Defending Internalism about Unconscious Phenomenal Character

1. Introduction

Two important questions arise concerning the properties that constitute the phenomenal characters of our experiences, notably of perceptual experiences: first, where these properties exist, and, second, whether they are by nature tied to our consciousness of them. The first question has received a lot of attention. Answers range from the suggestion that the redness featuring in a perceptual experience of an apple (say) inheres in the apple's surface, and that the apple itself contributes this redness to perceptual experience (at least in veridical perception), to the suggestion that the experienced redness is instead a property of a kind of internal, mental, object—a sense-datum. Many sorts of answer lie between these two extremes, and there is even a radical alternative outside this spectrum, which would eliminate phenomenal characters and their constituting properties altogether (Frankish, 2016). The second question receives less attention,[1] but it is at least as important, though perhaps even more controversial. In influential work, David Rosenthal (e.g., 2005) has long argued that what he terms 'mental qualities', notably sensory and perceptual qualities, can exist unconsciously, and that these unconscious mental qualities play important functional and theoretical roles, especially as regards subjects' perceptual capacities. As against Rosenthal, and a small band of sympathisers,[2] by far the majority of contemporary

[1] Though it is hard to know whether this is because most philosophers simply assume the relevant properties are indeed tied to consciousness, as opposed to deeming the issue less important.

[2] See, e.g., Lockwood (1989), Clark (1993), Pitt (2004, p. 3 n. 4), Marvan and Polák (2017), Polák and Marvan (2019), Coleman (2024b).

philosophers believe that the qualities of experience cannot exist except as experienced; that is, consciously.[3]

There is some terminological difficulty involved in expressing the views at issue as regards this second question, but we will construe the question as being whether *phenomenal qualities*, or, equivalently for us, *phenomenal character* can exist unconsciously. For many, though not all, theorists 'phenomenal' connotes 'conscious', however. So, then, such theorists can construe our question as being whether the qualitative nature of phenomenal character is essentially, or by nature, phenomenal—that is, whether such qualitative characters could exist outside the subject's consciousness of them. For our purposes, 'phenomenal qualities/character' serves a *reference-fixing* purpose: it picks out qualitative properties that are familiar from conscious experience, though without any commitment that their consciousness—their being experienced—is inseparable from them; the terminology leaves this issue open. This is analogous to considering the everyday features of water—its potability, transparency, and so on—as reference-fixers, without a commitment that they are inseparable from water (i.e., H_2O) as such. This, in turn, allows us to talk about 'unconscious phenomenal qualities/character'—meaning that certain of the kinds of qualitative property familiar from conscious experience could exist unconsciously. At any rate, this terminological policy is not one to which our opponents should object, since they too discuss, and indeed posit, unconscious phenomenal qualities.

The two questions about phenomenal qualities, though distinct, intersect. For example, a naïve realist believes the red phenomenal quality of a veridical perception of an apple belongs to the apple—this philosopher is an externalist about perceptual phenomenal qualities.[4] This externalism could, it seems, naturally accompany the view that such

[3] See, e.g., Galen Strawson (1994), Uriah Kriegel (2009), Charles Siewert (1998), Angela Mendelovici (2018), Ned Block (2011, p. 424).

[4] Soteriou (2020) calls this view "naïve realism about phenomenal character".

phenomenal qualities can exist without consciousness. One would only have to believe that the apple's redness persists, as such, when no one is perceiving it. On this view the redness of the apple one perceives is independent of the consciously experienced redness; what the experienced redness and the unexperienced redness have in common is simply that redness quality (of the apple). Alternatively, one could deny that, when unperceived, the redness exists on the apple in the same way it does when perceptually experienced by a subject. One might even say that, strictly, the apple was not red, or at least not in the same sense, when unperceived. For this theorist the quality of redness is dependent on conscious experience, in that it only strictly exists when such experience is occurring. Hence, we can summarise their positions by saying that the first naïve realist philosopher is an 'independence externalist', and the second a 'dependence externalist' about the properties constitutive of qualitative perceptual phenomenal characters: what they agree on is the location of phenomenal qualities, what they disagree on is their relation to consciousness. Within this framework, Rosenthal is an independence internalist, since he holds that mental qualities are internal to the subject.[5] A qualia theorist who believes that the qualia belonging to mental sense-data can only exist as experienced would exemplify dependence internalism.

Considering the prevalence of dependence, it is perhaps surprising that arguments for it, or against independence, are rare.[6] But what is usually presumed is that the independence theorist can embrace internalism or externalism as they wish, making a wide range of views about the structure of perceptual experience, in particular, open to them. In recent work, though, Paweł Zięba (2022) argues that there are reasons why independence theorists would do better to embrace externalism. If he is correct, this would create difficulties for independence theorists who want to endorse anything

[5] See Rosenthal (1991).

[6] See Coleman (2022b) for a critical survey of such arguments.

resembling traditional internalist sense-datum theory, and more modern descendants.[7] In this paper we will defend independence internalism (henceforth abbreviated as 'II'), by refuting Zięba's argument. The upshot is that the independence theorist remains free to embrace internalism or externalism as they see fit, though we will also briefly give some reasons why internalism may well be preferable on balance.[8] To be clear: our defence of II consists, in the main, in blocking Zięba's argument. We will not offer much in the way of reasons for endorsing independence in general. Rather our arguments are directed to those who already accept, or entertain, independence, and who may be swayed by the sorts of consideration Zięba raises when it comes to the issue of whether to be an internalist or an externalist. We will not, in particular, be aiming to persuade those who are dependence internalists of the truth of independence.[9]

Though issues about dependence/independence and internalism/externalism go far wider than debates over perceptual phenomenal qualities, these qualities, and especially colours, make a good test case. Zięba mostly focuses his arguments on phenomenal colour qualities, and we follow suit.[10] Zięba proposes a novel account of

[7] For example, it would cast doubt on Rosenthal's internalist 'quality-space theory' of mental colour qualities (Rosenthal, 2005; 2015), and on the internalist proposals of Marvan and Polák (2017) and Coleman (2024b).

[8] Zięba himself uses the terms 'inequivalence' and 'equivalence' where we talk of independence and dependence, respectively. We do not find his terminology maximally clear, hence we substitute our own. This does not affect any substantive issue.

[9] For positive arguments in favour of II, especially of independence, see Coleman (2024b) and Marvan (2024).

[10] Interesting questions, which we will not explore here, concern whether independence internalism holds good of emotional qualitative characters (see Coleman, 2024a), cognitive qualitative characters (if such there be—see, e.g., Pitt, 2004 and forthcoming, Coleman, 2022a). Arguably, the issue of whether colour qualitative character is independent bears on questions about whether there can be unconscious mental imagery (Coleman MS).

unconscious colour perception, combining a variant of primitivism about colours (see, e.g., Byrne & Hilbert, 2007) with an unorthodox relationalist account of perception. The primitivist view Zięba propounds is that colours are *sui generis* (i.e., non-reducible) mind-independent properties of external objects. Relationalism, in turn, construes perception as a state of being directly related to perceptual objects and their properties (see, e.g., Nanay, 2017). So, for instance, your perceiving a tree is accounted for by relationalists by saying that you are perceptually related to an external object, a tree. Zięba's innovation is that, in contrast to how relationalism is standardly understood, he broadens the scope of perceptual states to include instances of non-conscious perception. In his independence version of relationalism, non-conscious colour-perception episodes are explained in a way that parallels the explanation of conscious colour-perception episodes. If you unconsciously perceive a particular colour, you are unconsciously related to its mind-independent phenomenal character. Though this phenomenal character is not something of which you are consciously aware—you do not experience the colour in question—nonetheless your perceptual state, on this view, has every bit as 'colourful' a content as its conscious counterpart.

Now, *being unconsciously related to phenomenal characters* will surely sound suspect to many readers. Isn't all phenomenality invariably conscious? Zięba disagrees. By doing so, he is able to constructively address certain vexing issues in the philosophy of perception. For example, he shows that independence views are better suited to account for unconscious colour perception than dependence theories. Equally importantly, he notes that if unconscious phenomenal characters are admitted, perceptual relationalists no longer face the unpalatable choice between rejecting unconscious perception and denying that the relationalist analysis pertains to episodes of unconscious perception (cf. Berger & Nanay, 2016).[11]

[11] By admitting unconscious phenomenality, Zięba does not want to say that unconscious states have 'what-it's-like-ness'. It doesn't feel like anything to be in unconscious states. However, this does not imply that unconscious perceptual states lack all qualitativity. As Zięba notes, we can treat the relation

We share Zięba's conviction that qualitativity can be unconscious, and thus are committed to independence (see Author 1, 2017; Author 1, 2019; Author 2, 2015; forthcoming). That is advantageous because we can focus on the issue of the location and basis of unconscious phenomenal character, leaving the more controversial topic, of whether phenomenal character can exist unconsciously at all, to one side. Against Zięba, we will defend the more standard, internalist reading of the (potentially unconscious) phenomenal character of colours: this character is, in our view, produced internally to the subject, presumably in her brain. To repeat, we will not be concerned here to persuade dependence internalists (or externalists) of independence.

We will do two things to support II: we will first explain, and reject, Zięba's novel and ingenious argument against II. Then we will briefly present certain difficulties for independence externalism (IE) which II does not face—these concern cases where the perceptual relation fails, but a perceptual experience still results. In tandem, these points suffice to defuse Zięba's critique of our view, and support II as the most plausible account of unconscious perception of colours (UPC), as against IE. Thus we aim to turn the tables against IE, in favour of II. First we turn to Zięba's argument against II.

2. The dilemma

Zięba (2022) builds his case against II on a dilemma he thinks besets internalist, but not externalist accounts of unconscious colour perception. In this section we briefly

between phenomenal character and what-it's-like-ness *dispositionally*: one can say that phenomenal character determines what-it's-like-ness just in those cases where the state having phenomenal character becomes conscious. This is not to say, however, that the unconscious states only have as-if or dispositional phenomenal character. What's dispositional is not the phenomenal character, but one's being conscious of it.

summarize the dilemma, and in subsequent sections show that it does not in fact threaten II.

The starting point of Zięba's dilemma is Block's (1995) distinction between access and phenomenal consciousness, and the possibility of 'phenomenal overflow' the distinction makes possible. Whereas a state is phenomenally conscious when it feels like something to be in it, a state is access conscious when we can report the state and use it to guide our reasoning and actions. Phenomenal overflow is the thesis that there are situations in which the subject is phenomenally conscious of a stimulus or some of its aspects, but she is not access-conscious of the stimulus or those aspects—in a nutshell, though she has a relevantly contentful experience, she cannot do anything with the experience or its content cognitively.

Zięba, inspired by Phillips's (2018) scepticism about unconscious perception, argues that the plausibility of there being unconscious perception of colours (UPC) is "inversely proportional" to the plausibility of phenomenal overflow. The more probable the overflow hypothesis, the less probable that there can be UPC. Hence, when UPC proponents appeal to overflow, they actually undermine UPC, because the behaviour of apparently 'unconsciously perceiving' subjects may in fact be guided by "residual or transient" conscious, but unreportable—phenomenally overflowing—visual information.

What has this to do with II, precisely? Zięba (2022, p. 28) claims that phenomenal overflow and II share the assumption that visual phenomenal character is produced in the visual cortex. On the first horn of his dilemma, the internalist advocate of UPC who accepts overflow cannot offer any evidence to support his position, since any evidence he adduces will just be evidence for phenomenal overflow instead. If this is right, accepting II actually puts one at odds with UPC. And that is bad, of course, because II vigorously asserts the existence of unconscious phenomenal colours.

Zięba's second horn argument has two parts. In the first he argues that denying overflow—the first horn turns on the UPC advocate's accepting overflow, as above—entails that visual phenomenal character is not produced exclusively in the visual cortex. Since independence internalists tend to assert that it is, this is bad news for them. Then he uses this result in the argument's second part to establish that there is *no possible neural basis* for visual phenomenal character, given UPC. That would falsify II, on the grounds that there is nowhere inside the subject's brain that independence internalists could house the unconscious phenomenal characters they posit. For this reason, he concludes, proponents of unconscious phenomenal characters are better off with the externalist variant of the independence hypothesis.

It is important to note that Zięba does not claim to have given a decisive *proof* that independence can only be combined with an externalist account of phenomenal colours. Strictly what he claims is that IE is dialectically and evidentially in a better position than II, in that it does not have to confront the possibility that alleged cases of UPC are in fact cases of phenomenal overflow. So when Zięba says one should prefer IE over II, he does not wish to imply that internalist accounts of phenomenal character are wholly unmotivated or incoherent. His worry is that any difference between genuinely unconscious perception and overflow the II-theorist can point to won't be sufficiently robust to remove all reasonable doubt that these situations are in fact the same: hence that independence genuinely obtains. IE, by contrast, is better placed to make good on independence, since it can avoid the 'overflow challenge', as we might call it. In short, motivated by Phillips-style worries about unconscious perception in the light of the possibility of overflow, Zięba's response is to seek to protect independence by siding with independence externalism, which he sees as less vulnerable to Phillips's overflow-based critique.

However, fortunately for independence internalists, Zięba's arguments fail. Moreover, reflection on wider connections between independence and theories of perception may if anything strengthen the hand of the independence internalist. Below we consider each horn in turn.

2.1 *The first horn*

The first horn of the dilemma flows from the independence internalist's accepting overflow. The argument is that phenomenal overflow and II are supported by the same type of evidence, and that that is a problem because these two views are in tension. The two hypotheses will tend to have the same evidence base, one might think, because in both cases one will cite examples of subjects who show signs of successful perception despite not being able to report any conscious state involving colours. Therefore, either there is an unreportable (hence inaccessible) conscious perceptual state present, a case of overflow, or, alternatively, there is an unconscious (hence inaccessible) perceptual state present, which is guiding the subject's responses. Hence Zięba's claim that any evidence adduced for internalist UPC is evidence, too, for overflow, a rival hypothesis. Overflow is a rival to II, clearly, since it is consistent with the falsity of independence. And since the kinds of case the independence internalist might point to as evidence of independence are exactly those where it is plausible that overflow occurs, II advocates will struggle to provide evidence for their position, if Zięba is correct.

Put in neural terms, the first horn argument is as follows: Phenomenal overflow plausibly entails that the phenomenal character of visual perception is produced in the visual cortex, and is independent of the neural basis of access consciousness. That is because conscious phenomenal character can exist inaccessibly, given overflow, and access consciousness is widely believed to have its neural basis in frontal areas—hence

conscious visual phenomenal character cannot also have its basis in frontal areas. The other standard alternative for the proposed neural basis of visual phenomenal character is the visual cortex. If II is true, visual phenomenal character will also be independent of access, since it can exist unconsciously. That means the independence internalist must house visual phenomenal character in the visual cortex, just like the overflow theorist. But this, in turn, means that any neuroscientific evidence internalist UPC advocates offer for their theory will also support overflow, a thesis associated with the opposing position which denies visual phenomenal character can exist unconsciously. If the evidence for II is just evidence for overflow, II is in trouble, since overflow can be used to deny UPC but II asserts it.

However, this argument does not touch II, on reflection. First, the argument doesn't settle anything by itself regarding the truth of II. In fact, one might think, the argument cuts both ways, if it cuts at all. If evidence for II is evidence for phenomenal overflow, then, by parity, the reverse would seem to apply too: evidence for overflow is also evidence for II. That is, apparent cases of overflow could be diagnosed, instead, as instances of unconscious colour perception, undermining the case for overflow. At worst, then, neither the II-opposed overflow theorist nor the independence internalist would gain support from any evidence about phenomenal qualities being produced inaccessibly in the visual cortex: such evidence would simply be neutral, on balance, as to whether such qualities were inaccessible because they were unconscious or inaccessible in spite of being conscious. This result doesn't count against II as a thesis at all. It does nothing to the debate, in fact, except perhaps neutralise a source of evidence for both sides.

But, anyway, second, Zięba is wrong to think that II and phenomenal overflow share an evidential base. We *can* in fact say what would (and does) support II over overflow, empirically speaking. As concerns behavioural evidence, subjects' reports in overflow experiments indicate that they were phenomenally conscious of the stimuli they were

presented with. As Zięba himself notes, subjects in Sperling's (1960) experiments, the classic source of evidence for phenomenal overflow, claim that they phenomenally experienced all or almost all of the briefly presented letters, even if they could only recall a fragment of them (i.e., in Block's terms, they were not access-conscious of the letters they could not recall). In contrast, II posits *truly unconscious* phenomenal qualities. Truly unconscious phenomenal qualities should elicit the report: "I did not see anything", not the report: "I saw the letters but cannot name all of them now." The behavioural evidence for phenomenal overflow is therefore not at all likely to be the same as the behavioural evidence for the unconscious phenomenal characters II posits.

Zięba points out that according to Cova et al. (2021), evidence for the reports in overflow studies is questionable. Cova et al. found that most subjects in the Sperling experiments (and in the more recent Landman experiments—see Landman et al., 2003) did not claim to have seen all aspects of the presented stimulus, but rather opted either for "partial" or "generic" reports on their phenomenology, where a partial phenomenology report is something like 'I saw only some of the stimuli presented' and a generic phenomenology report is something like 'I saw all stimuli but only generically, not in their fine perceptual details'. According to Cova et al. it is an 'urban myth' that overflow experiment subjects report having seen all the presented stimuli. This is a fine point, but we fail to see how it could cast doubt on II. Cova et al. do not wish to argue that it is difficult to disentangle overflow and other variants of conscious phenomenology from cases of genuinely unconscious perception. To the contrary, they note that pretty much no subjects in their experiments reported having *no* phenomenology of the presented stimuli, which is the type of report one would associate with genuinely unconscious perception, as we noted earlier.[12] To repeat, then, the behavioural evidence for overflow, and that for II, are importantly different.

[12] See Table 2 at p. 434 in Cova et al. (2021).

It is equally clear, on reflection, that Zięba is wrong to assume that overflow and II share a neuroscientific evidence base. For independence internalists, it would be very helpful to have a neuroscience-backed way of distinguishing between activations in the visual cortex that correlate with unconscious perception of colours, and those that correlate with the conscious perception of colours.[13] With that distinction we could distinguish states that are phenomenal but unconscious from states that are phenomenal and conscious (including the states that overflow). As a matter of fact, the visual areas in the occipital part of the brain do activate in ways that enable such discrimination. These areas activate in both conscious and unconscious perceptual conditions, but with important differences. To begin with, their activation profiles in conscious and unconscious conditions differ in activation strength. Unconscious activations brought about by the same stimulus are weaker than the conscious activations (Fontan et al., 2021; Moutoussis & Zeki, 2002; Stein et al., 2021; Tong et al., 1998).

This does not mean that there is an absolute quantitative threshold, and that all activation above that threshold corresponds to conscious vision (and activation below it is unconscious). It can happen that a consciously seen stimulus evokes a smaller neural response than another, stronger stimulus which is not registered consciously. For instance, a high contrast visual target can remain consciously undetected, whereas the same target with lowered contrast, and therefore weaker neural response, can be consciously detected (see Vugt et al. 2018, esp. fig. 2 at p. 539). However, it still holds that when we are comparing activation responses for two visual stimuli that are of

[13] The category of conscious visual states includes episodes of overflow as subset, for all episodes of overflow are phenomenally conscious states, but, plausibly, not vice-versa. Although we concentrate on overflow, the neural argument spelled out in this section does not apply exclusively to overflow episodes but to conscious visual episodes in general (that is, to phenomenally conscious visual episodes with or without access consciousness). The relevant contrast the argument draws upon is thus not 'unconscious visual episodes' vs. 'episodes of overflow', but 'unconscious visual episodes' vs. 'conscious visual episodes (including episodes of overflow)'.

exactly the same kind and share all their relevant properties, such as contrast and luminance levels, the activation for the consciously non-registered stimulus will be somewhat weaker than for the consciously registered one. Keeping track of the differences in activation strength is therefore one way of distinguishing colour stimuli that were experienced from ones that were not.[14]

In addition, relevant higher, extra-visual brain areas in the parietal and frontal cortex co-activate with the more posterior visual areas only when visual perception becomes conscious (Rees & Frith, 2017; Dehaene & Naccache, 2001). So we have two, systematically different, activation patterns here. On the one hand, occipital visual areas activate in subtly different ways in both unconscious and conscious conditions. On the other hand, higher areas only activate in conscious conditions. Thus the relevant higher areas in parietal and frontal areas are probably needed for the visual character of a mental state to be expressed consciously. However, this in no way precludes that the visual character that can be consciously expressed is produced in the visual areas and not elsewhere, as the independence internalist tends to believe. These higher areas may simply help the already established unconscious visual character to become conscious. They may do this in the capacity of, e.g., higher-order representational states (Rosenthal, 2005; Coleman, 2015), or another type of state or process that provides the unconscious visual qualities with whatever they need to become conscious (see Marvan, 2024).

Zięba notes the possibility that his dilemma for II may result from an over-simplistic view of how the brain works (p. 31). Still, he believes that complicating the picture with more neural details will not allow us to "trace the neural factory of phenomenal

[14] It might seem that we beg the question here, in assuming there can be unconscious perception at all. But that is not so: bear in mind that we are discussing Zięba's argument here, and he accepts that there is unconscious perception. His claim is that, given the possibility of overflow, UPC is in strongest shape if combined with externalism. That is what we are disputing.

qualities". The foregoing sketch of a neural story shows that, *pace* Zięba, it will indeed allow this, for the story is perfectly consistent with II. The independence internalist will hold that visual character is prepared in the visual areas first unconsciously, and then it is, in the same areas, expressed consciously, with the help of higher areas in parietal and frontal cortex (using, e.g., higher-order representational states). Accordingly, we can precisely distinguish episodes of overflow (and other conscious visual episodes) from episodes of truly unconscious perception of colours. Because it is an instance of visual consciousness, phenomenal overflow needs a stronger response in visual centers than unconscious activation (for the same stimulus), and it further needs involvement of the relevant parts in the parietal and frontal cortex.[15] In contrast, unconscious visual character needs only a somewhat weaker activation in the visual parts of the cortex and no involvement of the relevant higher areas. Recall that our dispute with Zięba is not about the locus of *conscious* visual character but about the locus of unconscious visual character (although on our independence internalist account this ground of visual character can be shared with corresponding episodes of conscious vision, of course). Even if the neural substrate of conscious colour vision spills over to non-visual brain areas (up to the prefrontal cortex), unconscious visual character can remain confined to the visual cortex, hence completely internal to the subject.

Recall now that Zięba claims that IE is preferable over II because it avoids the overflow challenge. However, it is not obvious that this is the case. Note that Zięba himself frames the issue in terms of our *perceptual relation* to unconscious qualities. And it is unclear that as an account of the perceptual relation involving qualities, IE has any real advantage over II in dealing with the overflow challenge. For phenomenal overflow to

[15] While access consciousness presumably has its neural basis in the frontal cortex, it would be an oversimplification to claim that all frontal activations accompanying perceptual states imply access consciousness. The activations in frontal areas we speak about here are necessary for *phenomenal* consciousness, not for access consciousness. At least this is how we interpret the findings referred to in the previous paragraph.

obtain, note, it is not in any way required that the visual character of colour be located internally to the subject. It is perfectly conceivable that a subject mistakenly believes that she does not perceive any colour qualities consciously, while being in a state of phenomenal overflow involving *external* colour qualities. That is, she could be perceiving external colour qualities phenomenally consciously without being able to report on her experience. In that case, IE will also face the overflow challenge, in so far as it is a challenge: purported cases of unconscious, but perceived, external colour qualities could be taken, instead, as cases of overflowing conscious external colour qualities. So, so far no advantage of IE over II is in sight; the possibility of overflow seems to be equally challenging for internal and external unconscious phenomenal characters.

Presumably, Zięba thinks that the game changes when we zoom in on the neural implementation of visual phenomenal character. With II, the thinking goes, the same brain centers activating in unconscious perceptual conditions are also involved when colour perception becomes conscious. Because of that, one has grounds to worry that the allegedly unconscious internal colour qualities in fact overflow, i.e., are never unconscious. It might further be objected that such differences in conscious and unconscious neural activation profiles as we indicated above are not robust enough to exclude the possibility of phenomenal overflow beyond all reasonable doubt. By keeping the phenomenal qualities far away from the neural substrate of consciousness—outside the head—IE does not generate this worry and is thus more suitable for accommodating UPC.[16] This apparent advantage for IE can be put in terms of the modal profile of the theory with respect to overflow: if overflow is false, IE and II are on a par as regards accommodating and supporting the thesis of independence. But if overflow is true, IE is better placed to accommodate and support independence,

[16] We thank one of the reviewers for pressing us on this point.

since it avoids the overflow challenge. Hence, overall, given the possible scenarios, IE is better placed to accommodate and support independence than II.

However, recall again that the debate between II and IE is about the (unconscious) perceptual relation involving colour qualities. And in order to get into such a perceptual relation, one's perceptual brain centers will have to be involved. For visual contents, these centers will presumably be located in the visual cortex. Therefore, even unconscious perception of externally located phenomenal characters requires the involvement of the visual cortex. Hence, even if IE locates phenomenal characters themselves 'far away' from the substrate of consciousness, it can hardly locate the perceptual mechanisms of such characters far away from the substrate of consciousness. This implies that even in its neuroscientific setting the problem of overflow can affect IE to just the extent that it allegedly affects II. So, in fact, II and IE have the same 'modal profile' with respect to accommodating and supporting independence, given the possibility of overflow. There is no advantage here for IE.

To summarise: if our reasoning is correct, the first horn argument does not provide IE with any advantage over II. The battleground must shift: to adjudicate between IE and II, one must turn to a different type of argument because the "dialectical advantage of IE over II" argument does not work. And that different argument is indeed provided by Zięba in his second horn (section 6.3 of his paper). Here one finds an attempt to go beyond the overflow challenge and motivate the expulsion of visual qualities from the brain and into the external world. To this novel argument we therefore turn in the next section.

2.2 *The second horn*

Our interim conclusion, regarding the first of Zięba's two horns, is that the possibility of phenomenal overflow casts no doubt on II, since overflow and truly unconscious perception of colours manifest in behaviourally and neurally distinct ways. Moreover, even granting Zięba's assumption about the shared evidence of II and overflow does not give him an effective first horn, either, not least since IE would also face any overflow challenge. Upshot: the independence internalist need have no fear of phenomenal overflow.[17]

Horn two is posed for the II advocate who *rejects* overflow. The idea is that if the neural basis of visual phenomenal character is not exclusively in the visual cortex, but also at least partially in the prefrontal cortex (the seat of access consciousness), as Zięba claims the rejection of overflow entails, II is false. As already noted, the argument has two parts. Here are the premises and conclusion of the first part:

(1) the phenomenal overflow hypothesis is false (assumption);
(2) there is no phenomenal consciousness (PC) without access consciousness (AC) (from 1);
(3) the neural basis of PC is the same as the neural basis of AC (from 2);
(4) the neural basis of AC is not located exclusively in the visual cortex (an empirical fact);
(5) the neural basis of visual PC contains (or is identical to) the neural basis of the phenomenal character of visual perception (from PC's definition);

[17] Although the neural argument spelled out in the previous section allows us to distinguish conscious from unconscious perceptual states, it doesn't tell us, as such, what the markers specifically of overflow are going to be. For that some further empirical work would be needed. If one identifies frontal areas implementing access consciousness, distinguishes them from other frontal areas coactivating with phenomenally conscious visual states, and further shows that in some situations the first are inactive and the latter active, one obtains a marker of overflow. So, in theory at least, such markers are possible.

(6) the phenomenal character of visual perception cannot be produced exclusively in the visual cortex (from 3, 4 and 5).

And those of the second part:

(6) the phenomenal character of perception cannot be produced exclusively in the visual cortex;
(7) the phenomenal character of [visual] perception is consciousness-independent (from UPC);
(8) the activity in the neural basis of AC does not suffice for the [visual] phenomenal character to occur (from 7 [and from 3]);
(9) the phenomenal character of colour perception is not produced in the subject (from 6 and 8).

And here is our gloss on the reasoning of each part:

Part 1: If overflow is false (1), phenomenal consciousness and access consciousness always co-occur (2), hence they have the same total neural basis—that is to say, the same things happen neurally when either occurs, since they always go together (3).[18] Now, it is widely held that the neural basis of access consciousness involves frontal areas, hence visual cortex activity alone does not suffice for access consciousness (4). We can also say that the neural basis of visual phenomenal consciousness contains that of visual phenomenal character either as a proper or improper part—since visual phenomenal consciousness features such phenomenal character as a component (5). But, given the rejection of overflow, the neural basis of access consciousness coincides with that of phenomenal consciousness, and we know from the above that this neural

[18] We grant this premise for the sake of the argument. Note, though, that Block (2005) insists that the neural bases of phenomenal and access consciousness are importantly distinct, and that we need to distinguish two kinds of neural correlates of consciousness—PC-correlates and AC-correlates.

basis outstrips the visual cortex. Hence, Zięba concludes, the neural basis of visual phenomenal character, given the rejection of overflow, also outstrips the visual cortex (6). But this is bad for II, since its proponents typically assume that visual phenomenal character *is* in fact based in the visual cortex. Hence advocates of II who reject overflow cannot say what they want to about the neural basis of visual phenomenal character.

But worse is to come, as per the reasoning of part 2: If visual phenomenal character is not produced exclusively in the visual cortex (6), then clearly it must be produced elsewhere in the brain, if one is an internalist about it. The obvious, perhaps only, candidate in the offing would be prefrontal neural areas. However, the proponent of II is not just an internalist about visual phenomenal characters, of course; she also holds that they can occur phenomenally unconsciously (7). But now she is in a bind, thanks to her rejection of overflow on this horn. That rejection means the neural area that supports access consciousness also supports phenomenal consciousness. But the neural region in question is, again as widely held, precisely the prefrontal areas that are now the independence internalist's only remaining candidate for the basis of unconscious visual phenomenal character. Yet, by hypothesis, the activity of these areas produces phenomenally conscious states, so is unavailable for independence internalists' posited unconscious qualitative characters (8). Hence, there is nowhere in the brain for independence internalists to house unconscious phenomenal characters. So II cannot be true—if there is unconscious phenomenal character it can only occur outside the brain, hence, given reasonable assumptions, outside the subject altogether (9)—hence IE would be motivated, assuming the truth of independence.

We can now see that Zięba's argument actually does double duty: it not only serves to cast doubt on II, but, additionally, it provides a positive argument for IE, at least on its second horn. And since advocates of II tend to reject overflow, it would seem that their own assumptions put II in doubt, and indicate that IE is in a better position to account for UPC. Even if horn one fails, the fact that the II advocate is likely to reject overflow

means we can more or less treat horn two as a standalone argument against II. And, taken as such, horn two constitutes an ingenious and on the face of it powerful argument against II, and for IE.

However, on closer examination the second horn argument is also unsuccessful. Consider the first part first. Presumably, by "phenomenal character of visual perception" in (5) and (6) Zięba means *consciousness-independent* phenomenal character, i.e., that which on II can exist unconsciously.[19] But if so, (6) does not follow from premises (3), (4) and (5). We can reformulate premise (5) as:

(5)' the neural basis of visual PC contains (or is identical to) the neural basis of *consciousness-independent* phenomenal character of visual perception

And the conclusion (6), accordingly, as:

(6)' the *consciousness-independent* phenomenal character of visual perception cannot be produced exclusively in the visual cortex.

On this reading, the first part of the second horn argument just does not touch the issue of where the consciousness-independent phenomenal character may or may not be produced, and therefore poses no threat to II. Zięba assumes that if the hypothesis of overflow is false, visual qualities simpliciter cannot be produced in the visual cortex.

[19] Suppose Zięba meant conscious visual qualitative character, instead. Then (6) would say nothing about where the independence internalist's posited consciousness-independent visual character was produced, and would not impact II. Nor would the second part of the horn two argument work. And clearly, if horn two varies, sometimes meaning conscious visual qualitative character, and sometimes consciousness-independent visual character, by "phenomenal character of visual perception", it will come out as invalid by virtue of equivocation. Hence Zięba must mean consciousness-independent visual character by "phenomenal character of visual perception", throughout horn two. Thanks also to a reviewer, here, for advocating this reading.

But this just does not follow. As per the neural argument in the previous section, it is empirically plausible to assume that both conscious and unconscious qualities depend on activations in the visual cortex. The key difference is that whereas conscious visual character further depends on co-activations in more frontal areas, there is no reason to think that this is also true of the consciousness-independent visual character. Schematically, overflow = conscious visual character is produced in the visual cortex, whereas II = consciousness-independent visual character is produced in the visual cortex. (5), read as (5)', can be true if the component of PC responsible for visual character, that is, consciousness-independent visual character, has its neural basis exclusively in the visual cortex. A full state of PC, involving as it does frontal/pre-frontal activity, can contain the relevant visual cortex activity as a proper part, even though in cases of unconscious visual perception II will hold that the visual cortex alone is responsible for visual character. This is all totally consistent with II, so (6) simply will not follow. Hence, for all part one of the second horn argument says, the basis of consciousness-independent visual phenomenal character can still be confined to the visual cortex.

The second part of Zięba's argument (6–9) is equally ineffective, for much the same reason. The argument starts with a false premise as far as II is concerned, as just explained—'the *consciousness-independent* phenomenal character of visual perception cannot be produced exclusively in the visual cortex'—and thus ends up being unsound. On this reading (6) will fail to exile phenomenal character from the visual cortex—the aim of the argument being to leave II's possibly-unconscious visual phenomenal character with no home in the brain. Hence the two-step second horn argument, overall, fails. Rejecting phenomenal overflow does not commit one to expelling visual character from the visual cortex, let alone from the brain. Unconscious visual phenomenal character can be produced in the visual cortex, hence internally to the perceiving subject, as per II.

In light of the second horn's two-step argument, one may become tempted to think that according to Zięba, whatever holds for conscious versions of the visual qualities must equally hold for their unconscious counterparts. But this thought must be resisted, not least because Zięba himself exposes its incorrectness in a response to one of the objections he imagines his opponents to raise against his criticism of II (Zięba 2022, p. 30). In the end, what seems to do the main work in the two-step argument is the background assumption that rejecting horn two will lead us back to horn one, and that this will reinstate the challenge of phenomenal overflow.[20] Hence it is the possibility of overflow that is meant to constitute the main threat to II. However, given that horn one provides no distinctive problem for II that IE does not face, as we have shown in the previous subsection, Zięba's arguments do not seem to show that IE has any edge over II.

3. Concluding remarks

We have shown that Zięba's apparently powerful dilemma argument against II fails—advocates of II are free to embrace horn one or two without fear. Nor does horn two provide any positive support for IE.

This might seem to leave the debate at a stalemate—we have rejected the arguments against II and for IE, so all is equal. But let us close by suggesting some reasons why, in fact, II might well be preferable to IE, given that it is, if our argument is correct, still a live option in the debate for those who endorse UPC.

The problem for IE, as Zięba characterises it, comes due to its embroilment with a philosophy of perception, something II by itself stays largely clear of. IE, by contrast, is

[20] We thank one of the reviewers for suggesting this interpretation.

committed to *relationalism*, also called direct realism, about colour perception.[21] Zięba's view is that in visual consciousness it is external environmental colours that provide the phenomenal colour qualities of visual experience; they, as naïve realists say, directly constitute or form part of such experience.[22] Where IE differs from mainstream relationalism (dependence externalism) is in suggesting that these same environmental colour qualities can exist outside the subject's conscious perception of them, hence unconsciously.

Now, relationalism faces well known challenges concerning what goes on in suboptimal cases of perceptual experience, notably in illusion and hallucination, where the would-be perceptual object either is not as it seems to be or does not exist at all. *Prima facie*, if the colours present in my visual experience are contributed by external objects, it is hard to see how I could either misperceive those properties, or perceive there to be something coloured in my environment when there is in fact nothing relevant there at all. One might expect illusions and, even more so, hallucinations to be impossible. But of course they are not, hence relationists have to try to explain them away.

Two strategies predominate: relationalists either accept that in suboptimal cases phenomenal colour qualities are generated within the subject, or deny that in such cases there really is any visual phenomenal character in play. The first move is associated with what is called 'positive disjunctivism' (e.g., McDowell 1982). This is the view that in the good case the colour qualities of consciousness are contributed by external coloured objects, but in the bad case it is the subject herself who generates these colour qualities, or something appearing very much like them, in effect

[21] It might seem II is incompatible with all forms of direct realism. However, there are representationalist versions with which it may well fit, depending on the details. We will not explore that issue here. II is of course also compatible with internalist representationalist views of various sorts, a sense-datum theory, adverbialism, etc.

[22] Soteriou (2020).

mistakenly 'projecting' them, or taking them to be instantiated, outside herself. Positive disjunctivism does not seem a viable option for Zięba, since part of his case against II is that he doubts visual phenomenal colour characters can be produced *anywhere* in the subject's brain. But that means that such colour qualities cannot be produced in the subject's brain when she is subject to an illusory or hallucinatory experience, either.

Thus it looks like Zięba's only realistic option is to deny that illusory or hallucinatory experience features genuine phenomenal character. This view is sometimes called 'negative disjunctivism', or 'quietism', and is associated with Michael Martin.[23] The suggestion is that all that genuine conscious visual perception and bad cases have in common is that the subject cannot tell them apart, which is typically cashed out in terms of her being unable to discriminate the two situations, or judging (wrongly) that she is having an experience with such-and-such visual phenomenal character when in fact she is not.

There is not an outright contradiction involved, for independence externalists, in embracing this option, but it would seem ill-suited to their wider commitments. What the quietist seeks to safeguard is the thesis that colour qualities only ever really exist in the hum of the subject's conscious visual perceptual *relation* with the environment—hence the name 'relationalism'. That is their prime motivation for denying that, in suboptimal cases, there is any visual phenomenal character present at all, which in turn requires the quietist line focused on subject judgements. If the subject is not in the conscious visual perceptual relation with her environment, then there are no relevant phenomenal qualities present.[24] Independence externalists of course believe in visual

[23] E.g., Martin (2004).

[24] Cf. Kalderon's (2011) move here: he denies illusions are possible, on the grounds that they are in fact always veridical perceptions. Johnston (2004) deals with hallucinations by positing uninstantiated clusters of colour universals—perhaps Zięba could embrace a disjunctivism along these lines, but

colour qualities that exist outside this conscious perceptual relation (as Zięba says, on independence externalism phenomenal qualities exist "out there in the world, waiting … to be perceived"; p. 24). So IE seems to be in some tension with quietism, which is a purer form of relationalism about phenomenal colour qualities than the former positive disjunctivist forms of direct realism. The independence externalist might be seen as trying to have his cake and eat it here: insisting, with his fellow relationalists, that conscious colour qualities subsist in the hum of the world-involving perceptual relation, while maintaining that *the very same properties*, those selfsame colour qualities, can nonetheless still exist, when unconscious, outside this relationship.

Moreover, the relationalist is not in general a fan of phenomenal qualities, as compared to the independence internalist. II has time to explore the theoretical value of the phenomenology of dreams, for example, which if anything seems to be internally produced. Relationalists, especially quietists, tend, rather implausibly, to deny that dreams have any phenomenology at all. One tack, here, is for theorists to ascribe the supposed phenomenology of dreams instead to false memories of having dreamed.[25] Yet it is unclear what advance this makes for someone concerned to deny that phenomenal qualities are internally produced, since a subject's memories, and any phenomenology those memories involve, are so plausibly internally generated within the brain.[26] Quietists, of course, will swap the alleged phenomenology of dreams for false subjective judgements of having dreamt. We find that line implausible, again, but

---

Johnston's thesis is widely seen as extravagant. Surely II would be a more plausible option, given the common ground of UPC.

[25] E.g., Dennett (1976) suggests we confabulate memories of having dreamed between the putative end of the dream and waking.

[26] Hardcore direct realists can analyse memory, too, as involving relations to external things only, as a kind of time-traversing direct perception. However plausible this is for memories of external phenomena (a relationist could say that memory, like perception, represents or presents the outside world, albeit at a different time to the present), memories of dreams are *memories of experiences*, and do not relate the subject to anything outer.

that is not quite the point here. The point is that if the IE advocate is forced into quietism they must forgo talking about many sorts of mental episodes where phenomenal qualities seem to be involved, as with dreams. Yet fans of unconscious qualities tend to be fans of qualities *as such*. Put it this way, if one finds oneself 'going quietist' about many putative phenomenally qualitative mental episodes, the question arises of why one's quietism would not extend to posited unconscious phenomenal qualities. People who endorse independence do not tend to be very selective about which phenomenal qualities they allow in which circumstances; quite the reverse, they will tend to emphasise and extend the theoretical role of these properties.[27] So, again, IE seems to sit awkwardly with the desire to posit unconscious qualities. Given the failure of Zięba's critique of II, and of his case for IE, combined with the problems we have just highlighted for IE, we humbly suggest that, as a fan of unconscious qualities, Zięba, and those sympathetic to his argument, should join us in the independence internalist camp.

References

Berger, J., & Nanay, B. (2016). Relationalism and Unconscious Perception. *Analysis*, 76(4), 426–433.

Block, N. (1995). On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences*, 18, 227–287.

Block, N. (2005). Two Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, 9(2), 46–52.

[27] As with Rosenthal's take on the importance of perceptual phenomenal qualities and their functions—see, e.g., Rosenthal (2005). See also Coleman (2022a, 2022b, 2024a, 2024b, MS; Coleman and Montero 2023).

Block, N. (2011). The Higher-Order Approach to Consciousness is Defunct. *Analysis*, 71 (3), 419–431.

Byrne, A., & Hilbert, D. R. (2007). Color Primitivism. *Erkenntnis*, 66(1), 73–105.

Clark, A. (1993). *Sensory Qualities*. Oxford, New York: Oxford University Press.

Coleman, S. (2015). Quotational Higher-order Thought Theory. *Philosophical Studies*, 172, 2705–2733.

Coleman, S. (2022a). The Ins and Outs of Conscious Belief. *Philosophical Studies*, 179, 517–548.

Coleman, S. (2022b). Intentionality, Qualia, and the Stream of Unconsciousness. *Phenomenology and Mind* 22: 42-53.

Coleman, S. and Montero, B. G. (2023) Unconscious Transformative Experience. *Synthese* 202, 122 (https://doi.org/10.1007/s11229-023-04332-x).

Coleman, S. (2024a). A Feeling Theory of Unconscious Emotions. In *Conscious and Unconscious Mentality. Examining their Nature, Similarities and Differences*, edited by Juraj Hvorecký, Tomáš Marvan, & Michal Polák (pp. 207–226). London and New York: Routledge.

Coleman, S. (2024b). An Agument for Unconscious Mental Qualities. *Australasian Journal of Philosophy*.

Coleman, S. (MS). Mental Imagery and Unconscious Mental Qualities: How to Prevent Mental Imagery from Being a Cognitive Luxury.

Cova, F., Gaillard, M., Kammerer, F. (2021). Is the Phenomenological Overflow Argument Really Supported by Subjective Reports? *Mind and Language*, 36(3), 422–450.

Dehaene, S., & Naccache, L. (2001). Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, 79(1–2), 1–37.

Dennett, D. C. (1976). Are Dreams Experiences? *The Philosophical Review*, 85(2), 151–171.

Fontan, A., Lindgren, L., Pedale, T., Brorsson, C., Bergström, F., & Eriksson, J. (2021). A Reduced Level of Consciousness Affects Non-Conscious Processes. *NeuroImage*, 244, 118571.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, 23(11–12), 11–39.

Johnston, M. (2004). The Obscure Object of Hallucination. *Philosophical Studies*, 120, 113–183.

Kalderon, M. E. (2011). Color Illusion. *Noûs*, 45(4), 751–775.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press.

Landman, R., Spekreijse, H., Lamme, V. A. F. (2003). Large Capacity Storage of Integrated Objects before Change Blindness. *Vision Res*earch, 43(2), 149–164.

Lockwood, M. (1989). *Mind, Brain, and the Quantum: The Compound 'I'*. Oxford: Basil Blackwell.

Martin, M. G. F. (2004). The Limits of Self-Awareness. *Philosophical Studies*, 120, 37–89.

Marvan, T. (2024). The Brain-based Argument for Unconscious Sensory Qualities. In *Conscious and Unconscious Mentality. Examining their Nature, Similarities and Differences*, edited by Juraj Hvorecký, Tomáš Marvan, & Michal Polák (pp. 157–173). London and New York: Routledge.

Marvan, T., & Polák, M. (2017). Unitary and Dual Models of Phenomenal Consciousness. *Consciousness and Cognition*, 56, 1–12.

McDowell, J. (1982). Criteria, Defeasibility and Knowledge. *Proceedings of the British Academy*, 68, 455–479.

Mendelovici, A. (2018). *The Phenomenal Basis of Intentionality*. New York: Oxford University Press.

Moutoussis, K., & Zeki, S. (2002). The Relationship between Cortical Activation and Perception Investigated with Invisible Stimuli. *Proceedings of the National Academy of Sciences*, 99(14), 9527–9532.

Nanay, B. (2017). Philosophy of Perception: A Road Map with Lots of Bypass Roads. In *Current Controversies in Philosophy of Perception*, edited by Bence Nanay (pp. 1–20). London and New York: Routledge.

Phillips, I. (2018). Unconscious Perception Reconsidered. *Analytic Philosophy*, 59(4), 471–514.

Pitt, D. (2004). The Phenomenology of Cognition. Or *What Is It like to Think That P? Philosophy and Phenomenological Research*, 69(1), 1–36.

Pitt, D. (forthcoming). *The Quality of Thought*. Oxford: Oxford University Press.

Polák, M., Marvan, T. (2019). How to Mitigate the Hard Problem by Adopting the Dual Theory of Phenomenal Consciousness. *Frontiers in Psychology*, 10: 2837.

Rees, G., & Frith, C. D. (2017). Methodologies for Identifying the Neural Correlates of Consciousness. In *The Blackwell Companion to Consciousness*, edited by Susan Schneider & Max Velmans (pp. 589–606). Chichester: John Wiley & Sons, Ltd.

Rosenthal, D. M. (1991). The Independence of Consciousness and Sensory Quality. *Philosophical Issues*, 1, 15–36.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford: Oxford University Press.

Rosenthal, D. M. (2015). Quality Spaces and Sensory Modalities. In *Phenomenal Qualities: Sense, Perception, and Consciousness*, edited by Paul Coates & Sam Coleman (pp. 33–65). Oxford: Oxford University Press.

Siewert, C. (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.

Soteriou, M. (2020). The Disjunctive Theory of Perception. *Stanford Encyclopaedia of Philosophy* (available at: https://plato.stanford.edu/entries/perception-disjunctive/#WaysFormDisj).

Sperling, G. (1960). The Information Available in Brief Visual Presentations. *Psychological Monographs: General and Applied*, 74, 1–29.

Stein, T., Kaiser, D., Fahrenfort, J. J., & Gaal, S. van. (2021). The Human Visual System Differentially Represents Subjectively and Objectively Invisible Stimuli. *PLOS Biology*, 19(5), e3001241.

Strawson, G. (1994). *Mental Reality*. Cambridge, MA.: The MIT Press.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular Rivalry and Visual Awareness in Human Extrastriate Cortex. *Neuron*, 21(4), 753–759.

Vugt, B. van, Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., & Roelfsema, P. R. (2018). The Threshold for Conscious Report: Signal Loss and Response Bias in Visual and Frontal Cortex. *Science*, 360(6388), 537–542.

Zięba, P. J. (2022). Seeing Colours Unconsciously. *Synthese*, 200(3), 260, 1–36.