism seems especially counterintuitive.

In §9, I argue that as a *rich-illusion view* (Kammerer 2019c), the inference theory accounts for the cognitive impenetrability of the illusion.

2 Introspective Opacity

When a subject views a ripe tomato, a neural signal is transmitted into her brain, which, under normal conditions,[1] engenders what I shall call a *red* state*. This is the sensory state in the brain—entirely physical in nature—generated upon viewing a red object under normal conditions. A subject undergoing a red* state would typically form a perceptual judgment, reporting that the viewed object, e.g., a ripe tomato, is red-coloured. Upon introspecting this red* state, she would report it using phenomenal language by saying her experience is qualitatively *like something*, e.g., that she is 'experiencing a subjectively red feel'.[2]

Red* state

The sensory state—entirely physical in nature—engendered in the brain when a subject views a red object under normal conditions. A subject undergoing a red* state would typically form a perceptual judgment, reporting that the viewed object, e.g., a ripe tomato, is red-coloured. A subject would introspectively report the state using phenomenal language by saying, e.g., she is 'experiencing a subjectively red feel'.

[1] 'Normal conditions' obtain when there are normal viewing conditions and no failure of the sensory organs.
[2] Some subjects might say instead, e.g., 'this looks red to me'. There are few standard locutions to speak about putative subjective feels.

Although a subject may use phenomenal language to introspectively report her red* state, there are no phenomenal states or properties. While sensory states can be red* or green*, nothing is *phenomenally red* or *phenomenally green*, and phenomenal properties such as *phenomenal redness* or *phenomenal greenness* (henceforth 'p-redness' and p-greenness') do not exist.

Phenomenal properties

Phenomenal properties, a.k.a. 'qualia', 'raw feels', or 'subjective feels', are non-existent properties such as *p-redness* or *p-greenness* that a subject erroneously judges to be instantiated and to be qualitatively *like something* with which she is directly acquainted that is ineffable, atomic, intrinsic (non-relational), private, and non-physical.[3]

Each type of sensory state has unique physical/neural properties. Thus, a red* state will differ physically/neurally from a green* state.[4] A subject lacks introspective access to details of the physical/neural constitution of these states. However, she does have access, i.e., she is *access conscious*, with respect to the *types* of sensory states they instantiate, such that she can report this information (typically using phenomenal language), forming *introspective classificatory reports*, or 'ICRs'. For example, a subject might introspectively report the instantiation of her red* state by saying she is 'experiencing a subjectively red feel'.

Introspective classificatory report or 'ICR'

The introspective report, typically expressed in phenomenal language, that a sensory state *token* instantiates a particular sensory state *type*.

An ICR *i* uttered by subject *s* about her visual sensory state token *t* is *accurate* **iff** *i* correctly identifies the colour whose perception causes, under normal conditions, the instantiation in subject *s* of sensory state tokens of the type instantiated by *t*. For example, a subject's ICR about her red* sensory state that she is 'experiencing a subjectively *red* feel' would be *accurate*, for the instantiation of red* sensory states in that subject is indeed caused by the perception of red objects under normal conditions.[5] On the other hand, her ICR about her red* sensory state that she is 'experiencing a subjectively *green* feel' would be *inaccurate*, for the instantiation of red* sensory states in that subject is *not* caused by the perception of *green* objects under

[3] Untutored subjects may be unaware they have intuitions about ineffability, privacy, non-physicality, etc., which I discuss in §7. Awareness of such intuitions must typically be elicited through intuition pumps or thought experiments, e.g., the inverted spectrum. Although there is some debate about how widely these intuitions are shared, I shall presuppose, along with Chalmers (2018) and most illusionists, that they are widely if not universally shared.

[4] I suspect that sensory states are multiply realizable, neurally instantiated differently across individuals.

[5] The capacity to make ICRs does not require possession of a colour-word vocabulary. In principle, an agent could identify a colour in terms of the objects typically bearing the colour, or by ostension to a coloured object as an instance of the colour. Being in a particular sensory state is also tied, in each individual, to specific affordances, behavioural repertoires, inferential dispositions, and associations by which a subject might identify the sensory state type that a sensory state token instantiates.

normal conditions.[6] The *content* of an ICR is the classificatory claim it makes about the sensory state token it is about. That is, an ICR expresses the proposition that the sensory state token it is about instantiates a particular sensory state type.

How, according to the inference theory, does a subject form ICRs about her sensory states? A brain process sensitive to the physical/neural properties of sensory states 'reads' their physical/neural signatures, distinguishes and classifies the states by type, and makes this classificatory information available to the subject for report. However, the subject lacks introspective access to the physical/neural constitution of her sensory states and the operations of the brain process reading their signatures. She has access only to the *results* of this brain process, i.e., the information about the classifications of her sensory states by type. This lack of introspective access to the workings of brain processes is *introspective opacity.*

Introspective opacity

A brain process is introspectively opaque *iff* the subject cannot learn about it through introspection.[7] With respect to an ICR, a subject lacks introspective access to the physical/neural constitution of her sensory states and the operations of the process in her brain reading their physical/neural signatures.

Explaining the operations of this introspectively opaque brain process is an *easy problem* task for cognitive science, for it is strictly a question about information-processing mechanisms.

Closely related to the ICR is the *introspective classificatory judgment* or 'ICJ'. Whereas an ICR is a report *actually* articulated by a subject, an ICJ is the subject's *disposition* to issue an ICR about her sensory state.

Introspective classificatory judgment or 'ICJ'

A subject's *disposition* to issue an ICR about a sensory state.

For example, a subject forms the ICJ *I am experiencing a subjectively red feel* if she is merely *disposed* to report that she is experiencing a subjectively red feel, whether or not she actually issues this report (and without regard to the particular language she might use to do so).

3 The Infallibility Intuition

When a subject views a ripe tomato and a red* sensory state is instantiated, she typically forms the perceptual judgment that the tomato is red-coloured. She treats this judgment as highly reliable but fallible, realizing she may have misperceived the tomato's colour. By contrast, she intuits her introspective classificatory judgment about having a red* sensory state—the ICJ *I am experiencing a subjectively red feel*—to be

[6] Given the reliability of introspection, inaccurate ICRs would likely be rare.

[7] That is: for any brain process p and any subject s: p is introspectively opaque iff s cannot learn about the operations of p via introspection.

infallible. She intuits: *I may be wrong about the tomato being red, but I cannot be wrong about it subjectively seeming red to me.* This is *the infallibility intuition*.

The Infallibility intuition

A subject intuits her ICJs about the classifications of her sensory states to be infallible.[8,9]

The infallibility intuition is false, for no judgment is infallible. Conceivably, the brain process reading the physical/neural signature of a sensory state could *mis*read it, generating an inaccurate ICJ/ICR, although such errors would likely be rare.

ICJs and perceptual judgments differ with respect to their intuitive revisability. Perceptual judgments are intuitively revisable, for one might plausibly say: 'I judged at time T^1 that the tomato was red-coloured, but now I judge at time T^2 that it is orange-coloured, so my first judgment was erroneous'. The perceptual judgments at T^1 and T^2 are incompatible, for the tomato is either red-coloured or orange-coloured; the agent must revise one of her judgments. By contrast, ICJs are intuitively non-revisable, for no one would say: 'The tomato subjectively seemed red to me at time T^1, but now it subjectively seems orange to me at time T^2, so in fact it did not subjectively seem red to me at T^1'. The ICJs are about distinct subjective seeming episodes at T^1 and T^2 and do not contradict one another.

Both ICJs and perceptual judgments are introspectively opaque, meaning that a subject cannot learn, through introspection, about the brain processes forming them. Importantly, introspective opacity differently impacts a subject's ability to check her perceptual judgments and her ICJs. It does *not* impair her ability to check (and, if appropriate, revise) perceptual judgments. She might, for example, take a second look at a tomato to check her initial judgment that it was red-coloured. She would *not* need to introspect the brain process forming that judgment in order to check it. By contrast, the introspective opacity of ICJs blocks the subject's ability to check them. To check an ICJ, she *would* have to check the brain process that formed it for errors. Taking a second look at the object to check its colour would be to no avail, for, as discussed in §2 *supra*, the accuracy of an ICJ is not a function of the object's colour but rather depends on whether the subject's brain process has formed a correct judgment

[8] I deny that agents typically conceptualize ICJs under the mode of presentation *disposition to make an introspective classificatory report about a sensory state*, even if that is what ICJs in fact are. It is a truism that agents may conceptualize entities under erroneous modes of presentation. Therefore, I deny that, if asked to say whether their introspective classificatory judgments were infallible, agents would typically say they are. In fact, subjects do not typically think about their brain states and would likely have no idea what introspective classificatory judgments are. Nevertheless, I claim that agents do intuit their judgements about how things seem to them subjectively to be infallible, and these judgments are in fact classificatory judgments about their sensory states, even if agents may not typically conceptualize them this way.

[9] Whether we intuit our introspective judgments about our *emotions* to be revisable is a complex question. Judgments about emotions tend to express our beliefs about their causes and their connection to other emotions. Emotional states are typically non-atomic and complex, consisting of various parts (perhaps atomic ones) put together to form unique wholes. A judgment about an emotional state will be far more complex than one about a simple and atomic quale of a colour. Therefore, I focus on the simpler case of perceptual and sensory judgments.

about the sensory state type instantiated by her sensory state token. However, due to introspective opacity, the subject cannot introspectively check whether her brain process has made an error. And at the same time, she cannot check her ICJ *non*-introspectively (apart from, at least in theory, using a brain scan to do so—a prospect I shall revisit two paragraphs below).

I shall now propose a speculative hypothesis about the cause of the infallibility intuition. An intuition of *fallibility* arises vis-à-vis a class of judgments when a subject: (1) has the means, in principle, to check them, and (2) has the behavioural disposition to implement this means to do so. For example, a perceptual judgment is intuited as fallible because a subject has the means to check it and the behavioural disposition to do so, e.g., by taking a second look at the object the judgment is about. An *infallibility intuition* arises vis-à-vis a class of judgments when a subject: (1) lacks the means to check them, and (2) lacks the behavioural disposition to do so. Perceptual judgments and ICJs differ in these characteristics, and this explains the intuitive fallibility of the former and the intuitive infallibility of the latter. We have the means and the behavioural dispositions to check perceptual judgments, but we lack them with ICJs.[10,11]

Of course, as mentioned two paragraphs above, a subject might (at least in theory) use a brain scan to check her ICJs non-introspectively. However, I hypothesize that our intuitions about fallibility/infallibility are generated in a *cognitively impenetrable module* in the brain (i.e., a module informationally encapsulated with respect to higher-level cognition involved in theoretical belief formation). In this module, there are no mechanisms to check an ICJ, and therefore, no behavioural dispositions grounded in the module to do so. The informational encapsulation of the module means that no higher-level reasoning or beliefs can affect its operations. Therefore, a subject learning from a brain scan or through theorizing about the mind that her ICJs are fallible might shake her *belief* that they are infallible, but it would leave her *intuition* that they are infallible untouched.[12] In other words, beliefs about fallibility/infallibility formed in module(s) of the brain responsible for higher-level cognitive functions—such as those beliefs generated upon viewing a brain scan or involved in theorizing about the mind—cannot penetrate or affect the operations of the more primitive module in which the infallibility intuition is generated, such that it might counteract that intuition. While this hypothesis is speculative,

[10] Pereboom (2011: 23–24) likewise suggests the infallibility intuition arises from the fact that we cannot check our introspective judgments**.** See also Frankish (2019): '…we find it hard to doubt what introspection tells us. But this might be because there is no easy way of checking its accuracy'.

[11] One possible objection to this explanation of the infallibility intuition: we lack the means to check our judgment that we are not brains in vats, yet we do not intuit that we know infallibly that we are not brains in vats. However, I think the objection fails. Being a brain-in-a-vat would involve being subject to a *massive perceptual illusion*, and I claim we *do* have the ability and disposition to check perceptual judgments. Therefore, we intuit that it is conceivable that we could be wrong to judge that we are not brains in vats, although we deem the possibility that we are brains in vats to be extremely unlikely on inductive grounds. See §8 for my discussion of the similar 'Matrix' illusion in the context of my discussion of Kammerer's illusion meta-problem.

[12] See Fodor (1983) for discussion of modularity, cognitive impenetrability, and informational encapsulation.

its plausibility is supported by the growing body of evidence for the modular nature of the mind.

I think evolutionary considerations explain why our brains lack a system to check ICJs. The brain process forming them is likely reliable, so there would be slight adaptive advantage in such a system. It would have been costly and might have slowed cognitive processing. In any case, a system to check ICJs would itself be subject to error. We would require a second-order system to check it, and a third-order system to check that second-order system, ad infinitum. The possibility of error cannot be eliminated. The most efficiently designed mind would treat its judgments about the states it finds itself in as primitive or basic data playing 'a foundational role for the system'. (Chalmers 1996: 291).

The above considerations explain why we evolved the cognitively impenetrable intuition that we have infallible knowledge about one aspect of reality—the classification of our sensory states, and highly reliable but fallible beliefs about another—the nature of the external world. Again, my hypothesis to explain the infallibility intuition is speculative. But whatever theory ultimately explains it,[13] the viability of the inference theory does not depend on the specific nature of the explanation. The inference theory requires only *that* we have the intuition, and it is very plausible that we do.

4 The Justification Constraint

The justification constraint is an epistemic constraint on judgment/belief. If a subject judges that *p*, she necessarily believes, either explicitly or tacitly, that she is justified in judging that *p* at the time she judges that *p*. Thus, the absurdity/incoherence of: *p, but I have no reason to believe that p.*

The Justification Constraint

A subject judging that *p* necessarily believes, either explicitly or tacitly, that she is justified in judging that *p* at the time she judges that *p*.[14,15]

Consider John Gibbons' discussion of a non-standard case of Moore's paradox (2005):

> 'If you think about Moore-paradoxical statements from the normative perspective, then the same kind of incoherence that is involved in the standard cases [of Moore's paradox] also seems to infect the following: *p, but I have no reason to believe that p…*'

[13] See, e.g., Clark et al. (2019) and Schwarz (2019), explaining the intuitive infallibility of phenomenal judgments in terms of Bayesian belief updating mechanisms.

[14] That is: for any agent *a*, any judgment *that p*, and any time *t*: *a* judges *that p* at $t \rightarrow a$ tacitly or explicitly believes at *t* that she is justified in judging *that p*.

[15] Similarly, Chalmers (2018: 24) suggests that a cognitive system will naturally hold it has justification for its judgments about its internal states: 'If the system is a reasoner that requires reasons for its claims, it will be natural for it to hold that it has basic or immediate evidence that it is in this state'.

I want to claim that the incoherence of the non-standard Moore-paradoxical statement is explained by its violation of the justification constraint. No rational agent would utter the non-standard statement precisely because it is a conceptual truth that any agent who believes that *p* necessarily believes that she has a reason to believe that *p*.[16]

Three clarifications about the justification constraint are in order. First, I do not claim a subject can *identify* or *articulate* a justification, or that any justification must actually exist, for a judgment she makes or a belief she holds. Instead, I claim more modestly that in judging that *p*, the subject believes there exists *some justification or other* justifying her judgment. Second, the subject's belief about justification will often be *tacit* (or *implicit*). Tacit beliefs are *non-occurrent attitudes*, or dispositional states. According to Schwitzgebel (2019: § 2.2), 'one believes P implicitly (or tacitly) if one believes P, but the mind does not possess, in a belief-like way, a representation with that content. … Implicit beliefs are, perhaps, necessarily dispositional [i.e., non-occurrent]… if occurrently deploying a belief requires explicitly tokening a representation of it'. According to the inference theory, when a subject's belief that her judgment that *p* is justified is *tacit*, she does not, upon her judging that *p*, token any occurrent representation with the content *my judgment that p is justified*, nor any occurrent representation of her justification for judging that *p*. Third, it is sufficient for the inference theory that the justification constraint holds as a *contingent matter*, although it seems to me it is a conceptual truth.

There is potentially a regress objection to the justification constraint. Per the justification constraint, for any judgment *j*, the subject believes that *j* is justified. The objector might suppose that the subject would then believe that her belief that *j* is justified is justified; and then believe that her belief that her belief that *j* is justified is justified, is justified, ad infinitum. The objector's worry is that we are faced with an infinite series of tokenings of representations of ever higher-order attitudes following from the subject's (initial) belief that *j* is justified. However, as mentioned above, the justification constraint provides an escape from regress insofar as the subject's belief that her judgment *j* is justified will often be *tacit* (or *implicit*). Tacit beliefs are *non-occurrent attitudes*, or dispositional states (Schwitzgebel, ibid). Thus, a subject tacitly believing that *j* is justified does not token any occurrent representation with the content *j is justified*, nor any second-order occurrent representation with the content *my belief that j is justified is justified*, etc. She is merely disposed to do so under certain hypothetical circumstances. In no case will she token an actual infinity of occurrent representations, thus avoiding regress.

One might also object that the justification constraint is too strong, instead advocating for a different epistemic constraint along the lines Kieran Setiya (2008: 43) suggests in the context of discussing Bernard Williams' argument (1976) that it is impossible to acquire beliefs at will:

> 'It is impossible to believe that *p* or to be confident that *p* while believing that this degree of confidence or belief is not epistemically justified'.

[16] Moore paradoxical statements (of either the standard or non-standard sort Gibbons suggests) do not involve the speaker in inconsistency. They are paradoxical precisely because the speaker is not inconsistent, but at the same time says something incoherent.

This constraint is easier to satisfy than the justification constraint. It amounts to the claim that for every belief that *p*, the agent either believes her belief is justified (either explicitly or tacitly) *or* suspends judgment about justification. The only thing the agent cannot do is affirmatively believe that her belief that *p* is not justified while believing that *p*.

Despite this objection, I think that the justification constraint I propose is plausible, even if not quite as plausible as Setiya's constraint. A fulsome discussion of the justification constraint would require more space than I can devote to it here. I recognize that it remains a speculative hypothesis. Nevertheless, given its intuitive plausibility, it shall be my working assumption here that it is true.

5 Solving the Illusion Problem: The Inference Theory

I claim that phenomenal properties are *justificatory properties* that a subject erroneously unconsciously infers to be instantiated. This inference occurs upon introspection, i.e., upon the formation of ICJs. She intuits each type of justificatory property to be *like something* different from other types of justificatory properties, i.e., she intuits that each type of justificatory property has its own what-it's-likeness or 'WIL' distinct from the WIL of every other type of justificatory property—which WIL she intuits to be directly apprehensible, ineffable, atomic, intrinsic (non-relational), private, and non-physical.

I set out the inference theory as follows. In §5.1 below, I explain why, when a subject introspects and forms ICJs, she infers that justificatory properties (which I subsequently argue in §5.4 are identical to phenomenal properties) are instantiated. In §5.2, why she judges each type of justificatory property to be distinct from every other. In §5.3, why she judges each type of justificatory property to be *qualitatively like something*—to have a what-it's-likeness or WIL distinct from the WIL of every other type of justificatory property. In §5.4, why she judges she is *directly acquainted* (Chalmers 2010) with the WILs of justificatory properties on introspection. Finally, also in §5.4, I argue that we should identify these justificatory properties with phenomenal properties—for judging that an aspect of the mind (or in the mind) is *like something* with which we are directly acquainted upon introspection (i.e., *known non-inferentially*) is the hallmark of the phenomenal.

5.1 Judging Justificatory Properties to Be Instantiated

Through rational reconstructions of a subject's inferential process with respect to classificatory introspective judgment tokens, viz. ICJ-Red ('Introspective Classificatory Judgment Red') and ICJ-Green ('Introspective Classificatory Judgment Green'), I illustrate why she judges, upon introspection, that justificatory properties are instantiated.

ICJ-Red *I am experiencing a subjectively red feel*
ICJ-Green *I am experiencing a subjectively green feel*

Two clarifications are in order. First, ICJ-Red and ICJ-Green are ICJ *tokens*, although they represent ICJ tokens *of* different types. Second, the rational reconstructions below

model, from the first-person perspective, the erroneous inference patterns occurring upon a subject's formation of ICJs. However, this inference pattern is *unconscious*, so that the subject is not consciously aware that she is following it, inferring (iii), unaware of (and without the capacity to introspect) how she has inferred (iii) from (i) and (ii).

ICJ-Red: Rational Reconstruction

(i) I infallibly know that ICJ-Red is true.
(ii) ICJ-Red is true → There is a property instantiated—'p-redness'[17]—justifying judgment ICJ-Red.
(iii) *Therefore*, p-redness is instantiated.

ICJ-Green: Rational Reconstruction

(i) I infallibly know that ICJ-Green is true.
(ii) ICJ-Green is true → There is a property instantiated—'p-greenness'—justifying judgment ICJ-Green.
(iii) *Therefore*, p-greenness is instantiated.

On the supposition that the infallibility intuition (according to which a subject judges her ICJs to be infallible) characterizes human psychology, as I claimed in §3 *supra*, we should expect a subject to hold premise (i). And on the supposition that the justification constraint (according to which a subject believes her ICJs to be justified) characterizes human psychology, as I claimed in §4 *supra*, we should expect a subject to hold premise (ii). If the subject has judged her sensory state to be red*—e.g., reporting that she 'experiences a subjectively red feel'—she necessarily believes (at least tacitly) that there is *something about the world* in virtue of which she is justified in so judging, which is to say that she believes *there is some property instantiated* in virtue of which she is justified in so judging. (N.B.: She cannot believe this justificatory property to be the neural properties of her sensory state whose signatures her brain process has read to form her ICJ, for introspective opacity blocks her access to these neural properties.) I called these justificatory properties 'p-redness' in the first rational reconstruction and 'p-greenness' in the second. Accordingly, whenever a subject holding (i) and (ii) forms an ICJ classifying her sensory state, she infers (iii). She would have a very high level of confidence in the truth of (iii). After all, if she is subject to the infallibility intuition, the subject would take herself to know infallibly that (i) is true. And if the justification constraint is a conceptual truth or a basic principle of human psychology, she would have a very high degree of confidence in the truth of (ii).[18]

[17] Here, 'p-redness' refers to a justificatory property. I argue *infra* in §5.4 that this justificatory property is identical to a phenomenal property. For exegetical simplicity, I use 'p-redness' for p-redness both *qua* justificatory and phenomenal property.

[18] Without the infallibility intuition, I speculate that the illusion would be much weaker. An introspecting subject would have less confidence in the truth of her ICJ, so she would have less confidence in the instantiation of a justificatory property putatively justifying that ICJ. Hence, she would be more readily disposed to doubt the instantiation of phenomenal properties (which I claim in §5.4 *infra* are identical to phenomenal properties) and more disposed to accept illusionism on an intuitive level.

The subject does not *conceive* p-redness or p-greenness as justificatory properties, for she is not conscious of the justification constraint playing a role in the generation of her judgment in (iii).[19] Her inference is *unconscious* so that she cannot introspectively access the fact *that* she arrived at (iii) via inference, nor *how* she has drawn the inference. (Cf. von Helmholtz's (1867) appeal to unconscious inference to explain the Lyer-Müller illusion, where subjects are supposed to lack introspective access to the unconscious inferential process generating that illusion).

Some subjects would judge these justificatory properties not to be located *in the mind*, e.g., intuiting that p-redness is a property instantiated in the *ripe tomato* viewed rather than a mental property. Nevertheless, upon sufficient *philosophical reflection,* we tend to judge p-redness to be a property located in the mind. First, the inverted spectrum pumps our intuition that phenomenal properties are instantiated in the mind. Specifically, the intuition pumped is that two subjects may call a colour by the same name but differ in their subjective experience of it. However, their verbal reports about features of external, publicly observable, non-mental reality cannot communicate the nature of their subjective experiences, such that they might determine whether there is a spectrum inversion between them. Therefore, the essential natures of their subjective experiences—what those experiences are like, their WILs—are intuitively not constituted by features/properties of external, publicly observable, non-mental reality. Second, subjective feels demonstrate stark independence from external-world, non-mental reality. Considering this fact generates the intuition that these subjective feels are in the subject rather than in the external world. For example, it is apparent that even when no red object is present, we may experience (what we deem to be) a subjectively red feel in an afterimage, in our imaginations, or in dreams. And subjectively red feels can be caused by objects that are not themselves red-coloured. For example, a subject may look at a tomato and judge that she experiences a subjectively red feel but subsequently learn upon closer inspection that the tomato that caused the subjectively red feel was orange-coloured.

With other sensory modalities, such as *pain*, subjects would typically judge the phenomenal property of painfulness (*p-painfulness*) to be located in the subject, for there are no external objects onto which one might plausibly project p-painfulness. Pain generally persists after the event causing it (e.g., a stone falling on one's foot) and often has no apparent external cause. Therefore, p-painfulness regularly presents itself to us as independent of external reality. Of course, some subjects might conceive p-painfulness as a property located in a *body part* rather than in the mind. However, with both colour perception and pain sensation, subjects will tend, upon philosophical reflection, to conceive phenomenal properties as located in the subject (whether in the mind or the body).

[19] In other words, she does not take these properties under the mode of presentation *justificatory property*. At the same time, agents' disposition to refer to phenomenal properties to justify their judgments about the types of sensory states they are experiencing constitutes some evidence in favour of my contention that they are justificatory properties. For example, if asked how she knows she is in pain, an agent may say, e.g., 'I know I am in pain because *it feels* painful to me'. Or she may say 'I know I am seeing red because things *look* red to me'.

In any case, the inference theory is neutral on the question about where subjects judge these justificatory properties to be located, consistent either with subjects judging them to be instantiated inside or outside the mind.

5.2 Judging each Type of Justificatory Property to Be Distinct

A subject not only infers that p-redness and p-greenness are instantiated (as argued in §5.1, *supra*) but also necessarily judges p-redness and p-greenness to be *distinct* types of properties, as shown by *reductio*. Suppose she judged p-redness and p-greenness to be the *same* property: *p-gredness*. She would, therefore, judge that the instantiation of p-gredness could constitute the justificatory basis *either* for forming an ICJ classifying her experience as subjectively red *or* as subjectively green. Now, suppose she formed an ICJ classifying her experience as subjectively red. How could she judge that this ICJ was justified if she judged that the evidence for it—the instantiation of p-gredness—was equivocal, equally well supporting an ICJ that her experience was subjectively green? She would have to judge her ICJ that her experience was subjectively red to be *arbitrary*, for she could just as easily have judged her experience to be subjectively green. However, the *justification constraint* bars her from judging any ICJ to be arbitrary, for it would lack justification. Therefore, if she has judged her experience to be subjectively red, she cannot *but* judge that there is a property instantiated, p-redness, that justifies this judgment to the exclusion of judging her experience to be subjectively green.

The *infallibility intuition* likewise explains why a subject judges p-redness and p-greenness to be distinct properties. If, in the scenario described above, the agent judged her ICJ that her experience is subjectively red to be arbitrary and unjustified, she could not judge that she knew it *infallibly*. Given that she *does* in fact judge that she knows infallibly that her ICJ is true (per the infallibility intuition), she necessarily excludes the possibility that an equivocal justificatory property, such as p-gredness, could be the justificatory property backing her judgment. She must judge that her ICJ that her experience is subjectively red is supported by a justificatory property, p-redness, whose instantiation *guarantees* its truth—such that she knows *infallibly* that it is true—and categorically excludes her from judging her experience to be subjectively green.

5.3 Judging each Type of Justificatory Property to Be *Like Something*

In §5.1 and §5.2 above, I have explained thus far:

- why a subject judges that there are justificatory properties, such as p-redness and p-greenness, instantiated; and
- why she judges each type of justificatory property to be distinct from every other; and, therefore, why she judges that p-redness is distinct from p-greenness.

How do we now explain why she judges each type of justificatory property to be qualitatively *like something* distinct from what every other type of justificatory property is like? Let us define the what-it's-likeness of an entity, or its 'WIL', as follows:

What-it's-likeness or 'WIL' of an entity

The essential properties borne by an entity

From Leibniz's law (Leibniz 1969: 308), it follows that for any two entities X and Y, X = Y iff the WIL of X = the WIL of Y; here, X and Y differ in no essential property. And X ≠ Y (X and Y are 'distinct') iff the WIL of X ≠ the WIL of Y; here, X and Y differ in at least one essential property.[20]

Identity

X = Y iff the WIL of X = the WIL of Y (i.e., X and Y have exactly the same essential properties).

Distinctness (i.e., non-identity)

X ≠ Y (X and Y are non-identical or 'distinct') iff the WIL of X ≠ the WIL of Y (i.e.,X and Y differ in at least one essential property)

Because Distinctness is an obvious truth, a subject's inferential process can be reasonably expected to conform to it. This is captured in the principle I call 'distinctness as a rule of inference in judgment' or 'Distinctness-RIJ':

Distinctness as a Rule of Inference in Judgment ('Distinctness-RIJ')

A subject judges that entities X and Y are non-identical or 'distinct' iff she judges that X and Y have distinct WILs, i.e., X and Y differ in at least one essential property.

With the above Distinctness-RIJ principle, we can explain why a subject judges that p-redness and p-greenness have distinct WILs. As argued in §5.2, a subject necessarily judges each type of justificatory property to be distinct from every other. Accordingly, she judges that p-redness ≠ p-greenness, which, in turn, instantiates the antecedent of Distinctness-RIJ. Since she judges the antecedent to be true, she judges the consequent to be true as well—that the WIL of p-redness ≠ the WIL of p-greenness.

5.4 Judging *Direct Acquaintance* with Justificatory Properties, which I Claim Are Identical to Phenomenal Properties

In §§5.1-5.3, *supra,* I explained:

- why a subject judges that there are justificatory properties, such as p-redness and p-greenness, instantiated; and

[20] Leibniz's law encompasses both *essential* and *accidental* properties. I deviate from Leibniz's formulation and include only essential properties, for one ought not to differentiate entities *by type* based on accidental properties.

- why she judges each justificatory property type to be distinct from every other; and therefore, why she judges that p-redness is distinct from p-greenness; and,
- why she judges each justificatory property type to have its own WIL distinct from that of every other justificatory property type; and therefore, she judges that the WIL of p-redness ≠ the WIL of p-greenness.

I now must explain why a subject judges that she is *directly acquainted* with the WILs of the justificatory properties justifying her ICJs. She does not merely judge *that* the WIL of p-redness ≠ the WIL of p-greenness. Over and above this, she judges that she *knows these WILs intimately*, as if she were in immediate contact with the WILs of p-redness and p-greenness so that their natures are *revealed* to her as they truly are. First, I explain why the subject judges that she is directly acquainted with these WILs. Then I argue that being (putatively) directly acquainted with WILs upon introspection is the hallmark of the phenomenal; and therefore, we ought to identify these justificatory properties with phenomenal properties.

It is important to mention at the outset that we intuit phenomenal properties to be *atomic*—to consist not of a bundle of properties, but to be non-composite, unanalysable, 'simple and homogenous' (Dennett 1988), with no parts. Now I claim that phenomenal properties do not exist. Thus, the question of whether they are *in fact* atomic, or merely intuited as such, cannot be sensibly raised within the framework of the inference theory. (See § 7, *infra*, for discussion of the problem intuition of atomicity).

Let us now consider a principle, which I take to be a conceptual truth, that applies exclusively to atomic entities: 'Apprehension of Distinctness of Atomic Entities', or 'ADAE'.

Apprehension of Distinctness of Atomic Entities ('ADAE')

A subject judges that the WIL of (atomic) X ≠ the WIL of (atomic) Y *iff* she judges that she *apprehends* the WILs of X and Y *in their entireties*, compares them, and discerns they differ.

Importantly, ADAE is *false* with respect to *non*-atomic entities. To apprehend the WIL of an entity *in its entirety* is to apprehend *all* of its essential properties. With *non*-atomic entities, one need *not* apprehend the WILs of two entities, P and Q, in their entireties to judge them to be distinct. It is sufficient to judge that they differ in just *one* of their potentially numerous essential properties. By contrast, the WILs of *atomic* entities intuitively consist in *one single unanalysable property*. Therefore, a subject's judgment that she has apprehended the WIL of p-redness and the WIL of p-greenness, compared them, and discerned *one* difference amounts to having apprehended their WILs *in their entireties* (the single atomic property that each of these WILs intuitively consists in) and discerned them to be different.

We can see that ADAE must be true by *reductio*. Suppose a subject judged that she had apprehended the WIL of neither X nor Y, nor discerned any difference between them. She would have to judge her judgment about the distinctness of X and Y to be arbitrary and unjustified, and therefore no judgment at all. Therefore, a subject judging that the WIL of X ≠ the WIL of Y necessarily judges that she apprehends their WILs

individually (what X and Y individually 'are like'), compares them, and discerns a difference between their WILs. N.B.: ADAE does not provide that she *in fact* discerns the difference between the WILs of X and Y, only that she *judges* that she does (for these WILs are, I claim, non-existent).

ADAE explains why the subject judges that she apprehends the WILs of p-redness and p-greenness. I argued in §5.3 that a subject judges that the WIL of p-redness ≠ the WIL of p-greenness. This judgment, in turn, *instantiates the antecedent of ADAE*. Therefore, the consequent of ADAE is true as well: she judges that she *apprehends* the WILs of p-redness and p-greenness individually, compares them, and discerns that they are different. This explains the intuited apprehensibility of the WILs of p-redness and p-greenness.

Importantly, this apprehension of WILs is intuitively *direct*, i.e., *intuitively non-inferential*, because the inferential process behind her judgment that she has apprehended these WILs is *unconscious*. She is aware only of the *product* of this inferential process, her judgment that she has apprehended these WILs.

Given the foregoing discussion, I contend that we should identify these justificatory properties with phenomenal properties. We intuit not only that they are instantiated, but further, that we directly apprehend, or are directly acquainted with, what each property *is like*, its WIL, through introspection. We judge our knowledge of this WIL to be immediate and non-inferential because the inference is unconscious, and we lack introspective access to it. This description captures precisely what a phenomenal property is supposed to be: a property such that we are, through introspection, directly acquainted with what it is like.

6 Avoiding the 'Phenomenal Representation Problem'

The 'Phenomenal Representation Problem' (Frankish 2018: slide 16), or 'PRP', presents a potential worry for illusionist theories of phenomenal consciousness. Derk Pereboom articulates the PRP in his 2016 (185):

> … Illusionists agree that the what-it's-like features of sensory states, what I call the qualitative natures of phenomenal properties, are illusory in that they don't exist. But what of the illusion itself? **Experiencing an illusion requires that the illusions themselves, which are kinds of introspective representations, will exist. However, these illusions themselves will have phenomenal properties, or so it would seem** [Boldface mine]. Illusions of phenomenal properties would appear to differ in a key respect from other sorts of illusions, for instance from the illusions of psychokinesis … Illusions of psychokinesis will not have psychokinetic properties, while illusions of phenomenal properties would seem to have phenomenal properties. Accordingly, the objector may claim that there will be something it is like to have an illusion of phenomenal greenness, and it is the same as what it is like to have a sensation of green. If this is right, then the illusionist's strategy won't get us very far.

Keith Frankish articulates the PRP in his 2019:

> [One common objection to illusionism is that] the very idea of illusionism [is] confused. To be under the illusion of seeing an apple is to have an experience exactly like that of seeing an apple, even though there's no apple present. How

> then could we be under the illusion of having an experience? If you are having an experience exactly like a pain experience, then you are having a pain experience…. This looks like a serious objection, but in fact it is easily dealt with… **When illusionists say that phenomenal properties are illusory, they mean that we have introspective representations like those that we would have if our experiences had phenomenal properties. And we can have such representations even if our experiences don't have phenomenal properties. Of course, this assumes that the representations themselves don't have phenomenal properties. But … representations needn't possess the properties they represent** [Boldface mine]. Representations of redness needn't be red, and representations of phenomenal properties needn't be phenomenal.
> **But how does a brain state represent a phenomenal property? This is a tough question** [boldface mine].

According to Pereboom and Frankish, the illusion of a phenomenal property necessarily involves an introspective representation of it. An introspective representation of pain would capture, from the first-person point of view, the subjective feel of pain. And an introspective representation of red would capture the subjectively felt redness of the red. The problem is that, intuitively, if our introspective representations managed to capture these subjective feels, it would be for an experiencing subject tokening these representations exactly as if she were experiencing the *genuine* subjective feels of pain or red, i.e., the phenomenal properties p-painfulness and p-redness. To introspectively represent what it is like to feel pain would *be* to experience actual pain. And to introspectively represent what it is like to see red would be to experience the actual way red subjectively looks 'from the inside', as it were. Our introspective representations of phenomenal properties would themselves turn out to have phenomenal properties. And this would be problematic for illusionism because the illusionist would not have *eliminated* phenomenal properties but merely *relocated* them from our sensory states to our introspective representations of them.

Of course, intuitions can be unreliable. It is too quick, Frankish argues, to credit the intuition that introspective representations of phenomenal properties would themselves have phenomenal properties. Representations need not possess the properties they represent. Still, Frankish concedes on the last line of the above quotation that the question of how a brain state could represent a phenomenal property is quite difficult. How can a purely subjective, ineffable feel be represented if no such feels are instantiated in the world? As Frankish recognizes (2016: 18–19), representations of phenomenal properties cannot get their content from causal connections to phenomenal properties, for non-existent entities cannot enter into causal relations.

The simplest way to avoid the PRP—the approach I advocate—involves leaning on the distinction between our *conceptual* representations of phenomenal properties on the one hand and *introspective* (or *quasi-perceptual*) representations of them on the other. We *do* form conceptual representations of phenomenal properties—which constitute what we *judge* to be true about them as a result of the erroneous inferential process I proposed above in §5. For example, we judge that they are instantiated, that they have distinctive WILs with which we are directly acquainted, that they are typically caused by visual experiences of certain kinds of objects, etc. However, contrary to Pereboom's

claim that illusions are 'kinds of introspective representations', or Frankish's claim that the illusion of phenomenal consciousness involves 'introspective representations like those that we would have if our experiences had phenomenal properties', I maintain that this illusion does not involve, nor is it caused by, introspective representations. Phenomenal properties are not like anything, nor do we introspectively represent them as being like anything. When we look at a red surface and contemplate the subjectively red feel in us, we are contemplating nothing whatsoever. We are firmly convinced that we are contemplating something, but we are mistaken. Frankish (2018: slide 16) notes that this move, which is typical of what he calls 'negative misrepresentation theories', cleanly avoids the PRP, for such theories deny the existence of any introspective or quasi-perceptual representations whose content would need to be explained. (By contrast, '*positive* representation theories' (ibid., slides 17 & 18) posit that introspective representations of phenomenal properties cause the illusion, so such theories owe us an account of the content of said representations). The inference theory avoids the PRP by gainsaying Pereboom's claim that '[e]xperiencing an illusion requires that the illusions themselves, which are kinds of introspective representations, will exist'. The inferential process described in §5 generates the illusion, and we need not posit any introspective representations of phenomenal properties to explain it.[21]

How, one might wonder, can a subject distinguish p-redness from p-greenness if neither property exits, nor do we introspectively represent to ourselves what each is like? I shall make some proposals that may go some way towards an answer. As discussed above, we conceptually represent certain relational aspects of phenomenal properties. A subject associates the non-existent phenomenal properties p-redness and p-greenness with the distinct objects typically causing their (putative) instantiations, e.g., ripe tomatoes and strawberries with the former, grass and unripe bananas with the latter. She might distinguish these non-existent phenomenal properties based on their distinct associations. A subject also likely associates p-redness and p-greenness with different *affordances*—the distinct ways she might interact with or respond to the environment when in a red* versus a green* sensory state. A subject also associates p-redness and p-greenness with her red* and green* sensory states, respectively. She *can* and *does* distinguish these sensory states, for she is access conscious vis-à-vis the distinctions that her introspectively opaque brain process draws between them. Furthermore, the subject may associate each phenomenal property with the distinct ICJs she forms upon introspecting her sensory states and the distinct ICRs she is disposed to utter. Thus, she might conceive p-redness as the phenomenal property she infers to exist whenever she is in a red* state and forms the ICJ *I am experiencing a subjectively red feel*. And she might conceive p-greenness as the phenomenal property she infers to exist whenever she is in a green* state and forms the ICJ *I am*

[21] Three promising contemporary illusionist theories, those of Kammerer, Pereboom, and Humphrey, aim to explain our erroneous beliefs about phenomenal consciousness in terms of our representations (or misrepresentations) of our mental states. Francois Kammerer's ERM theory (Kammerer, 2019b) aims to explain problematic intuitions 'by putting our representations of phenomenal states in perspective within the larger frame of the cognitive processes we use to conceive of evidence'. Like the inference theory, Kammerer's theory focuses on justification/evidence to explain the illusion. Derk Pereboom's Phenomenal Inaccuracy Theory (2019) claims the mind misrepresents our sensory states as having phenomenal properties that they do not have. He addresses the PRP in his 2016 and 2019, although his approach to the problem is different from mine. Nicholas Humphrey argues that evolutionary considerations explain why we represent our inner states as having phenomenal properties (although Humphrey now classifies his view as 'phenomenal surrealism' (2016, 2019) rather than as a species of illusionism, as he had in previous writings.)

experiencing a subjectively green feel. Finally, I speculate that when we contemplate a phenomenal property such as p-redness in the imagination (when not contemporaneously viewing a red object), we token a 'faint copy' of a red* sensory state and an accompanying faint copy of the ICJ *this is a subjectively red feel*. Perhaps we conceptualize p-redness *indexically* as the property *this* ICJ is about—the one we are faintly tokening contemporaneously with our thought about p-redness; and likewise for p-greenness. So our thoughts about p-redness and p-greenness would be distinguishable based on being indexically linked to faint copies of distinct sensory states, one red* and the other green*.[22]

7 Explaining 'Problem Intuitions' about Phenomenal Properties

The inference theory explains our 'problem intuitions' about consciousness (Chalmers 2018). These problem intuitions include the intuition that our sensory states bear phenomenal properties that are *like something* with which we are directly acquainted (discussed in §5) and that what they are like is *ineffable*, *atomic, intrinsic (non-relational)*, *private*, and *non-physical*.

Except for the problem intuition of intrinsicality/non-relationality (discussed at the end of this section), I shall explain the other four problem intuitions—ineffability, atomicity, privacy, and non-physicality—in terms of the fact that we do not represent (in a sense to be qualified) the WILs of phenomenal properties. In §6, I denied that we introspectively represent WILs. As discussed in §5 and §6, we conceptually represent WILs in the *limited* sense that we judge that they exist, judge that each type of phenomenal property has a distinct WIL, and judge that we are directly acquainted with these WILs.[23] However, I claim that we do not represent WILs *robustly*—in a way that would allow us to differentiate the WIL of one phenomenal property type from the

[22] Papineau (2008: 105), invoking Hume, articulates a similar idea, which he employs with respect to his phenomenal concepts strategy approach (which is not an illusionist theory): '… a similar point applies to imaginative uses of phenomenal concepts. When I later think imaginatively about some earlier experience, like seeing red ('that experience . . .'), I won't actually have an experience of seeing red, but my experience is likely to bear some phenomenal similarity to the experience of seeing red, to be 'a faint copy', as Hume put it. So in both cases the use of phenomenal concepts to refer to some experience will standardly involve the thinker actually having the experience itself, or a faint copy of it'.

[23] I categorize our conceptual representations about phenomenal properties roughly into three types. The first type includes the judgments *that* phenomenal properties such as p-redness and p-greenness are instantiated, *that* they have different WILs, and *that* we are directly acquainted with these WILs. These judgments are false, as discussed in §5. No WILs exist, nor are we are directly acquainted with any WILs; we merely judge that we are. The second type is about relational properties holding between phenomenal properties and the external world. For example, we represent that ripe tomatoes and strawberries cause the instantiation of p-redness, and that unripe bananas and grass cause the instantiation of p-greenness. The third type is about relational properties holding *between* phenomenal properties. For example, we judge p-redness and p-orangeness to be more similar to one another than to p-greenness. This conceptual representation is based on a judgment formed by an introspectively opaque brain process about the greater similarity of a red* sensory state to an orange* sensory state than to a green* one. We are access conscious with respect to this judgment of similarity/dissimilarity even though the brain process forming it is introspectively opaque. One need not suppose that phenomenal properties exist, nor that we introspectively represent what they are like, in order to explain how we form these judgments about them. These judgments are in a sense really judgments about the relational properties of our *sensory states*, and not about phenomenal properties, for our sensory states are their *truthmakers*. We are disposed to mistakenly judge these representations to be about phenomenal properties, rather than our sensory states, because introspective opacity blinds us to the very existence of our sensory states.

distinct WIL of another and articulate the difference between them. If we robustly represented the WILs of p-redness and p-greenness, it would be possible to articulate (at least in principle) *how* we differently represented them such that we could determine whether two agents, *a* and *b*, see colours the same way or are inverted. But according to the intuition pumped in the inverted spectrum, we cannot do this.[24] Similarly, if the WILs of p-redness and p-greenness were robustly represented, we would be able to explain to a man blind from birth what it is like to see red and distinguish this from what it is like to see green. But this is impossible. In short, I claim that we represent the WILs of phenomenal properties neither introspectively nor robustly. And furthermore, I argue that this lack of representation generates the problem intuitions that WILs are ineffable, atomic, private, and non-physical. In what follows, for ease of exposition, I shall say that WILs 'are not represented', with the understanding that they are conceptually represented in the limited sense mentioned above but not represented introspectively or robustly.

Let us begin with the problem intuition of *ineffability.* Ineffability is the thesis that our power to represent the world outstrips our capacity to communicate, through language, how we represent it. With respect to phenomenal consciousness, ineffability is exemplified by the claim that we represent the WIL of p-redness but are unable, even *in principle*, to communicate *how* we represent it using language. I consider ineffability a *problem* intuition because I take the thesis of ineffability to be false: any entity *e* that is represented is such that it is in principle possible to communicate how *e* is represented, even if *e* is non-existent. According to the inference theory, we cannot articulate how we represent the WILs of phenomenal properties not because they are ineffable, but because we do not represent them. If the WILs of p-redness and p-greenness were non-existent but at least represented, it would be possible to articulate *how* they were differently represented, much in the way that unicorns and leprechauns are non-existent but the differences between them are represented and are therefore effable. The intuition of ineffability arises out of the clash of two powerful forces—our (erroneous) conviction that we vividly and clearly represent the WILs of phenomenal properties on the one hand (as explained in §5), and our recognition, on the other, that words fail us completely when we attempt to articulate how we represent them.

[24] Conceptual representations about the different kinds of objects typically causing instantiations of p-redness and p-greenness shed no light on their WILs and cannot be appealed to in order distinguish the WIL of p-redness from the WIL of p-greenness. Such representations are intuitively about contingent or accidental matters. After all, if you woke up one day and found your spectrum inverted with respect to the previous day, red would still subjectively 'look' precisely the same way to you—it would have the very same WIL; it's just that different kinds of objects would now look red to you and cause subjectively red feels in you. So representations about which kinds of objects cause subjectively red feels do not get at the essential nature of what it's like to see red, its WIL. Its essential nature, its WIL, is whatever is *invariable* across spectrum inversions. (This does not entail that we cannot associate phenomenal properties with the objects we take to cause their instantiations, since it is commonplace to associate entities with their contingent properties. However, such associations cannot inform us about their essential properties.) To be clear, as an illusionist I do not hold that WILs exist, nor that spectrum inversions are genuinely conceivable (for if there are no WILs, there is nothing that might be inverted). But the inverted spectrum is nevertheless a good intuition pump through which we can explore our intuitions about phenomenal consciousness. See my explanation of the problem intuition of non-relationality, at the end of this section, for a more fulsome discussion of how the inverted spectrum exposes our representations of the relational properties of phenomenal properties to be intuitively contingent or accidental, and the WILs of phenomenal properties as independent of such relational properties.

Next, we have the problem intuition of *atomicity*—the intuition that the WIL of a phenomenal property such as p-redness is atomic: simple and homogenous, non-composite, unanalysable, and lacking parts, components, or internal structure. According to the inference theory, we intuit that the WIL of p-redness is atomic because it lacks components (because it is non-existent) and is not represented as having components (because it is not represented). If, hypothetically, the components of the WIL of p-redness were represented, it should be possible to say *how*, regardless of whether or not p-redness existed—much as we can articulate how we represent unicorns in terms of the components of the concept *unicorn* (horse-like, horn-bearing, etc.) despite the non-existence of unicorns.

Next, we have the intuition of *privacy*. A subject intuits the WILs of phenomenal properties to be *private*—i.e., inter-subjectively incomparable, consistent with folk intuitions in the inverted spectrum—because the WILs of phenomenal properties are not represented. I cannot compare what p-redness and p-greenness are like for me to what they are like for you because there is no represented content on which to base any such comparison.

Next, we have the *anti-physicalist intuition*, which I claim is widely misconceived. According to the conventional wisdom, we positively and affirmatively represent phenomenal properties as non-physical. However, the conventional wisdom is wrong. We represent them as non-physical only in the *weak sense* that we fail to represent them as physical.[25] In fact, we fail to represent them either as physical *or* as non-physical. To see why the conventional wisdom is wrong, consider a toy *non-physicalist* theory of phenomenal consciousness, T^{NP}. According to T^{NP}, the following identities hold: non-physical property R = p-redness and non-physical property G = p-greenness. As shown in the *reductio* arguments below, when we subject T^{NP} to the inverted spectrum, these identifications are undermined. And when we subject T^{NP} to the zombie argument, T^{NP} is undermined as a genuinely explanatory theory of phenomenal consciousness.

Inverted spectrum applied to T^{NP}

Per T^{NP}, non-physical property R = p-redness and non-physical property G = p-greenness. Nevertheless, it is perfectly conceivable that this could be inverted for subject *s*, so that for *s*, non-physical property G = p-redness, and non-physical property R = p-greenness. *So non-physical property R ≠ p-redness and non-physical property G ≠ p-greenness after all.*

Zombie argument applied to T^{NP}

Suppose that T^{NP} provides that subject *s* satisfies all non-physical necessary and sufficient conditions to be phenomenally conscious. Nevertheless, it is conceivable that *s* is a zombie. *So T^{NP} does not explain phenomenal consciousness after all.*

If the above *reductio* arguments are correct, non-physicalist theory T^{NP} confronts an explanatory gap equally as unbridgeable as any physicalist theory. We would have just

[25] Chalmers (2018: 23) points out that the failure of introspection to represent phenomenal properties as physical does not adequately explain why we positively represent them as non-physical: '… there are any number of cases where one does not perceive that some phenomenon is physical, without perceiving that it is non-physical'. I avoid Chalmers' worry by arguing that we do *not* positively represent phenomenal properties as non-physical.

as much trouble grasping how a non-physicalist theory, such as T^{NP}, could explain phenomenal consciousness as any physicalist theory. The anti-physicalist intuition turns out to be merely a special case of what I shall call the *anti-explanatory intuition*, according to which we resist the identity of a phenomenal property with *any* property proposed to explain it, whether physical or non-physical.

The inference theory explains this anti-explanatory intuition in terms of our failure to represent the WILs of phenomenal properties. *A non-existent entity that is not represented, such as a WIL, cannot resemble anything.* Therefore, if some theory proposed property R, whether physical or non-physical, to explain the WIL of p-redness (or property R as identical to it), property R would never seem to us to resemble the WIL of p-redness any more than property G or any other property. Hence, we would persist in thinking it conceivable that property R might not be identical to the WIL of p-redness. Furthermore, if some *overall* theory of phenomenal consciousness proposed property P, whether physical or non-physical, to explain phenomenal consciousness *in general* (or property P as identical to phenomenal consciousness), property P would never seem to us to resemble the WIL of phenomenal consciousness—what it feels like from the inside to be phenomenally conscious. We would persist in thinking it conceivable that creatures instantiating property P might be zombies.

Non-physicalist theories aiming to explain phenomenal consciousness will fare no better than physicalist theories. The hard problem of consciousness is not the question of how phenomenal consciousness could arise from physical matter but how it could arise from *anything* whatsoever. As Thomas Nagel put it, 'The question of how one can include in the objective world a mental substance having subjective properties is as acute as the question how a physical substance can have subjective properties'. (1979: 201). What then explains the attraction of non-physicalist theories such as dualism, idealism, or panpsychism? At present, they are exercises in hand-waving that leave consciousness a comfortable mystery. They do not constitute *detailed* proposals positing *specific identities*. I suspect that any specific and detailed non-physicalist theory of phenomenal consciousness would inexorably come up against the same unbridgeable explanatory gap that physicalist theories face.

Finally, we have the problem intuition of intrinsicality or non-relationality, to which I shall refer henceforth as 'non-relationality'. Non-relationality is the intuition that what phenomenal properties are like, their WILs, are independent of their relational properties. An example of a relational property is the property of ripe tomatoes and strawberries causing the instantiation of p-redness, and unripe bananas and grass causing the instantiation of p-greenness. The inverted spectrum pumps the intuition of non-relationality—that these sorts of relational properties are not essential elements of phenomenal properties. On one version of the inverted spectrum, you wake up one day and find your spectrum inverted—with ripe tomatoes now causing subjectively green experiences and grass now causing subjectively red experiences. Before the inversion, you conceptually represented p-redness as being as caused by ripe tomatoes. After the inversion, you now conceptually represent p-redness as being as caused by grass. Despite the change in how you conceptually represent the relational properties of p-redness with respect to the environmental causes of its instantiation, you judge that what it is like to see red—the WIL of p-redness—remains unchanged from before the inversion event. Red still subjectively 'looks' precisely the same way; it is just that different kinds of objects now look red to you and cause subjectively red feels in you.

This reveals that we intuit that such conceptual representations about the relational properties of p-redness are not about the essential nature of p-redness, its WIL. Intuitively, they are about accidental or contingent features of phenomenal properties.

How do we explain what generates the problem intuition of non-relationality? I shall speculate here. In §2, I claimed that a subject forms ICJs classifying her sensory states through a brain process that reads their physical/neural signatures. Reading these physical/neural signatures does not involve the apprehension of any relational properties holding between the subject and the external world. The accuracy of an ICJ is not a function of the colour of the object that caused the instantiation of the sensory state it is about. The ICJ *this is a subjectively red feel* is accurate as long as it is about a red* state, even if an orange-coloured object caused the instantiation of that red* state.[26,27,28] So the brain process forming ICJs is not sensitive to the relational properties holding between a red* sensory state and the external object that caused its instantiation. On the other hand, it is plausible to suppose that there are relational properties holding between the neural signatures of sensory states and the brain process reading these signatures to form ICJs, and that these relational properties are crucially involved in the formation of ICJs. However, due to introspective opacity, the subject cannot introspect the neural signatures of her sensory states or the operations of her brain process. If there are any such relational properties involved in the formation of ICJs, the subject is blind to them. Accordingly, the subject is unaware of the brain process forming her ICJs depending on any relational properties whatsoever—either relational properties holding between her sensory states and coloured surfaces of objects in the external world (because relational properties of that sort are not involved in the brain process forming ICJs), or relational properties holding between the neural signatures of her sensory states and the brain process reading them (because introspective opacity blocks the subject's access to them). Therefore, it seems to the subject that no relational properties are involved in the formation of her ICJs.

8 The Illusion Meta-Problem

François Kammerer (2016, 2018, 2019a) poses the *illusion meta-problem*: why does illusionism strike us as an especially counterintuitive, inconceivable, and incoherent view? Other sorts of illusions do not strike us this way. Consider the illusion of the external world depicted in the film *The Matrix* and call it 'the Matrix Illusion'. The average person inevitably finds the notion that she is subject to the Matrix illusion (i.e., that she is in the Matrix) wildly implausible. She does not take the possibility seriously. At the same time, she judges the Matrix illusion to be *coherent* and *conceivable*. She can imagine what it would be like to be in the Matrix. By contrast, she judges the non-existence of phenomenal consciousness to be *incoherent* and *inconceivable*.

[26] Of course, the instantiation of a red* sensory state upon viewing an orange-coloured object will typically cause the subject to form an erroneous *perceptual judgment* about the colour of the object, judging it to be red-coloured rather than orange-coloured.

[27] This explains why we intuit ICJs to be infallible, as discussed in §3. We cannot check the accuracy of ICJs against the outside world, and introspective opacity prevents us from checking them introspectively.

[28] An inaccurate ICJ would be, for example, the ICJ *this is a subjectively orange feel* formed about a red* sensory state.

Contrasting intuitions about coherence/conceivability distinguish the illusion of phenomenal consciousness from the Matrix illusion. To solve the illusion meta-problem, we must explain these contrasting intuitions.

We typically intuit incoherence/inconceivability whenever a proposition strikes us as *contradictory and logically impossible* and not merely physically (nomologically) impossible or extremely improbable. I propose that the illusionists' claim that phenomenal properties do not exist strikes us as contradictory and logically impossible—even if we may not be pre-philosophically disposed to put it this way—because the proposition that we know that justificatory (i.e., phenomenal) properties are instantiated is unconsciously inferred from, i.e., *logically entailed by*, premises we (erroneously) take to be fundamental truths, as I argued in §5. By contrast, our judgment that we are not subject to the Matrix illusion is not logically entailed by our beliefs. We disbelieve that we live in the Matrix on *inductive grounds* because the enormous scale of the misperception that would be involved makes it extraordinarily improbable, with a non-zero probability, but one so close to zero that we treat it as if it were zero. This is the relevant contrast between these illusions.

Perceptual illusions, such as optical illusions, can also be distinguished from the illusion of phenomenal consciousness. We find it relatively easy to accept the former (but not the latter) to be illusions. For example, with the Lyer-Müller illusion, once we measure the lines, we quickly accept that they are of equal length (despite continuing to be subject to the cognitively impenetrable illusion that they differ in length) because perceptual judgments are not subject to the infallibility intuition and are intuitively revisable, as I argued in §3. By contrast, as I also argued in §3, our ICJs are subject to the infallibility intuition and are intuitively unrevisable. As shown in the rational reconstructions in §5.1, this results in us drawing the inference that phenomenal properties are instantiated, and we take ourselves to know this with a very high degree of certainty. So illusionism posits that what we take to be highly certain and practically unrevisable—our judgments about the existence of phenomenal properties—are systematically false. Therefore, it is far more difficult to accept that phenomenal consciousness could be an illusion than to accept that the Lyer-Müller illusion is an illusion.

Optical illusions differ from the illusion of phenomenal consciousness in another important respect. With the Lyer-Müller illusion, we are told that the introspective representation of the lines having different lengths is false and that a different introspective representation—a representation of them being the same length, is true. Crucially, we can introspectively represent *both* the true and the false states-of-affairs—picturing in our imaginations the lines being the same length and being different lengths. And we are not told that our introspective representation of the lines being different lengths does not exist, only that it is false. By contrast, with the illusion of phenomenal consciousness, the illusionist does not claim that our introspective representation that it is like *this* to see red (where 'this' putatively denotes what it is like to see red) is false. Instead, the illusionist tells us that this introspective representation *does not exist*, for 'this' denotes nothing whatsoever. And he does not furnish our imaginations with a *second* introspective representation, a true one, to take the place of the one he has purged as non-existent. We cannot pictorially or imagistically represent the true situation. Instead, to comprehend the true situation—the non-existence of the

(putative) introspective representation that it is like *this* to see red—we must employ *philosophical reasoning* to override what our intuitions tell us is the case. And trusting reason (especially philosophical reasoning) over intuition does not come naturally.[29]

Finally, phenomenal properties, though non-existent, are the *only* properties of our sensory states with which we take ourselves to be acquainted. Our sensory states are introspectively opaque, and we do not see them. In denying the existence of phenomenal properties, the illusionist denies the existence of the only features of our sensory states we take ourselves to apprehend and by which we take ourselves to recognize them *as* sensory states. Therefore, it seems as if the illusionist were denying that we had any sensory states at all, which would appear to deny the data rather than to explain them.

9 The Inference Theory as a *Rich-Illusion View*

Recently, Kammerer has argued that the fact that the illusion of phenomenal consciousness is *cognitively impenetrable*[30] entails that the illusion must be *rich* rather than *sparse* (2019c:1).

> 'By "rich illusion", I mean an illusion in which an object is *positively presented in an incorrect way* by a cognitively impenetrable representational process; by "sparse illusion" I mean an illusion in which the object is *not positively presented in an incorrect way* by a cognitively impenetrable representational process – but presented in a merely *partial* or *incomplete* way, which in turn leads us to *infer* incorrect beliefs about the object.'

For example, optical illusions such as the Lyer-Müller illusion are rich illusions. The parallel lines are positively presented to us in an incorrect way by a cognitively impenetrable process, so that no matter what we come to believe about the lines being equally long, the illusion persists. We continue seeing the lines as having different lengths. By contrast, most magic tricks, such as the headless woman illusion described by Armstrong, are examples of sparse illusions.

> 'To produce this illusion, a woman is placed on a suitably illuminated stage with a dark background and a black cloth is placed over her head. It looks to the spectator as if she has no head. The spectators cannot see the woman's head. But they gain the impression that they can see that the woman has not got a head. … Unsophisticated spectators might conclude that the woman did not in fact have a head. What the

[29] Furthermore, if the illusion of phenomenal consciousness is *sui generis*—fundamentally distinct from other sorts of illusions such as optical illusions—we would have no other illusions to serve as a model to guide our understanding, no analogous illusions that we understand and through which we might conceptualize this illusion.

[30] An illusion is cognitively impenetrable *iff* one remains subject to it even after recognizing that it is an illusion.

> example shows is that, in certain cases, it is very natural for human beings to pass from something that is true: 'I do not perceive that X is Y', to something that may be false: "I perceive that X is not Y"'. (Armstrong, 1968, p. 48)

The woman is not positively represented in an incorrect way by a cognitively impenetrable process. We do not see anything that is not there; we simply *fail* to see her head. We infer (erroneously) from the fact that we do not see her head that she has no head. This erroneous inferred belief, which constitutes the illusion, is cognitively penetrable, so that once we come to understand the trick and believe that the woman actually has a head, the illusion quickly loses its force.

According to Kammerer, the illusion of phenomenal consciousness must be a rich illusion because of its strength. It persists regardless of what an agent might come to believe about the reality (or unreality) of phenomenal consciousness. By contrast, a sparse illusion such as the headless woman illusion quickly loses its grip on us once we come to believe it is an illusion and understand how it is generated. Kammerer criticizes Dennett's and Graziano's theories for being sparse-illusion views (2019c: 6–8). On Dennett's 'user illusion' view (Dennett 2017: 271),[31] the mind represents its internal processing in a schematic and partial way. On Graziano's (2013) sparse illusion account, the mind creates a schematic and simplified representational model of the process of attention. According to Kammerer, on these sparse-illusion views, the illusion of phenomenal consciousness consists in a *cognitively penetrable* erroneous inference (which results from a correct but schematic cognitively impenetrable representation of cognitive processes). The fact that the illusion consists in a *cognitively penetrable* erroneous inference entails that the illusion ought to be quite weak. It should disappear with a bit of belief revision, as the headless woman illusion does. But the illusion of phenomenal consciousness is powerful, persisting no matter what one comes to believe about the reality of phenomenal consciousness. Therefore, contends Kammerer, a sparse theory such as Dennett's or Graziano's cannot explain the illusion.

I believe Kammerer is right, in light of the cognitive impenetrability of the illusion of phenomenal consciousness, to claim we need a rich-illusion view. Although Kammerer (2019c: 8) classifies the inference theory as a sparse-illusion view (based on an earlier draft of this paper), I think it is better characterized as a rich-illusion view in its current form. As explicated in §5, the illusion is constituted by an erroneous cognitively *impenetrable* unconscious inference. It is not, as on a sparse theory such as Dennett's[32] or Graziano's, caused by a cognitively *penetrable* inference. Furthermore, on the inference theory, the illusion is not generated by partial or schematic representations of brain processes, as on Dennett's or Graziano's theories. Instead, it is generated by introspective opacity, the *absence* of representation of brain processes.

[31] Dennett describes the illusion of phenomenal consciousness as a 'user illusion'. He posits that we are subject to a user-illusion vis-à-vis computer icons to the extent that we are blind to the hidden computational complexity activated upon clicking on them. Analogously, Dennett suggests we are subject to a 'user-illusion' with respect to consciousness when we mistake the mind's schematic and simplified representations of its mental goings-on for an accurate depiction of the whole story, and we fail to realize that what transpires in the mind is far more complex and detailed.

[32] Kammerer (2019c: 8) considers and rejects a possible interpretation of Dennett's user-illusion view as a rich-illusion view: '… [I]nterpreting this mistaken inference as a non-cognitively penetrable inferential process (which I think is not the natural interpretation here) would make Dennett's view a rich-illusion view'.

10 Conclusion

Suppose that our sensory states were neither qualitatively like anything nor (robustly or introspectively) represented as such. Could there be a subject who *judged* her sensory states not to be like anything? I claim no such subject is possible. A subject forming ICJs distinguishing her red* and green* sensory states who simultaneously judged there was no reason to distinguish them would not *really* distinguish them in light of the justification constraint. Given introspective opacity, or lack of access to the distinct physical/neural properties of her sensory states—which, hypothetically, if she had introspective access to them, she might take as the properties on the basis of which she is justified in distinguishing her sensory states—the subject unconsciously infers the existence of justificatory/phenomenal properties, p-redness and p-greenness, as stand-ins. Although she does not (robustly or introspectively) represent what these phenomenal properties are like, their WILs, she is firmly convinced that she does. This conviction generates problem intuitions vis-à-vis ineffability, non-physicality, etc. The illusion is explained as an inevitable non-adaptive side-effect of introspection. The inference theory has the virtue of solving the illusion problem and the illusion meta-problem, while at the same time avoiding the PRP and, as a rich-illusion view, avoiding the worries Kammerer (2019c) identifies for sparse-illusion views.

References

Armstrong, D.M. 1968. The headless woman illusion and the defence of materialism. *Analysis* 29 (2): 48–49.

Chalmers, D.J. 1996. *The conscious mind: In search of a theory of conscious experience*. New York: Oxford University Press.

Chalmers, D.J. 2010. *The character of consciousness*. New York: Oxford University Press.

Chalmers, D.J. 2018. The meta-problem of consciousness. *Journal of Consciousness Studies* 25 (9–10): 6–61.

Clark, A., K. Friston, and S. Wilkinson. 2019. Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies* 26 (9-10): 19–33.

Dennett, D.C. 1988. Quining Qualia. In: Marcel, A. & Bisiach, E. (eds.) Consciousness in Modern Science, Oxford University Press.

Dennett, D.C. 2017. From Bacteria to Bach and Back: The evolution of minds. W. W. Norton & Company.

Fodor, J.A. 1983. *Modularity of mind: An essay on faculty psychology*. Cambridge: MIT Press.

Frankish, K. 2019. The consciousness illusion. Retrieved 29 November 2020 from aeon.Co: https://aeon.co/essays/what-if-your-consciousness-is-an-illusion-created-by-your-brain

Frankish, K. 2018. What kind of illusion is the illusion of consciousness? *Workshop*: *Illusionism as a Theory of Consciousness*. LABEX IEC, Institut Jean Nicod, Paris, 25 May 2018. [PowerPoint presentation,

retrieved 29 January 2019 from]: https://onedrive.live.com/view.aspx?resid=56819066950EDA39!471885&ithint=file%2cpptx&authkey=!AP3U-qyiCZvNFfY
Frankish, K. 2016. Illusionism as a theory of consciousness. *Journal of Consciousness Studies* 23 (11–12): 11–39.
Gibbons, J. 2005. Knowledge, the priority of. In *Encyclopaedia of Philosophy*. Retrieved 29 Jan 2019 from Encyclopedia.com: http://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/knowledge-priority
Graziano, M.S.A. 2013. Consciousness and the social brain. Oxford University Press.
Humphrey, N. 2019. Easy does it. *Journal of Consciousness Studies* 26 (9–10): 105–114.
Humphrey, N. 2016. Redder than red illusionism or phenomenal surrealism? *Journal of Consciousness Studies* 23 (11–12): 116–123.
von Helmholtz, H. 1867. *Handbuch der physiologischen Optik*. **3**. Leipzig: Voss.
Kammerer, F. 2016. The hardest aspect of the illusion problem—And how to solve it. *Journal of Consciousness Studies* 23 (11–12): 124–139.
Kammerer, F. 2018. Can you believe it? Illusionism and the illusion meta-problem. *Philosophical Psychology* 31: 44–67.
Kammerer, F. 2019a. The illusion of conscious experience. *Synthese*, online first.URL: https://doi.org/10.1007/s11229-018-02071-y.
Kammerer, F. 2019b. The meta-problem of consciousness and the evidential Approach.*Journal of Consciousness Studies* 26 (9-10):124-135 (2019).
Kammerer, F. 2019c. How rich is the illusion of consciousness? *Erkenntnis,* online first.URL: https://doi.org/10.1007/s10670-019-00204-4.
Leibniz, G. W. 1969. Discourse on metaphysics, section 9. In *G. W. Leibniz: Philosophical Papers and Letters, 2nd ed*, ed. & trans. L. Loemker. Dordrecht: D. Reidel.
Nagel, T. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.
Papineau, D. 2008. *Thinking about consciousness*. Oxford: Clarendon Press.
Pereboom, D. 2016. Illusionism and anti-functionalism about phenomenal consciousness. Illusionism*: special issue*, ed. K. Frankish. Exeter, UK: Imprint Academic.
Pereboom, D. 2019. Russellian monism, introspective opacity, and the illusion meta-problem of consciousness. *Journal of Consciousness Studies* 26 (9–10): 182–193.
Schwarz, W. 2019. From sensor variables to phenomenal facts. *Journal of Consciousness Studies* 26 (9-10): 217–227.
Schwitzgebel, E., 'Belief', *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL: https://plato.stanford.edu/entries/belief/#OccuVersDispBeli
Setiya, K. 2008. Believing at Will. *Midwest Studies in Philosophy* 32 (1): 36–52.
Williams, B. 1976. Deciding to believe. In *Problems of the Self*. Cambridge: Cambridge University Press.