8

The Real Trouble with Armchair Arguments against Phenomenal Externalism

Adam Pautz

> *The intrinsicness of phenomenology is self-evident to reflective introspection.*
>
> —Terry Horgan and John Tienson

> *Every argument has its intuitive bedrock.*
>
> —John Hawthorne

According to *reductive externalist theories* of phenomenal consciousness, the sensible qualities (colours, sounds, tastes, smells) reduce to physical properties out there in the external environment, contrary to the seventeenth-century Galilean view that they are somehow only in the mind-brain. Further, if we are to be conscious of these external qualities, we must be appropriately physically connected to them (e.g., our sensory systems must causally detect or 'track' them). This leads to *phenomenal externalism*: intrinsically identical subjects can differ phenomenally due to external differences. Examples of reductive externalism include 'tracking intentionalism' (Dretske, Lycan, Tye), active externalism (Noë and O'Regan) and reductive versions of naive realism (Fish). Many think such externalist theories represent our best shot at reductively explaining phenomenal consciousness.[1]

The standard arguments against such theories invoke armchair intuitions about far-out cases such as brains in vats, swampmen and inverts (Block, Chalmers, Hawthorne, Levine, Shoemaker). They often presuppose *phenomenal internalism*: phenomenology supervenes on intrinsic character. They play a crucial role in the burgeoning *phenomenal intentionality program* (Horgan, Tienson, Kriegel). A central plank in this program is that reductive externalism fails because armchair reflection establishes phenomenal internalism. Indeed, phenomenal intentionality

theorists not only reject reductive externalist theories; they often reject all reductive approaches, adopting instead what I call a *consciousness first* approach: phenomenal consciousness is not something that must be reductively explained in other terms (e.g., tracking plus cognitive/rational accessibility) but rather a starting point from which to explain other things (e.g., cognition, rationality, value).

My sympathies lie with phenomenal internalism and the phenomenal intentionality program.[2] In particular, I defend an internalist, neo-Galilean (or 'Edenic') version of 'intentionalism' (Pautz 2006; Chalmers 2006). But here my aim is negative. I criticize *three* main armchair arguments against rival reductive externalist theories. Externalists like Dretske, Lycan and Tye have raised their own objections to such arguments. But I show that they do not quite hit the nail on the head; I identify what I take to be the real problems, arguing that the much-discussed armchair arguments are in fact without merit. The moral: the case for phenomenal internalism must depend on *empirical* arguments.

1 Preliminary: why take phenomenal externalism seriously?

Before considering arguments against reductive externalist theories of consciousness, I want to briefly explain why such theories should be taken seriously in the first place. I develop what I consider the best argument. On the face of it, the best explanation for the perception of spatial features invokes interactions with things in external space.

Consider an example, which I use throughout this chapter. Let *R* be the phenomenal property you have when you view a tomato on a certain occasion. *R* is essentially *externally directed*; it necessarily exhibits a *rich spatial intentionality*. Necessarily, if you have *R*, then – even if you are hallucinating – you ostensibly experience phenomenal redness and roundness as bound together at a certain viewer-relative place. So you are in a state that 'matches the world' only if a round object is present there. Call this *the spatial datum* about *R*.

Now, against nominalism, I assume the existence of *properties*. Then the spatial datum implies that having *R* entails bearing a certain *relation* to the spatial property *being round*: roughly, the relation *x is in a state that matches the world only if property y is instantiated*. Call this relation the *phenomenal representation relation*. So if you have *R* while hallucinating, then the general *property* of being round exists, and you bear the phenomenal representation relation to it, even if you do not see an existing *object* that instantiates this property.

In my view, the best argument for externalist theories of phenomenal consciousness is that they might provide the *best explanation* of spatial experiential intentionality. I focus throughout this essay on *tracking intentionalism* (Dretske, Tye, others).

Tracking intentionalists accept *strong intentionalism*: phenomenology is *fully* constituted by the phenomenal representation of complexes of properties. Then they explain phenomenal representation in two steps. First, against internalists who often locate the sensible qualities in the brain (and so have trouble explaining the spatial datum), they maintain that sensible properties are *really instantiated in external space*. For instance, the apparent redness of the tomato (redness-as-we-see-it) is a 'light-reflectance property' of the tomato's surface. Second, you represent such properties because under optimal conditions your neural states suitably causally co-vary with ('track') their external instantiation and in turn lead to appropriate behaviour in space (e.g., behaviour appropriate to a *round thing at p*). For instance, in having *R*, you represent a certain phenomenal colour and shape as co-instantiated out there in space, by virtue of the fact that your neural state has the function of normally tracking their co-instantiation in space. You can represent these properties while hallucinating, because your brain state retains the function of tracking them. So the phenomenal representation relation *just is* a tracking relation – a kind of thermometer model of consciousness. How *else* might we explain the representation of qualities in space?[3]

The basic idea comes in different versions. Dretske (1995) reduces the phenomenal representation relation to this relation: *x* is in a state that satisfies a certain cognitive-rational access condition and that has the systemic biological function of tracking property *y*. The relevant cognitive access condition is what is supposed to turn mere subpersonal representation into genuine conscious, 'phenomenal' representation (it will not play a role here). The relevant notion of biological function is explained in historical-evolutionary terms (more on this in Section 6). Recently, Tye (2012) converted to a similar historical theory.

Tracking intentionalism is externalist. Phenomenology isn't fixed by subjects' intrinsic properties but by what external properties their sensory systems track. To illustrate, consider an *accidental, lone, lifelong* brain in a vat (*a bad-off BIV*) that is an intrinsic duplicate of your brain as you see a tomato. It lacks an evolutionary history. Its brain states, unlike your brain states, *lack* the function of tracking any external shapes (etc.) in any population. So on tracking intentionalism, it cannot represent *roundness*. But given the spatial datum, *R* is inseparable from representing roundness. So it cannot have *R*.

Tracking intentionalism is also reductive. Reduction is attractive. Everyone accepts the following *dependence claim: total (narrow and wide) physical duplicates must (as a matter of at least nomological necessity) bear the phenomenal representation relation to all of the same properties (shapes, orientations, phenomenal colours)*. Because it is reductive, tracking intentionalism nicely *explains* this as follows: (i) *total* physical duplicates trivially bear the *tracking relation* to the same properties (since it is a physical relation), and (ii) the phenomenal representation relation *just is* the tracking relation. This explanation for phenomenal-physical dependence bottoms out in a phenomenal-physical *identity* ('real definition'). This is appealing, because identities do not 'cry out for' further explanation. Indeed, identities are *explanation stoppers*: they don't admit of further explanation. What would it be to explain any identity?

Turn now to *phenomenal internalism*. I favour it in the end, but I admit it faces puzzles. For one thing, I think it requires a radically *non-reductive* account of experiential intentionality. Phenomenal internalists (including phenomenal intentionalists such as Horgan and Tienson) have overlooked the point. True, some phenomenal internalists (Block, Kriegel, McLaughlin, Prinz) incline towards reductive type-type neural identity theory, holding that monadic phenomenal properties such as *R* are identical with monadic internal neurofunctional properties.[4] But as Field (2001, 69–72) has argued, philosophers cannot rest content with reducing *monadic* mental properties to *monadic* neural or functional properties: they must also say something about *dyadic* mental *relations* between subjects and external items, such as the phenomenal representation relation. Tracking intentionalists reduce it to an externally determined *tracking relation*. Can phenomenal internalists reduce it?

The following argument (Pautz 2010b, §7; Tye [forthcoming a]) suggests not: phenomenal internalists (even if they accept type-type identity for *monadic* phenomenal properties) must accept primitivism about the *dyadic phenomenal representation relation*:

1. Given phenomenal internalism, the aforementioned bad-off BIV *does* have tomato-like experience *R*.
2. Having *R* is inseparable from having an experience of a *round* thing, an experience that 'matches the world' only if something is present that is *round*. (Spatial datum.)
3. So in having *R*, the BIV bears the dyadic phenomenal representation relation to the property *being round* – even if *R* is intrinsic, it entails standing in this relation to a particular shape (Horgan et al. 2004,

304–305). (As we saw, this follows from the spatial datum and the existence of properties.)

4. But, *ex hypothesi*, the BIV bears no *dyadic physical-functional* (e.g., *tracking*) relation to this property.
5. *Conclusion*: phenomenal internalism implies that the phenomenal representation relation is a *non-physical* and (presumably) *primitive* relation between individuals and *being round* and other sensible properties.

I think this argument is a threat to reductive materialism no less serious than the standard 'explanatory gap' arguments.

The case for premise 4 is simply that, while the BIV is conscious of roundness and other properties, it bears no actual or dispositional *physical* relations uniquely to those properties.[5] So phenomenal internalists must apparently say that the phenomenal representation relation is a non-physical relation. Phenomenal internalism goes naturally with *non-reductive, internalist intentionalism* (Chalmers 2006; Pautz 2006) in direct opposition to tracking intentionalism.

Some phenomenal internalists object to the move from premise 2 to 3. For instance, since they reject abstract objects, Kriegel (2011) and Mendelovici (2010) deny that having *R* essentially involves standing in any real *relation* to the general property, *roundness*, understood as a necessarily existing abstract object. If there is no such relation, internalists don't have to worry about reducing it. My reply is that the argument doesn't require that *R* is *essentially* relational.[6] For even if *R* is not *essentially* relational, having *R* certainly at least *contingently implies* standing in certain mental relations to properties (or tropes) and concrete types of things *in scenarios where those properties and types exist*. For instance, suppose that in the BIV scenario there happens to be a round tomato in front of the BIV as it undergoes its tomato-like hallucination. Then, even if the property of being round exists only in scenarios where it is instantiated (an 'immanent' conception of properties), the property *being round* exists in the scenario, and in having *R*, the BIV stands in the phenomenal representation relation to it. Further, the BIV bears the following relation to the concrete tomato that exists in the scenario *x has an experience that is accurate with respect to y*. Since the BIV bears no physical-functional relations to such things, phenomenal internalism apparently entails that these mind-world relations are primitive.[7]

The main drawback of primitivism is that it cannot provide any very appealing explanation of the aforementioned dependence claim: *total*

(narrow and wide) physical duplicates bear the phenomenal representation relation to all of the same properties (shapes, orientations, colours) and types of concrete things.

One option for internalist primitivists is a *dualist explanation*: there is a swarm of basic, contingent *phenomenal-physical laws* linking being in certain physical states (e.g., neural states) with bearing the primitive phenomenal representation relation to certain shapes, phenomenal colours, positions and so on. They would be 'nomological danglers', additional to the basic laws of physics. This is unappealing. The tracking intentionalist's explanation bottoms out an identity, which is an 'explanation stopper'. By contrast, the dualist's laws cry out for further explanation. *Why* do these laws obtain? We must admit *some* such basic modal truths, but we should keep them to a minimum.

Another option for internalist primitivists would be an *emergent materialist* explanation of phenomenal-physical dependence (Rosen 2010, 132). I argued that internalism implies that the phenomenal representation relation is *primitive* and *distinct* from the physical, *in the sense that* there is no reduction (real definition, metaphysical analysis) of it in physical-functional terms (no general completion of schema *what it is for x to stand in the phenomenal representation relation to y just is for x to bear physical-functional relation R to y*). *Emergent materialists* nevertheless claim that this primitive relation is always *grounded in* (instantiated *by virtue of*) the physical and hence depends on the physical with *metaphysical necessity*. (Grounding has received a great deal of contemporary interest; see Rosen 2010.) The only difference with dualism is modal: on dualism, the dependence is *contingent*. So on emergent materialism, there is a huge swarm of *phenomenal-physical 'grounding laws'* of this form: being in internal neural state *N* grounds bearing the primitive phenomenal representation relation to *X*. All these phenomenal-physical grounding laws are basic. I don't just mean that they are deeply a posteriori rather than a priori (most materialists would not object to a posteriori necessities). I mean they are *metaphysically* basic 'grounding danglers': they don't follow from any more basic truths. Emergent materialism, then, is analogous to a Moorean metaethical view on which *goodness* is *primitive* but always *grounded in* natural properties.

The innumerable grounding danglers of emergent materialism are just as unappealing as the nomological danglers of dualism. They cry out for explanation. If the phenomenal representation relation is *distinct from* all physical-functional relations, *why* is it *necessarily* connected with (and grounded in) the physical-functional facts? (Analogy: if an emergentist said that we are simple souls yet somehow necessarily grounded

in wholly distinct bodies, we would be mystified.) Just as many used to say 'supervenience' cries out for explanation, I suggest grounding – its contemporary replacement – cries out for explanation as well. Reductive explanations are better.

We phenomenal internalists face another problem. It appears totally *arbitrary* that a mere pattern of neural firing *B* should result in the phenomenal representation of *roundness* rather than some other shape, even in the BIV scenario in which it is not causally linked to any shape at all. (Also, why should it result in the experience of phenomenal redness *at this particular position* in the visual field?) Phenomenal externalists, such as tracking intentionalists, might avoid this arbitrariness (even if they cannot avoid the explanatory gap). On their externalist view, the physical ground of phenomenally representing *roundness* is not the mere neural state *B* but a wider, environment-involving state that is specified in terms of that very spatial property: the state of having some internal state or other that has the function of tracking *roundness* and causing behaviour appropriate to *round* things.

So phenomenal externalists *might* nicely explain the spatial intentionality of perceptual experience. By contrast, those of us favouring phenomenal internalism must grapple with some serious puzzles. Therefore our topic is important: can simple armchair arguments undermine reductive externalism and establish phenomenal internalism? My answer is no.

2 The argument from the internalist intuition

Many philosophers (Block, Chalmers, Burge, Hawthorne, Kriegel, Levine and Speaks) espouse the first argument I examine against reductive externalist theories. It has two steps.

Roughly, an *intrinsic property* of an individual is one whose instantiation by that individual does not constitutively depend on contingent items wholly distinct from that individual. We have empirically defeasible, pre-theoretical justification for believing *some* properties to be intrinsic. Consider, for instance, *shapes*. The first step of the argument claims that we likewise have a strong pre-theoretical justification for believing that *phenomenal properties* such as *R* and *having a headache* are intrinsic properties of *subjects*, so that

> *Phenomenal internalism*: all experiences supervene on subjects' intrinsic properties; total intrinsic duplicates of you must (as a matter of *metaphysical* necessity) be phenomenal duplicates of you.

This is the *internalist intuition*. Hawthorne (2004, 352) calls it *intuitive bedrock*, and Horgan and Tienson (2002, n. 23) call a similar claim 'self-evident'.[8] How could individuals differ concerning whether they have a headache or see red, unless they differ intrinsically? [9]

The second step uses thought experiments to show that certain reductive theories of consciousness violate phenomenal internalism. Consider again 'tracking intentionalism' (Section 1).

1. *The bad-off BIV*: We already saw that tracking intentionalism entails that the bad-off BIV cannot support phenomenal consciousness. This violates phenomenal internalism, since it is an intrinsic duplicate of your brain.

2. *Swampman*: Harold is an ordinary person. A total intrinsic duplicate of Harold materializes by chance in a swamp. This system, *Swamp Harold*, has no evolutionary history. As noted in Section 1, Dretske's and Tye's versions of tracking intentionalism entail that only systems with a selection history can represent the world. Since standard forms of phenomenology are inseparable from intentionality (e.g., standard visual experiences necessarily exhibit spatial intentionality), it follows that although Harold and Swamp Harold are total intrinsic duplicates, they differ phenomenally: in particular, Swamp Harold simply has no interesting experiences at all, contrary to phenomenal internalism.

3. *Inverted Earth*: On Earth, the sky is blue. When Harold looks at it, he gets receptor activity *A* on his retina and downstream neural state *S*. On Earth, among humans, *S* has the biological function of tracking external blueness.

Suppose that, on Inverted Earth, there evolved a species intrinsically identical to Homo sapiens. But objects there have inverted colours; for instance, the sky is yellow rather than blue. However, the ambient light is weird and has always been weird throughout the evolutionary process, so that yellow objects give off 'blue' light. So when Twin Harold looks at the *yellow* sky on inverted earth, he gets receptor activity *A* on his retina and downstream neural state *S*, the same neural state Harold gets when he looks at the *blue* sky on earth. (In another version of the case, twin humans naturally evolved with colour-inverting lenses in front of their eyes, so that yellow light is transformed into blue light.) Indeed, on viewing the sky, Twin Harold is a complete intrinsic duplicate of Harold on earth. But whereas in Harold *S* has the biological function of tracking *surface blue*, in Twin Harold it has the biological function of tracking *surface yellow*.[10]

On tracking intentionalism, this means that even though Harold and Twin Harold occupy the same total internal state, this internal state

enables them to phenomenally represent different external colours. Again, tracking intentionalism violates the internalist intuition.

3 Problem: no armchair support for phenomenal internalism

I think that the argument from the internalist intuition is a non-starter. However, before explaining why, I criticize standard externalist objections.

Externalists have been concessive. Dretske (1995, 151) concedes that phenomenal internalism seems 'obvious' and 'powerful'. Tye (2000, 120) admits that rejecting phenomenal internalism is 'deeply counterintuitive'. Their objections lie elsewhere. But I think armchair enthusiasts might offer somewhat convincing replies to those objections.

1. Dretske's main objection involves a *rebutting defeater* (1995, 151). Apparently, the best reductive materialist theories of consciousness and its intentionality violate phenomenal internalism, as I admitted myself (Section 1). So there is a strong theoretical argument against it. Maybe this beats the intuitive argument for it.

Possible reply: Maybe phenomenal internalism is compatible with reductive materialism after all; maybe my Section 1 argument fails. If not, perhaps the internalist intuition is so compelling that we must accept it despite its problematic consequences.

2. Dretske (1995, 148) and Tye (2012) produce *undercutting defeaters*. Offhand, you might have thought that every intrinsic duplicate of a heart is a heart or that every intrinsic duplicate of a gas gauge is a gas gauge. These intuitions are wrong, because the relevant properties depend on historical, extrinsic factors. Given the bad track record of internalist intuitions, maybe we are equally wrong about the intrinsicness of phenomenal character.

Possible Reply: Even if our initial, unreflective *opinions* about the intrinsicness of some quite different properties are wrong (as we see ourselves realize after a moment's thought), this provides little reason to doubt our persisting *intuition* about the intrinsicness of *phenomenal character*.

3. Lycan (2001, 24) says using the internalist intuition 'simply begs the question of [phenomenal] externalism in favor of internalism'.

Possible reply: Hawthorne (2004, n. 4) notes that this is a confused use of 'begging the question': it is not question begging to use pre-theoretical intuitions against a theory just because they conflict with that theory.

So I think externalists' objections to the argument from the internalist intuition fail. What then is the real problem?

Externalists like Tye and Dretske have been much too concessive in granting that phenomenal internalism enjoys pre-theoretical support at all. The real problem is that this is not so. A little reflection, indeed the whole history of human thought on perception, shows that phenomenal externalism is not absurd at all; indeed, if anything, it is pre-theoretically quite plausible.

For instance, many have accepted *naive realism*. On this view, the sensible qualities, or 'qualia', are really out in the world: colours, sound qualities, tastes and so on. For instance, when you view a tomato and have phenomenal property *R*, the red quality you are directly acquainted with is really 'spread out' on its surface. The distinctive claim of naive realism is that, at least in 'veridical' cases, *some* ordinary (non-burry, etc.) phenomenal properties such as *R* are grounded in nothing but standing in a relation of direct acquaintance to a concrete state, or condition, involving the instantiation of sensible properties (colours, shapes) by a mind-independent physical object. (These sensible properties include viewpoint-relative properties such as *being elliptical from here* – 'objective looks'.)

To handle hallucination, naive realists might appeal to non-normal objects such as sense data or Meinongian objects, so that *R* is always grounded in acquaintance with objects. Alternatively, they might accept a more extreme 'disjunctivism' (I think the best version is 'primitivist disjunctivism', discussed in Pautz 2010a, 275). So hallucination doesn't undermine naive realism.

Naive realism is an example of a *relational (act-object) theory of phenomenology*: *some* phenomenal properties are, in some cases, grounded in direct acquaintance with concrete objects and states wholly distinct from perceivers.

Naive realists have taken different views on the physical basis of acquaintance with the world. Pre-modern thinkers such as Plato, Euclid and Ptolemy accepted the *extromission theory*: we become acquainted with the world by way of rays *emanating from* the eye (perhaps with infinite velocity, like gravity in Newton's theory). Thus in his *Optika* Euclid wrote:

> Rays [proceed] from the eye [and] those things are seen upon which the visual rays fall and those things are not seen upon which the visual rays do not fall.

This entails phenomenal externalism. The phenomenal difference between your seeing a round thing and your seeing a square thing is not

an intrinsic difference in your head or soul. Indeed, your head might be exactly the same in both cases. The phenomenal difference is an extrinsic, relational difference in what shapes you are acquainted with out in the world via the ray; it is entirely 'outside the head'. And it is this merely relational difference that causally explains your different beliefs and behaviour in the two cases.

Today we accept intromission theory: vision is a causal process leading via light *from the object to the brain*. This might encourage acceptance of phenomenal internalism (see the 'simple empirical argument' below). But contemporary naive realists (e.g., Fish 2009, 137) combine it with naive realism. Their idea is that the long causal process going from external states to appropriate neural processing is the *supervenience base* of the mind's 'reaching out' and getting directly acquainted with those external states.

This view is also externalist, entailing that *intrinsic duplicates can differ phenomenally*. To see this, consider again Harold on Earth where the sky is blue and Twin Harold on Inverted Earth where the sky is yellow (Section 2). Their neural processes as they view the sky are intrinsically identical. Nevertheless, on naive realism, they enable Harold and Twin Harold to be acquainted with different external concrete colour states ('tropes') involving the sky, because they are appropriately caused by those different external colours.

Traditional sense datum theory is another relational theory of phenomenology. Indeed, I argue that if it is generally considered a paradigmatic version of phenomenal internalism, *it is another version of phenomenal externalism*.

By 'the sense datum theory' I mean the view that phenomenal properties are grounded in an acquaintance relation to a concrete state involving a sense datum *wholly distinct from the subject*, where a *sense datum* is an object generally having the properties external things appear to have. On a standard elaboration, there are contingent psychophysical laws whereby the subject's brain states cause certain sense data to come into existence for a short period of time and simultaneously cause the subject to stand in the acquaintance relation to those sense data.

I suspect many take sense datum theory to be internalist because they mistakenly think sense data are *parts* of subjects (souls or brains), so that differences in sense data would indeed be intrinsic differences. This is not the view I have in mind. When in hallucination you are acquainted with a literally round sense datum, that round thing is of course not *part of* your brain. And if you are a simple soul without any parts rather than a brain, then again the sense datum is not part of you. Then where are

sense data? On one version, sense data occupy a separate, private two-dimensional mental space. Even on this version they are *wholly distinct from* the subject (soul or brain) that observes them. On the different version defended by Price (1954, vii–viii) and Jackson (1977, 102), they are *three-dimensional objects in public physical space* alongside physical objects. Think of them as projections of the brain. On this view, even though a sense datum exists in *public space*, only the subject of the brain that causes it to come into existence there can be acquainted with it. So the sense datum theory is a relational theory like naive realism, with sense data as simulacra for physical objects.

Since sense data are wholly distinct from subjects, this theory is also externalist. Suppose as before that Harold has an 'intrinsic duplicate', Twin Harold, on inverted earth. Exactly what that means depends on the correct ontology of the human subject. If Harold and Twin Harold are physical things like brains or bodies (even if they bear non-physical acquaintance relations to non-physical sense data), then they are intrinsic duplicates in that these brains or bodies share all of their intrinsic properties. Alternatively, if they are simple, non-physical souls, they are intrinsic duplicates because these souls have exactly the same intrinsic properties, like *being happy* or *thinking*.

Either way, *just like naive realism and tracking intentionalism*, the sense datum theory implies that *Harold and Twin Harold can differ phenomenally*, contrary to phenomenal internalism. For suppose that Harold and Twin Harold live under different ('inverted') psychophysical laws connecting brain states with acquaintance with sense data, so that while Harold is acquainted with a blue sense datum on looking at the sky, Twin Harold is acquainted with a yellow one. Alternatively, suppose that they live under the same psychophysical laws but these laws are *probabilistic*, so that even though Harold and Twin Harold undergo an intrinsically identical brain state, this same brain state happens to cause them to be acquainted with these qualitatively different sense data.

Then although Harold and Twin Harold are *intrinsic duplicates*, they differ phenomenally. On the sense datum theory, just as on naive realism and tracking intentionalism, the phenomenal difference between these two brains or souls is *not* an intrinsic difference. Instead, it is a purely *relational, extrinsic difference*: they bear the acquaintance relation to different sense data wholly distinct from them. It is like the difference between sitting next to Mary and sitting next to Jane. Indeed, the *whole point* of the sense datum theory is that phenomenal differences are relational differences, contrary to rival 'non-relational' views such as 'adverbialism'. On the sense datum theory, this purely relational difference

between Harold and Twin Harold results in their having different perceptual and introspective beliefs and perhaps different behavioural dispositions. Hence, the sense datum theory is a version of phenomenal externalism no less than tracking intentionalism and naive realism. Experience isn't fixed by how you intrinsically are; it's fixed by your relations to things distinct from you.

Or again, suppose you have an experience of a tomato and consider a soul or a brain – call it BIV – that is an intrinsic duplicate of you in a vat. If phenomenal internalism is 'intuitive bedrock' (Hawthorne) or 'self-evident' (Horgan and Tienson), then BIV must have the very same experience. Sense datum theorists disagree. To see this, suppose that for some reason this brain cannot produce sense data according to the usual psychophysical laws. H. H. Price noted in a famous passage (1954, 3) that, on the sense datum theory, having a tomato-like experience (even in hallucination) essentially requires *accompaniment*: it requires 'that there exists a red patch of a round and somewhat bulgy shape' distinct from oneself and that this patch is present to one's consciousness. Price said *this* is 'self-evident'. If this is right, then BIV simply cannot have a tomato-like experience, because BIV is unaccompanied by a suitable sensible object. So no less than naive realism and tracking intentionalism, sense datum theory is externalist, entailing that an intrinsic BIV duplicate of a conscious subject might fail to be a phenomenal duplicate. Against internalism, awareness of sense data is a relational affair that doesn't supervene on a subject's intrinsic state with 'metaphysical' necessity (contingent psychophysical laws secure only 'nomic' supervenience).

Now for my main point. If phenomenal internalism were really self-evident or intuitive bedrock, we should be able to immediately rule out the basic relational conception of visual phenomenology (including extromission theory, naive realism and sense datum theory) *simply because it is externalist*. But we cannot. For instance, pre-theoretically, it is simply *not* counterintuitive that different visual experiences should be grounded in a purely extrinsic, relational difference involving what the subjects are acquainted with. So the argument from the internalist intuition against contemporary reductive externalist theories (e.g., tracking intentionalism and active externalism) is a non-starter. If we cannot rule out relational theories *just because they violate phenomenal internalism*, we also cannot rule out reductive externalist theories for this reason.

So my criticism of the argument from the internalist intuition is that pre-theoretical reflection fails to support for phenomenal internalism.

Although my criticism does not require it, I also believe something stronger: given what experience is like, pre-theoretical reflection supports *rejecting* phenomenal internalism and *accepting* phenomenal externalism (even though ultimately I accept internalism). The armchair enthusiasts have it backwards.

True, maybe *having a headache* seems like an intrinsic, non-relational property of oneself. It was such sensations that motivated 'adverbialism'. (Lycan [2001, 28] notes that even reductive externalists might accommodate the intuition that *some* phenomenal properties involving one's own body, like *having a headache*, are indeed intrinsic.) But suppose you look at a tomato and have visual phenomenal property *R* on a particular occasion. Bracket all of your detailed empirical knowledge about the role of the brain, the possibility of hallucination, the case for materialism and so on. What theory would you accept if you had only phenomenology to go on? Intuitively, the basic *relational theory* about *R* is correct. Intuitively, there exists a red and round object *wholly distinct from you*, and your having the phenomenal property *R* is grounded in your bearing an acquaintance relation to the concrete state of this object's being red and round rather than any intrinsic property. The character of your conscious state is grounded in the object's possessing these perceptible properties, together with the fact that you stand in a relation of acquaintance or direct awareness to this state.

Likewise, if you see a balloon changing shape while deflating, it is not *pre-theoretically* plausible that the phenomenal change in your experience is an *intrinsic* change inside of you (even if empirical investigation has shown it is *accompanied by* one), as the phenomenal internalists insist. Rather, what is *pre-theoretically* plausible is that it consists in your being acquainted with the concrete instantiation of a new shape by an object distinct from you (akin to the difference between sitting next to Mary and sitting next to Jane).

Indeed, scores of philosophers (sense datum theorists and naive realists) share the same basic relationality intuition for visual experience (Fish 2009, 20). And as we have seen, the relational theory entails phenomenal externalism, because intrinsic duplicates, occupying different environments or operating under different psychophysical laws, could be acquainted with different (mental or physical) items.

The armchair enthusiasts might offer a response to my more modest point that armchair reflection at least *fails to support* phenomenal internalism. Hawthorne (2004, 352) and Horgan, Tienson and Graham (2004, 302) emphasize that the internalist conviction is 'compelling',

'widespread' and 'persistent'. Maybe the best explanation is that we do in fact have some pre-theoretical justification for accepting phenomenal internalism in general, contrary to what I have suggested.

But this response fails. First, are most ordinary people *really* inclined to accept phenomenal internalism? This is an empirical question – a question for 'experimental philosophy' – but I doubt it. As I have just said, throughout history many have been inclined to instead accept externalist naive realism. And recent studies (Winer et al. 2002) suggest that even today many accept the primitive *extromission theory of vision*, which (as we saw) entails phenomenal externalism.

Second, maybe some other people, including philosophers and scientists, are inclined to accept phenomenal internalism. But the reason cannot be that it is pre-theoretically self-evident or intuitive bedrock; I have shown it is not. I offer an alternative explanation: the *real* source of their internalist conviction is not pre-theoretical reflection but what I will call the *simple empirical argument*, which has a long history (Russell 1927, 320ff.). Here's a recent statement:

> It is only inner states that matter for experience [phenomenal internalism], not anything relational. [Phenomenal externalism] flies in the face of the scientific evidence correlating experiences with neural responses: for every measurable change in experience, there is some measurable change in the nervous system. (Prinz 2012, 19)

Likewise, Kriegel declares that 'everything we know about the laws of neurophysiology suggest that a lifelong envatted brain with the same sensory-stimulation history of my brain would undergo the same experiential life as mine' (2011, 137). And Horgan and Tienson note that 'distal environmental causes generate experiential effects only by generating more immediate [neural] links in the causal chains between themselves and experience'. The simple empirical argument concludes that it is *metaphysically impossible* to make changes in experience except by making changes in intrinsic neural properties.

So in view of what our experience of the world is like, what is *pre-theoretically* plausible, before we learn any science, are externalist theories, such as naive realism: at least for visual experience, phenomenal differences *do not* necessarily require intrinsic differences inside the head. Granted, today many philosophers (Kriegel, Horgan, Tienson, Hawthorne) vehemently reject phenomenal externalism and find phenomenal internalism 'obvious' across the board. But the explanation, I conjecture, is that they have for most of their lives known the basic

scientific facts about the role of the brain in enabling conscious experience. Because of the seductive simple empirical argument, they have become totally convinced of phenomenal internalism: phenomenal differences do require intrinsic differences inside the head. Because their confident belief in phenomenal internalism has become so ingrained, they mistakenly take it to be something that is obvious or self-evident on a moment's reflection. But really it is just a high-level empirical belief, one that became widely accepted in the history of human thought only after detailed empirical investigation.

Now the phenomenal internalist might naturally respond, 'OK, the armchair argument from the internalist *intuition* fails, but why can't I just directly rely on *the simple empirical argument* to undermine reductive externalist theories like tracking intentionalism and naive realism?'

My focus here is on armchair arguments, but let me address this question. I favour certain empirical arguments (see Section 8), but I think that this very simple empirical argument fails. The quick way to see this is to note that it is equally true that 'for every change in *thoughts about natural kinds*, there is a measurable change in the brain' (to appropriate Prinz's language). But this does not entail that natural kind thoughts are fixed by intrinsic neutral properties: content externalism means this is not the case. The inference is equally fallacious concerning phenomenal states.[11]

The fallacy is obvious: the mere fact that it is *nomically necessary* in actual humans that phenomenal differences are correlated with intrinsic neural differences doesn't mean that this is *metaphysically necessary*. What Kriegel calls 'the laws of neurophysiology' might (like typical special science laws) obtain only relative to a background condition, one not satisfied in the case of the BIV. (For instance, a brain state might result in an experience of round *only if* it normally tracks round objects.) Indeed, Prinz is wrong that the simple correlational data even *raise the probability* of phenomenal internalism over phenomenal externalism, since all phenomenal changes are correlated with *both* changes in intrinsic neural states *and* changes in externally-determined content (e.g., when you go from seeing yellow to seeing blue, you go from a neural state that normally tracks yellow objects to a neural state that normally tracks blue objects). So the simple correlational data alone are entirely neutral between phenomenal internalism and phenomenal externalism (naive realism, tracking intentionalism, active externalism).

4 The argument from possibility intuitions

So the argument from the internalist intuition fails. But perhaps all is not lost for the armchair enthusiasts. Another argument is available: *the argument from possibility intuitions*. The argument specifically targets *reductive materialist* externalist theories, like tracking intentionalism. (It doesn't work against *dualist* externalist theories; see note 13.) Chalmers, Loar, Shoemaker and Levine suggest the argument but do not clearly distinguish it from the argument from the internalist intuition.[12] So let me explain the difference.

Again, consider tracking intentionalism. On tracking intentionalism, having *R* entails the obtaining *of a certain wide (non-intrinsic) physical condition*: having a state that under biologically normal conditions tracks – and thereby represents – the instantiation of redness (on this view, a reflectance property) and roundness in the external world. Having the experience in the absence of the wide physical condition is metaphysically impossible.

Therefore, to refute such a reductive externalist theory, it would be enough to establish from the armchair the mere *possibility* of having *R* in the absence of the relevant wide physical condition. For instance, it would be enough to show that in *some* possible world a BIV intrinsic duplicate of oneself has *R*. It would also be enough to show that certain *spectrum inversion scenarios* are possible (more on this below). The argument from possibility intuitions merely relies on such possibility claims. Thus it differs from the argument from the internalist intuition, which by contrast relies on a much stronger *necessitation* claim to the effect that *every* possible intrinsic duplicate of oneself (e.g., every possible brain in a vat duplicate) in *every* possible world is a phenomenal duplicate of oneself.[13]

For example:

> The intuition [that a BIV with experiences is *possible*] supports the view that my [experiences] are constituted independently of my actual situation in the world. (Loar 2003, 230)

> Focusing on … color, I say 'THIS is supposed to be a reflectance property of the surface of … a cloud of fundamental particles' … . Reflection on the disparity between the manifest and the scientific image makes inescapable the conclusion that the phenomenal character we are confronted with in color experience is due not simply to what there is in our environment … it seems intelligible [possible] that there are creatures who, in any given objective situation, are confronted with a

> very different phenomenal character than we would be in that same situation. (Shoemaker 1994, 293–294)
>
> It seems intuitively plausible that states with different qualitative character could nevertheless represent [track] the very same distal feature. (Levine 1997, 109)

On tracking intentionalism, having *R* consists in being in a state that normally tracks the colour red, which is identical with a certain surface reflectance *F*. But, Shoemaker notes, there is an *explanatory gap*. Why should tracking this reflectance *F* constitute a reddish experience as opposed to (say) a greenish experience? Therefore, against tracking intentionalism, it is intuitively possible that two individuals should normally track *F* and yet be spectrum inverted: while one has a reddish experience, the other has (say) a greenish experience. So tracking intentionalism is false.

In general, intuitively, phenomenology is modally independent of wide physical conditions, contrary to reductive externalism.

5 Problem: the argument is unavailable to materialists

My objection to the argument from the 'internalist intuition' was simply that we lack *pre-theoretical* justification for accepting phenomenal internalism. By contrast, I grant that, contrary to reductive externalist theories, it *is* intuitive that technicolor phenomenology is modally independent of wide *physical* conditions, such as *tracking a particular reflectance property* (as opposed to wide *non-physical* conditions, such as standing in a primitive acquaintance relation to a primitive external colour). This is just an instance of our more general antimaterialist intuitions. So the relevant scenarios are *conceivable*. They cannot be ruled out a priori.

So what's wrong with the argument from possibility intuitions? Tye's response is that conceivability does not *entail* possibility.[14] But this is not a strong criticism, because conceivability nevertheless provides *some* defeasible evidence for possibility, hence against reductive externalist (e.g., tracking) theories.

I think that the real problem is that most philosophers are *materialists*, including Loar, Shoemaker and Levine. And materialists cannot consistently invoke possibility intuitions against reductive externalist theories.

There are only two possible forms of materialism: *internalist materialism* (type-type identity theory, internalist functionalism) and *externalist*

materialism (tracking intentionalism, active externalism). Our possibility intuitions count equally against *both*, since we have general *antimaterialist intuitions* to the effect that experience is modally independent of *all* physical conditions (internal and external). Call this the *parity problem*.

To illustrate, consider Shoemaker. Shoemaker notes the explanatory gap between tracking a reflectance *F* and having a reddish experience. The connection looks *contingent*. So it is intuitively possible that tracking the reflectance *F* could be associated with having a greenish experience rather than a reddish experience, as in spectrum inversion.

But it is strange that Shoemaker uses this argument against the externalist materialism of Dretske and Tye. Equally robust possibility intuitions would undermine Shoemaker's *own internalist materialism*. For on Shoemaker's internalist materialism, having a reddish experience is constituted by some neural-functional state *N* involving soggy grey matter. And the explanatory gap between having *N* and having a reddish experience is *just as wide as* the explanatory gap between tracking reflectance *F* and having a reddish experience. The connection between neural-functional state *N* and the colour experience seems *just as contingent as* the connection between reflectance *F* and the colour experience. Consequently, contrary to Shoemaker's internalist materialism, it is intuitively possible that *N* should be associated with a reddish experience in humans and with a greenish experience in another population: there intuitively could be *spectrum inversion among individuals with the same narrow neural-functional states*, just as there intuitively could be spectrum inversion among individuals with the same wide physical states of the form *normally tracking reflectance F*.

Likewise, as Loar (quoted above) implies, intuitively, a bad-off BIV could have a reddish experience while not tracking reflectance *F*, contrary to reductive externalist theories, such as tracking intentionalism. But it is equally intuitively possible that an individual (say an alien or a robot) should have a reddish experience while lacking neural-functional property *N*, contrary to the internalist materialism Loar himself accepts.

Our possibility intuitions against externalist materialism are not any 'stronger than' our possibility intuitions against internalist materialism. So materialists like Loar, Shoemaker and Levine would need some other considerations or arguments (e.g., the empirical arguments to be mentioned in Section 8) in order to justify accepting our possibility intuitions against externalist materialism while ignoring our equally strong possibility intuitions against their own internalist materialist theories.

But then these other arguments would be doing all the justificatory work.

Materialists cannot use possibility intuitions against externalist materialism for another reason. I call it *the bad lot problem*. Consider an analogy: if you believe that the weather man is wrong in his predictions about *wind conditions* half the time (these predictions form a 'bad lot'), you should put hardly any stock in *any* of his predictions about wind conditions. But the materialist believes that our antimaterialist possibility intuitions about the relationship between the phenomenal and the physical also form a 'bad lot': whatever version of materialism turns out to be true, intuitions in this group must generally be false (e.g., if internalist materialism is true, all contrary possibility intuitions are false). So *if* you accept materialism, you must say that not only do they provide equal justification against internalist materialism and externalist materialism (the parity problem); they are also not to be trusted at all (the bad lot problem).

6 The argument from phenomenal localism

Previously, I criticized the argument from the internalist intuition against reductive externalist theories of consciousness (Section 3). For instance, it is simply not pre-theoretically intuitive that tomato-like experience *R* is *intrinsic*. But, I concede, it *is* pre-theoretically intuitive that your having *R* for a period is *temporally local*: that is, totally modally independent of everything outside *the total state of the universe during that period*. Thus it differs, for instance, from the property of *being a traffic signal that means stop*, whose instantiation now constitutively depends on past conventions (e.g., to stop when the light turns red). More generally,

> *Phenomenal localism*: Necessarily for any phenomenal property *P*, if a subject instantiates *P* for temporal period *p*, and proposition *C* specifies *all* of the intrinsic properties and relations instantiated in *the whole world* during period *p*, then, for any world *W* at which *C* is true for period *p*, the subject also has phenomenal property *P* in *W* during period *p*, no matter what *W* is like before and after period *p*.

Roughly, whereas phenomenal internalism is the claim that having a certain experience for a time supervenes on the intrinsic properties of the *subject alone* during that time, phenomenal localism is the weaker claim that it at least supervenes on the intrinsic properties and relations

instantiated in *the whole universe* during that time. To appreciate the difference, consider naive realism. On this view, the character of your experience is fixed by your standing in a primitive acquaintance relation to a state of the external world, so phenomenal internalism is false. But the holding of that relation at a time might be modally independent of the past and future (ignoring the time-lag argument), so that phenomenal localism is true. Likewise, on the sense datum theory, phenomenal internalism is false (as we saw), but phenomenal localism is true.

Phenomenal localism provides a promising argument against a *subset* of reductive externalist theories; namely, *historical* externalist theories that violate phenomenal localism.

Consider the tracking intentionalism of Dretske and Tye. Suppose you have reddish experience *R* for ten seconds. And suppose tracking intentionalism is true. On tracking intentionalism, you have *R* because you have a brain state, *B*, which has the *biological function* of tracking the red reflectance.

Now consider a world *W* that is intrinsically like the actual world for the ten-second period but in which everything came into existence ex nihilo at the start of the ten-second period (there is no past at all). In this world, since you have no evolutionary history, your brain state does not have the *biological function* of tracking the red reflectance. So according to Dretske and Tye, in *W*, even though the total state of the universe for the ten-second period is intrinsically same as in the actual world, you don't have *R* for that period, because your brain state represents nothing at all.

Or consider a world *Z* in which only our evolutionary history is different in such a way that your current brain state *B* now counts as having the function of tracking the *green reflectance*. (Compare: had only the past been appropriately different, the stoplight turning red in the present might have meant *go* rather than *stop*.) On tracking intentionalism, in *Z*, even though the total state of the universe for the ten-second period might be intrinsically the same, you have a greenish experience rather than a reddish one for that period.

Scores of philosophers, appealing to BIVs and swampmen, argue that such externalist theories are absurd because they violate *phenomenal internalism*. We have seen that this standard argument fails: phenomenal internalism is simply not a self-evident truth. *Many* perfectly coherent theories – extromission theory, naive realism and sense datum theory – violate phenomenal internalism. So I think philosophers shouldn't have focused on phenomenal internalism. Instead they should have focused on *phenomenal localism*. What is truly new and unusual about some

contemporary externalist theories is not that they violate phenomenal internalism but that they also violate phenomenal localism. Some past theories of phenomenal consciousness (sense datum theory, naive realism) violated phenomenal internalism but none violated phenomenal localism.

7 Problems with the argument from phenomenal localism

Nevertheless, I think that even the argument from phenomenal localism fails.

First, it undermines only a subset of reductive externalist theories: namely, those violating phenomenal localism, for instance the tracking intentionalism of Dretske and Tye. Other reductive externalist theories might accommodate phenomenal localism, even if they violate phenomenal internalism.

Here are some examples. (i) While Dretske's and Tye's versions of tracking intentionalism violate phenomenal localism because they appeal to history, maybe other possible versions accommodate phenomenal localism. True, devising such a theory might be difficult – since all standard theories of representation appeal to historical facts or forward-looking facts to help settle what external features our inner states have the 'biological function' of tracking or track under 'optimal conditions' in the present – but maybe not impossible. (ii) Likewise, maybe naive realists could reduce the acquaintance relation to a complex mind-world causal relation. And maybe, contrary to Humeanism about causation, causal facts themselves are *local facts* that do not depend on regularities in the past and future (Hawthorne 2004). The result would be a reductive externalism that accommodates phenomenal locality. (iii) Maybe 'active externalism' and other output-based versions of phenomenal externalism can accommodate phenomenal localism, if the relevant action-oriented facts do not depend on the past or future.

Historical externalists like Dretske and Tye might pursue a less conciliatory response: phenomenal localism is simply false, even if compelling. To soften the blow, they might say the following.

First, we also have locality intuitions about *x causes y* and *thinking about water*. But Hume and Putnam have convinced many that these intuitions are false. This bad track record might undercut *somewhat* our confidence in phenomenal localism.

Second, materialists need a theory of how we might be justified a priori in believing that a property is local and a theory of how such

beliefs might be generally reliable. But that is hard to come by. (A materialist cannot comfortably accept 'revelation': that we 'immediately grasp' the full essential nature of phenomenal properties just by being acquainted with them and can tell that those essential natures don't involve the past or future.) Absent such a theory, maybe we should be sceptical about our intuition favouring phenomenal localism.

Third, Tye and Dretske might explain away our localist 'intuition' as follows: since we do not have to look to the past or future to know that we have certain phenomenal properties now – one need only introspect – we might erroneously conclude that they are temporally local. To see that this inference is erroneous, consider another case: my three-year-old daughter can immediately tell just by looking that something is a heart, without knowing about its evolutionary history. But *being a heart* is a historical, non-local property: if an intrinsic duplicate of the heart formed by chance in a swamp, it would not also be a heart, because it would lack the right evolutionary history – it would be a 'fake heart'. The general point: you can know something without knowing all its a posteriori consequences. Likewise, maybe on Dretske and Tye's historical externalism my daughter (or an adult sceptical of evolution) can immediately know about her phenomenal properties, even if she doesn't know about her evolutionary history.

8 Conclusion: a plea for an empirical approach

Many prominent philosophers (Chalmers, Hawthorne, Horgan, Shoemaker) rely on armchair arguments against reductive externalist theories of experience (e.g., tracking intentionalism, naive realism, active externalism).

My aim has been to identify the 'real problems' with central armchair arguments, because I think the criticisms of Dretske, Lycan, Tye and others fall short. There are additional antiexternalist armchair arguments I have not examined: for instance, the argument from the locality of mental causation (Fodor), the argument from introspection (Levine) and the argument from slow switching (Chalmers). But there are plausible replies.[15] Indeed, although I have not shown this here, I believe that some (non-reductive) externalist theories (notably naive realism and sense datum theory) cannot be clearly ruled out on the basis of *any* a priori arguments (Pautz 2010a). Here I have suggested to the contrary that armchair reflection on phenomenology *supports* externalism.

Nevertheless, my own sympathies lie with phenomenal internalism and the 'phenomenal intentionality program' mentioned in the

introduction. In particular, elsewhere (2006) I have defended an internalist, neo-Galilean 'projectivist' version of intentionalism. Chalmers (2006) also defends such a theory, calling it the *Edenic Theory*. My disagreement with armchair internalists like Chalmers is just this: I think that the only good arguments for internalism and against an externalist rival like naive realism are *empirical*.[16] Elsewhere I have developed three empirical arguments: the *internal-dependence argument*, the *structure argument*, and the *explanatory argument*.[17] (They differ from the faulty *simple empirical argument* of Prinz, Kriegel, and Horgan and Tienson that I briefly criticized in Section 3.) To decide the important externalism-internalism issue, we must get out of our armchairs and look seriously at work in neuroscience and psychophysics.[18]

Notes

1. For reductive externalism, see Dretske (1995), Lycan (2001), Tye (2000), Noë (2004) and Fish (2009, 153). I will not explain 'reductive' here. See Sider (2011, 116–132) for clarification and defence of a general reductionism about the manifest image. See §2 of this chapter for a case for reduction over alternatives (e.g., basic grounding relations).
2. See Kriegel (2011), Horgan and Tienson (2002), Loar (2003) and Mendelovici (2010). Pautz (2013) defends in detail the following 'consciousness-first' picture: Consciousness grounds rationality because it is implicated in basic epistemic norms. (For a related view, see Smithies, Ch. 6 of this volume.) In turn, the facts about rationality help to constitutively determine belief and desire (Davidson, Lewis). So consciousness also ultimately grounds belief and desire. Chalmers (2012, 467) briefly defends a related two-stage view on which acquaintance grounds normative inferential connections and these in turn pin down content.
3. Another well-known argument for phenomenal externalism starts with what I have elsewhere (2007, 251) called the *properties version* of the 'transparency observation' (see, e.g., Tye [forthcoming b]). But I think this 'transparency observation' (unlike the 'spatial datum') is far from pre-theoretically obvious, due to problems (not considered by Tye) concerning hallucination, many-property situations and a priori constraints on attentive awareness (Pautz 2007, 517, 522 and n. 12).
4. See, e.g., Kriegel (2011, 167) and Prinz (2012, 286).
5. The bad-off BIV has no visual receptor system or motor output system (just the central nervous system). Granted, if the BIV *were* suitably connected to a human body, its current neural state *would* track round things and cause round-appropriate behavioural movements. Could the BIV's standing in this counterfactual relation to roundness constitute its standing in the phenomenal representation relation to *being round* rather than to any other shape (e.g., *being square*)? No, for by *differently* hooking up the BIV to the world and a motor output-system, we could get its brain state to be caused by (say) *square* things and to cause *square*-appropriate behaviour.

6. However Johnston (2004), Pautz (2007) and Tye (forthcoming) provide an argument (not addressed by Kriegel or Mendelovici) for this claim, based on the fact that having *R* would necessarily enable one to know what such properties or qualities are like (which requires that they exist and that one is perceptually related to them).
7. In a recent book (2012), Prinz presents his 'AIR' theory, a materialist theory of consciousness. The AIR theory entails that experiences are essentially *intermediate representations*, defined as *representations of 'view-point relative microfeatures'* (see 124–126, 286; but see 327 for a contradictory claim). He also defends internalism: a BIV might have *R* (19, 286). He might say (20) that in having *R* the BIV phenomenally represents a response-dependent 'shape appearance' – a view I haven't covered here. Could he avoid my argument that internalism leads to primitivism about representation relations? No; indeed, even though his AIR theory entails that experience is inseparable from sensory representation, he provides no theory of sensory representation in the book. This is like *Hamlet* without the prince. Formerly, Prinz accepted Dretske's externalist theory of representation (see Pautz 2010c). But as we have seen, Dretske's *externalist* theory cannot be applied to the BIV, since its internal states don't have the biological function of indicating *any* properties (including Prinz's 'microfeatures' and 'appearances'). Elsewhere (2010c) I also argue that Dretske's theory is *incompatible with* Prinz's general view that experience represents 'response-dependent properties'.
8. Horgan and co-workers actually invoke internalist theses somewhat different from the one I have formulated in the text. But I think they can be set aside because they are problematic. (i) Horgan, Tienson and Graham (2004, 302) say that, intuitively, 'a [arbitrary] *physical* duplicate of oneself would also be a phenomenal duplicate of oneself'; similarly, Kriegel (2009, 79) claims that, intuitively, a *physical* duplicate of me 'would *have to* undergo the same conscious experience I undergo' (my italics). Unlike the more basic thesis of phenomenal internalism I formulated in the text (which is neutral on whether experience depends on intrinsic *physical* or *non-physical* nature), the thesis these philosophers are expressing is that mere intrinsic *physical* duplication would have to result in phenomenal duplication. But intuition clearly *doesn't* support this. In fact, the reverse is true: intuitively, an intrinsic physical duplicate of you (e.g., a BIV) might be *spectrum inverted* with respect to you, or a *zombie* without any experiences at all (e.g., if dualism is true and the psychophysical laws are highly contingent). (ii) Horgan and co-authors suggest that what they call *narrowness* is obvious: 'phenomenology does not depend constitutively on factors outside the brain' (2002, 526–527; 2004, 299, 301). The problem with this *brain-based* narrowness thesis is that it is also too strong to be justified from the armchair. It rules out *substance dualism* and *sense datum theory*, for on these views conscious experience depends constitutively on the existence and character of particulars *wholly distinct from the brain* (neither non-physical souls nor non-physical sense data reside in the brain). It also rules out the ancient view that the physical basis of consciousness is the *heart*. Even if these views are false, mere intuition isn't enough to rule them out. (iii) Horgan and Tienson (2002, n. 23) also suggest that what they call *intrinsicness* is 'self-evident to reflective introspection': 'phenomenology is not constitutively dependent on anything outside phenomenology

itself'. What does this mean? If 'anything outside phenomenology' means *anything 'whose nature is describable in non-phenomenological language'*, in the words of Horgan and Tienson (2002, n. 23), then this thesis simply amounts to *dualism*, so it doesn't capture any obvious intrinsicality thesis. If, on the other hand, 'anything outside phenomenology' means *anything that phenomenology does not constitutively depend on*, then the thesis becomes a trivial analytic truth and so is even compatible with reductive externalist theories.

9. Chalmers (2006, 56), Hawthorne (2004), Kriegel (2007, 321) and Levine (2001, 113) explicitly claim that we have strong pre-theoretical reason to accept phenomenal internalism (but see Chalmers 2006, 78, for the opposite claim). Block (1990, 16) and Burge (2003, 444) accept phenomenal internalism without argument.
10. For these versions of the 'inverted earth' case, see Lycan (2001, 30–31) and Levine (2001, 113).
11. Tye (forthcoming a, §3) makes a similar point.
12. See Chalmers (2004, 168; 2006, 56) and the quotes below.
13. Here is another way to see that possibility intuitions differ from the internalist intuition. Consider the dualistic sense datum theory. Or consider a (somewhat strange) dualist version of naive realism, on which external qualities and our acquaintance with them supervenes only *nomically* on the physical character of objects and the causal process from objects to the brain. Since they are externalist, such theories are inconsistent with the *internalist intuition*. But since they are also dualistic, they are quite consistent with intuitions concerning the possibility of spectrum inversion and brains in vats: they agree that acquaintance with qualities – and hence phenomenology – can vary independently of wide *physical* conditions.
14. See Tye (forthcoming a, §3) and (2000, 110).
15. For these arguments, see Fodor (1991), Levine (2001, 117) and Chalmers (2004, 354–355). For replies to Fodor's causal argument, see Dretske (1995, 151ff.) and Tye (forthcoming a, §3). For a reply to Chalmers's argument from slow switching and indeterminacy, see Lycan (2001). As for Levine's introspective argument, I think it too fails. Suppose you have the confident introspective belief that two of your colour experiences radically differ (for short, the 'difference belief'). Levine argues that if an externalist theory like tracking intentionalism is true (or even if you merely *believe* it is true), then your apparently indefeasible difference belief is in fact *defeasible* by (perhaps misleading) empirical evidence that the colour experiences track the *same* external reflectance feature: this evidence should make you reject your own confident introspective belief! My reply: this is an issue for *everyone* (Byrne 2003, 645). For instance, if neural identity theory is true (or even if you merely *believe* it), then the 'difference' belief should likewise be defeasible by (perhaps misleading) evidence that your underlying *brain states* are the same. Levine (117) also supposes that tracking intentionalism absurdly entails that you could confidently have the difference belief, even while it is *actually* false, because the colour experiences *actually* track the same feature. But this too is an issue for everyone: why cannot there be a radical mismatch between one's *most basic, simple* introspective beliefs and the true character of one's experience (constituted by tracking, brain states, or whatever)? Pautz (2010c, 359) sketches an answer (one available to externalists).

16. This bears on modal epistemology. Chalmers's (2009) two-dimensional approach and 'modal rationalism' require that all necessities (formulated in non–Twin Earthable terms) are a priori. But I think a counterexample is the necessary falsehood of certain relational-externalist theories, specifically *sense datum theory* and *naive realism*. True, Chalmers (2004, 168; 2006, 56) thinks some *materialist* externalist theories can be ruled out a priori on the basis of antimaterialist conceivability arguments about spectrum inversion and the like. But Chalmers cannot use these a priori antimaterialist arguments against *non-materialist* (dualist or 'pansychist') externalist theories, such as a non-materialist version of naive realism (see n. 13); indeed Chalmers considers such theories a priori *plausible* (2006, 79). Against 'modal rationalism', the necessary falsehood of such theories is only knowable a posteriori (n. 17).
17. The internal-dependence argument and the structure argument are discussed in Pautz (2006, 2010c). The explanatory argument is briefly put forward in Pautz (2010c, n. 23). In more recent work, I use these empirical arguments against naive realism.
18. My thanks to Angela Mendelovici, Boyd Millar, Declan Smithies and Mark Sprevak.

References

Block, N. (1990) 'Inverted Earth'. *Philosophical Perspectives,* 4, 53–79.

Burge, T. (2003) 'Phenomenality and Reference: Reply to Loar'. In M. Hahn and B. Ramberg (eds), *Reflections and Replies*. Cambridge, MA: MIT Press.

Byrne, A. (2003) 'Color and Similarity'. *Philosophy and Phenomenological Research,* 66, 641–665.

Chalmers, D. (2004) 'The Representational Character of Experience'. In B. Leiter (ed.), *The Future of Philosophy*. Oxford: Oxford University Press.

Chalmers, D. (2006) 'Perception and the Fall from Eden'. In T. Szabo Gendler and J. Hawthorne (eds), *Perceptual Experience*. Oxford: Oxford University Press.

Chalmers, D. (2009) 'The Two-Dimensional Argument against Materialism'. In B. McLaughlin and S. Walter (eds), *Oxford Handbook to the Philosophy of Mind*. Oxford: Oxford University Press.

Chalmers, D. (2012) *Constructing the World*. Oxford: Oxford University Press.

Devitt, M. and Sterelny, K. (1987) *Language and Reality.* Cambridge, MA: MIT Press.

Dretske, F. (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Field, H. (2001) *Truth in the Absence of Fact*. Oxford: Oxford University Press.

Fish, W. (2009) *Perception, Hallucination and Illusion*. Oxford: Oxford University Press.

Fodor, J. (1991) 'A Modal Argument for Narrow Content'. *Journal of Philosophy,* 88, 5–26.

Hawthorne, J. (2004) 'Why Humeans Are Out of their Minds'. *Noûs,* 38, 351–358.

Horgan, T., and J. Tienson. (2002) 'The Intentionality of Phenomenology and the Phenomenology of Intentionality'. In D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

Horgan, T., J. Tienson, and G. Graham. (2004) 'Phenomenal Intentionality and the Brain in a Vat'. In R. Shantz (ed.), *The Externalist Challenge: New Studies on Cognition and Intentionality*. Amsterdam: de Gruyter.
Jackson, F. (1977) *Perception*. Cambridge: Cambridge University Press.
Johnston, M. (2004) 'The Obscure Object of Hallucination'. *Philosophical Studies,* 120, 113–183.
Kriegel, U. (2007) 'Intentional Inexistence and Phenomenal Intentionality'. *Philosophical Perspectives,* 21, 307–340.
Kriegel, U. (2009) *Subjective Consciousness*. Oxford: Oxford University Press.
Kriegel, U. (2011) *The Sources of Intentionality*. Oxford: Oxford University Press.
Levine, J. (1997) 'Are Qualia Just Representations?' *Mind and Language,* 12, 101–113.
Levine, J. (2001) *Purple Haze*. Oxford: Oxford University Press.
Loar, Brian. (2003) 'Phenomenal Intentionality as the Basis of Mental Content'. In M. Hahn and B. Ramberg (eds), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. Cambridge, MA: MIT Press.
Lycan, W. (2001) 'The Case for Phenomenal Externalism'. *Philosophical Perspectives,* 15, 17–35.
Mendelovici, A. (2010) *Mental Representation and Closely Conflated Topics*. PhD diss., Princeton University.
Niyogi, S. (2005) 'Aspects of the logical structure of conceptual analysis'. Proceedings of the 27th Annual Meeting of the Cognitive Science Society.
Noë, A. (2004) *Action in Perception*. Cambridge, MA: MIT Press.
Pautz, A. (2006) 'Sensory Awareness Is Not a Wide Physical Relation'. *Noûs,* 40, 205–240.
Pautz, A. (2007) 'Intentionalism and Perceptual Presence'. *Philosophical Perspectives,* 21, 495–541.
Pautz, A. (2010a) 'Why Explain Visual Experience in Terms of Content?' In B. Nanay (ed.), *Perceiving the World*. New York: Oxford University Press, 254–309.
Pautz, A. (2010b) 'A Simple View of Consciousness'. In G. Bealer and R. Koons (eds), *The Waning of Materialism*. Oxford: Oxford University Press.
Pautz, A. (2010c) 'Do Theories of Consciousness Rest on a Mistake?' *Philosophical Issues,* 20, 333–367.
Pautz, A. (2013) 'Does Phenomenology Ground Mental Content?' In U. Kriegel (ed.), *Phenomenal Intentionality: New Essays*. Oxford: Oxford University Press.
Price, H. (1954) *Perception*. 2nd edn. London: Methuen.
Prinz, J. (2012) *The Conscious Brain*. Oxford: Oxford University Press.
Rosen, G. (2010) 'Metaphysical Dependence: Grounding and Reduction'. In Hale and Hoffman (eds), *Modality: Metaphysics, Logic and Epistemology*. Oxford: Oxford University Press.
Russell, B. (1905) 'On Denoting' In The *Philosophy of Language,* 3rd edn. A. P. Martinich, (ed.), Oxford: Oxford University Press, 1996, 199–207.
Russell, B. (1927) *The Analysis of Matter*. London: Kegan Paul.
Schmidtz, D. (2006). *Elements of Justice*. Cambridge: Cambridge University Press.
Shoemaker, S. (1994) 'The Phenomenal Character of Experience'. *Philosophy and Phenomenological Research,* 54, 291–314.
Sider, T. (2011) *Writing the Book of the World*. Oxford: Oxford University Press.
Tye, M. (forthcoming a) 'Phenomenal Externalism, Lolita and the Planet Xenon'.

Tye, M. (forthcoming b) 'Transparency, Qualia Realism, and Representationalism'. In *Philosophical Studies*.

Tye, M. (2000) *Consciousness, Color and Content*. Cambridge, MA: MIT Press.

Tye. M. (2012) 'Thinking Fish and Zombie Caterpillars'. Interview with Richard Marshall, www.3ammagazine.com/3am/thinking-fish-zombie-caterpillars/.

Winer, G. A., J. E. Cottrell, V. Gregg, J. S. Fournier, and L. A. Bica. (2002) 'Fundamentally Misunderstanding Visual Perception: Adults' Beliefs in Visual Emissions'. *American Psychologist,* 57, 417–424.