[Forthcoming in 'The Relational View of Perception: New Philosophical Essays' (Routledge) eds. Beck & Masrour. This is an uncorrected draft – please cite from the published version]

What Can the Naïve Realist Say about Total Hallucinations? Riding the New Relationalist Wave

Heather Logue and Thomas Raleigh[1]

1. Introduction

In this chapter we will explore relatively new avenues for developing and defending Naïve Realism (also known as Relationalism). We will understand Naïve Realism as a thesis about the 'phenomenal character' of perceptual experience. There are other ways of formulating Naïve Realism that do not treat it as fundamentally about phenomenal character – for example, a version of the view also focused on epistemological features of perceptual experience (the role it plays in generating and justifying beliefs about one's surroundings). But as a thesis about phenomenal character, the core claim is that 'what it's like' for a subject who enjoys a normal, successful perceptual experience of her surroundings consists – at least partially – in her being directly consciously aware of mind-independent objects and features in her external environment. For example, the phenomenal character of an experience of looking at a banana is constituted, at least partially, by *the banana*, its yellow colour, its curved shape, and its position in the space before the subject's eyes.

It is widely agreed that the strongest challenge to Naïve Realism comes from the alleged possibility of total hallucinations. These are experiences that are supposed to be phenomenally identical and hence subjectively indistinguishable from normal perceptual experiences, but which are not relational in nature (e.g., the sorts of experiences that brains-in-vats (BIVs) or subjects in the Matrix are alleged to have). So the phenomenal character of such hallucinatory experiences would not constitutively involve anything outside of the subject's head. If we then add the 'Common Kind' assumption – that perceptions and perfectly matching hallucinations are fundamentally the same kind of conscious state – we get the basic 'argument from hallucination', whose conclusion is that normal perceptual experiences are not relational in nature either.

The orthodox line of response for Naïve Realists to the argument from hallucination has been to reject the Common Kind assumption, and maintain instead that a perception and a matching hallucination are fundamentally different kinds of conscious state, despite their subjective indistinguishability. This position is, of course, called 'Disjunctivism'. We will not discuss Disjunctivism in this chapter, as its strengths and weaknesses have been explored in great detail elsewhere (see, e.g., the papers collected in Haddock and Macpherson 2008). Instead, we will focus on an emerging alternative line of response, or family of responses, for Naïve Realism—which we'll call 'New Wave Relationalism' (NWR). For present purposes, the core commitment of NWR is to reject the initial assumption that there can be total

[1] This chapter is entirely co-authored. The order of the names is merely alphabetical.

hallucinations that are both non-relational and subjectively indistinguishable from genuine perceptions. The primary aim for this chapter will be to explore and defend this broad NWR strategy for responding to the threat posed by hallucinations to Naïve Realism.

We will argue that which specific version or species of NWR a Naïve Realist should adopt will depend on their background methodological and metaphysical commitments. We will not attempt to adjudicate between rival stances on these commitments here, as that would take far more space than we have at our disposal. Our aim is to provide a roadmap of sorts for Naïve Realists about perceptual phenomenal character by arguing for claims of the following form: given a certain set of background commitments, certain versions of NWR are more (or less) attractive in the light of those commitments.

In section 2, we will articulate the relevant commitments in detail. In section 3, we will clarify the notion of total hallucination and flesh out the NWR response to the argument from hallucination. Sections 4 and 5 articulate and defend two variants of NWR, what we call the 'Object-Supplying' and 'Indistinguishability-Denying' strategies, respectively. Along the way, we flag the ways in which a Naïve Realist's background methodological and metaphysical commitments influence which variants of these strategies one should embrace.

2. Background commitments

Note that Naïve Realism about perceptual phenomenal character holds that it consists in the obtaining of the perceptual relation between a subject and the mind-independent objects of perception. Since there are (at least) two relata that a Naïve Realist can appeal to in accounting for perceptual phenomenal character, a key question for Naïve Realism is: how much of perceptual phenomenal character is explained by the subject-end of the perceptual relation, and how much is explained by the object-end (the mind-independent things perceived)?[2] The answers can be arranged along a spectrum of different views concerning the relations between various features of the subject, the perceptible features in the subject's external environment and the phenomenal character of the subject's perceptual experience.

At one extreme of the spectrum, perceptual phenomenal character is entirely (or perhaps almost entirely) explained in terms of the mind-independent things perceived.[3] This end of the spectrum is roughly what David Chalmers memorably labelled the 'Edenic' position:

> In the Garden of Eden, we had unmediated contact with the world. We were directly acquainted with objects in the world and with their properties. Objects were simply

[2] Some naïve realists, such as Bill Brewer, hold that there is a third relatum: 'an index of the conditions of perception, which involve the subject's spatio-temporal point of view, and other relevant circumstances, such as lighting, and so on' (2008: 171-2).

[3] You might think that there has to be a subject for there to be perceptual phenomenal character in the first place; if so, a version of naïve realism at this extreme holds that that's the extent of the explanatory contribution made by the subject-end of the perceptual relation.

> presented to us without causal mediation, and properties were revealed to us in their true intrinsic glory. When an apple in Eden looked red to us, the apple was gloriously, perfectly, and primitively red. There was no need for a long causal chain from the microphysics of the surface through air and brain to a contingently connected visual experience. Rather, the perfect redness of the apple was simply revealed to us. The qualitative redness in our experience derived entirely from the presentation of perfect redness in the world. (2006: 49)

Such views would treat the phenomenal character of experience as determined (almost) entirely by the perceptible objects and features in the external scene. Furthermore, in Eden the Sellarsian 'manifest image' (1962) is *all there is*—there are no microphysical goings-on constituting a 'scientific image' of the world that underlies the manifest image. The most naïve possible version of Naïve Realism is at this end of the spectrum.

At the other end of the spectrum lies the familiar position that the phenomenal character of experience supervenes only on the internal state of the subject (e.g., her brain and nervous system)—that perceptual phenomenal character is entirely explained in terms of features of the subject. Almost all views of this sort tend to be non-Relationalist rivals of Naïve Realism, on which the perceptible features in the subject's environment play a merely causal role but do not play any constitutive part in the subject's experience. However, there is logical space for a view at this end of the Naïve Realist spectrum: roughly, a view on which the mind-independent objects of perception are constituents of perceptual phenomenal character, but are mere 'pegs' on which the internally-generated phenomenal features are 'hung'.

In between the Phenomenal Internalism at one end of the spectrum and the Edenic view at the other, we have 'middle-ground' versions of Naïve Realism.[4] These are views on which the phenomenal character of perceptual experience is (somehow!) *jointly constituted* by both the external objects and features perceived and features of the subject (and perhaps also the manner in which these two factors interact). We can get an intuitive grip on the idea of joint constitution by thinking of the familiar way in which an object's spatial appearance is the joint upshot of both its intrinsic shape and also the subject's specific viewpoint or perspective.[5]

---

[4] Note that the middle-ground we have in mind is very broad—it's basically any view that's neither pure Eden nor pure Phenomenal Internalism. So some 'middle-ground' views will be just shy of Eden, and some will be just shy of Phenomenal Internalism.

[5] However, it is not obvious how exactly this idea should extend to colour appearances or to sensory appearances in non-visual modalities. Also, arguably, 'middle-ground' Naïve Realists who endorse the idea of joint constitution of phenomenal character have yet to give fleshed out answers to the question of what limits are there (if any) on how an object and subjective way/manner can combine to produce a specific phenomenal character/property (and related questions in the vicinity). If the sorts of factors one admits as 'ways' are limited to, for example, just the subject's perspectival positioning with respect to the object/scene, then there certainly will be limits of this kind. No matter how much one plays around with the subject's perspective and position, one will not be able to get a sphere to spatially look the way that a cube looks from head on. But if the subjective way/manner of perceiving can include arbitrarily radical alterations to the subject's perceptual mechanisms and brain and also arbitrarily radical prosthetic and technological ways of being causally sensitive to the object, then it may be that there are few or no limits. Any Naïve Realist defending a middle-ground view owes answers to questions like this—at a minimum, the answer may be that such questions are empirical and so unable to be settled from the philosopher's armchair.

Alas, there is no denying that we have fallen from Eden. We have eaten from Tree of Science (2006: 49-50)—there is more to the world than is revealed in the manifest image. For example, scientific investigation has revealed that seeing an apple does indeed involve "a long causal chain from the microphysics of the surface through air and brain" (Chalmers 2006: 49), and more generally that perceptual experience is underpinned by microphysical phenomena such as neural activity, perceptual processing, surface spectral reflectances, chemical properties and so forth.[6] So the Edenic extreme is not tenable. And the Trees of Illusion and Hallucination (Chalmers 2006: 49) threaten to push us even further from Eden—when it appears to S that something is F, but there isn't anything F that S is perceiving, we can't explain the phenomenal character of S's experience in terms of them perceiving an F thing (because they aren't).[7]

But exactly how far have we fallen? Answering this question is complicated, because there are (at least) two ways to fall from Eden, and the further one falls in one way the less one may fall in the other.

One way to fall is by complicating the subject-end of the perceptual relation beyond what is manifest to the subject in Eden, by holding that features of the subject do a significant amount of the work in determining perceptual phenomenal character. The more a Naïve Realist does this, they inch towards the Phenomenal Internalist end of the spectrum. It's not always entirely clear where to locate individual naïve realists on the spectrum, but arguably, most versions of Naïve Realism/Relationalism on offer are closer to the Edenic end: for example, Campbell 2002, Martin 2004, Johnston 2007, Brewer 2008, Stoneham 2008, and Fish 2009. However, some explicitly occupy a more middle ground (e.g., Logue 2012, French 2014, 2016, 2018, Beck 2019), and some are quite close to Phenomenal Internalism (Masrour 2020: 741, fn. 9).

One can also fall from Eden by complicating the *object-end* of the perceptual relation beyond what is manifest to the subject in Eden—for example, by holding that the potential objects of the perceptual relation go beyond the familiar ordinary physical objects like bananas and desks, and include entities like complexes of uninstantiated properties

[6] It's worth noting that Naïve Realist primitivists about sensory qualities (e.g., Campbell 1993, Allen 2016) can maintain that the fruits of the Tree of Science don't force us all that far from Eden. For example, on such a view, "[t]he qualitiative redness in our experience is derived [almost] entirely from the presentation of perfect redness in the world" (Chalmers 2006: 49). It's just that perfect/primitive redness is determined in some way by certain microphysical properties. By contrast, Naïve Realists who are suspicious of primitivism must fall further from Eden (perhaps by conceding that the phenomenology associated with some sensory qualities is partly determined by facts about the subject's visual system—see Logue 2012, French 2014, 2016, 2018, Beck 2019).

[7] In recent years, many have argued that illusions don't require a retreat all the way to Phenomenal Internalism (e.g., Brewer 2008, Fish 2009, Antony 2011, Kalderon 2011, French & Phillips 2020). The basic move is to set up camp in the middle ground of the spectrum, in claiming that the phenomenal character of illusions (and many partial hallucinations) can be explained as the joint upshot of features of the objects of perception and features of the subject (e.g., the way the subject's perceptual system processes input in the prevailing perceptual conditions). By contrast, Naïve Realists who are wary of the middle ground have to adopt another account (perhaps a negative V v I/H Disjunctivism is the only option—see Byrne & Logue 2008).

(Johnston 2004), or objects with spatiotemporally scattered parts (Byrne & Manzotti 2022), or parts of a BIV or Matrix apparatus (Raleigh 2014, Ali 2016, Masrour 2020).

As mentioned above, these two ways of falling from Eden interact with one another—if a Naïve Realist formulates their theory to avoid falling in one way, they'll be under pressure to fall in the other way. For example, suppose one wants to defend a version of Naïve Realism as far away from Phenomenal Internalism as possible. This commits one to explaining perceptual phenomenal character in terms of the object-end of the perceptual relation as much as possible. So in cases where there are no suitable candidate ordinary objects of perception (e.g., BIV/Matrix scenarios), one will be under pressure to say that the object of perception is one of the 'non-standard' kinds mentioned above. On the flip-side: suppose that one wants to hew as close to the manifest image as possible in terms of what the objects of perception can be (i.e., only ordinary objects, like bananas and desks). Then in cases where there are no suitable candidate ordinary objects of perception, one will be under pressure to explain the perceptual phenomenal character in terms of what's going on in the subject-end of the perceptual relation.

Finally, there is another, broader kind of background commitment that will influence which versions of NWR a Naïve Realist might willing to embrace. This commitment has to do with the explanatory relations that can hold between the scientific image and the manifest image. The orthodox view is that the manifest image can be *fully explained* in terms of the scientific image—for example, that we can give a complete account of phenomena like consciousness, agency, and the mind in general in terms of the microphysical entities that constitute it. However, one might be sceptical of this view—perhaps on the grounds that we've been giving it a go for quite some time and it doesn't seem to be panning out. Such a theorist might be amenable to a heterodox metaphysical picture which allows for 'top down' explanation of aspects of the scientific image in terms of aspects of the manifest image. The version of NWR articulated in section 5 is a live option only for those who are willing to consider embracing a heterodox metaphysics of this sort.

3. Matching Hallucinations, Total Hallucinations

It is common to assume that 'causally-matching' or 'neurally-matching' hallucinations (the sort of hallucination that would be had by a BIV or a subject in the Matrix) would be total hallucinations as they're traditionally conceived (hallucinations in which the subject doesn't perceive any mind-independent things, but which is nevertheless subjectively indistinguishable from an experience in which they do perceive mind-independent things). This is precisely what the proponent of NWR rejects. In order to clarify this commitment of NWR, we need to clarify the notions of causally-matching hallucinations, neurally-matching hallucinations, and total hallucinations. We'll then proceed to sketch two broad forms NWR can take, before detailing them more fully in the subsequent sections.

'Causally-matching' or 'neurally-matching' hallucinations are supposed to pose the most worrying challenge for the Relationalist (see, e.g., Martin 2004). These are experiences in which the subject's brain receives exactly the same proximal stimulation as in a case of veridical perception (causal matching) or in which the subject's brain processes are neuro-

physically identical to those involved in a veridical perception (neural matching). So, for example, if we are imagining a subject who is a brain in a vat or who is trapped in the Matrix, we typically imagine either that they receive exactly the same proximal stimulation as in some normal perceptual scenario, or that their brain is physically identical to the brain of a subject throughout some normal perceptual scenario (or both). In principle, a hallucination need not be either causally-matching or neurally-matching. But if it isn't, it's less clear why we should accept that it has the same phenomenal character as the relevant veridical perception—it's open to the Relationalist to argue that the causal/neural differences make for a phenomenal difference. This move isn't available in the case of causally-matching and neurally-matching hallucinations, which is why they are supposed to be particularly problematic for the Relationalist.

However, the notions of causally-matching and neurally-matching hallucinations need to be handled with some care. First, what exactly is the *proximal* stimulation in a case of normal perception? The suggestion is usually that it is some brain state late in the chain of perceptual processing—a state that can be brought about directly by the vat/Matrix, as opposed to being brought about through normal perceptual processing. However, at least some Relationalists are sceptical of the idea of a 'last' brain state that causes perceiving. In the context of defending a non-disjunctivist version of Relationalism, Mark Johnston insists:

> Seeing the object is not the next event *after the visual system operates*. Seeing the object is an event materially constituted by *the long physical process* connecting the object seen to the final state of the visual system. Seeing the object is an event that is (as it actually turns out) constituted by a physical process that goes all the way out to the object seen…There is no such 'last' brain state that then causes seeing. (2004: 139, emphasis in text)

Arguably, by the Relationalist's lights, the neural activity at issue is better thought of as a constituent of perceptual experience. Hence, framing the challenge for Relationalism in terms of causally-matching hallucinations is a dubious dialectical move. For this reason, the Relationalist's opponent might as well just press their case solely in terms of neurally-matching hallucinations, since they play exactly the same dialectical role in the argument from hallucination anyway.

However, when it comes to neurally-matching hallucinations, we need to ask: how much neural matching is there supposed to be? The neural matching can't be total, since the early neural activity initiated by the stimulation of the sense organs in normal perception won't be happening in vat/Matrix cases. Fortunately, we can give an answer to this question that both Relationalists and their opponents should be happy with.

Let's say that neurally-matching hallucinations are perceptual experiences that *only* involve the 'last' stage of neural activity that *constitutes* a veridical perception of some kind—i.e., *not* the neural activity involved in early perceptual processing. Roughly, the 'last' stage of neural activity that constitutes a veridical perception is the neural activity typically caused by early perceptual processing of input to the sense organs, and which is typically the

proximate cause of immediately post-perceptual mental states (e.g. perceptual judgments and perceptually-based intentions).[8]

For example, a Matrix-generated perceptual experience of a specific sort of banana is supposed to involve the last stage of neural activity involved in a veridical experience of that sort of banana. It doesn't involve the neural activity immediately resulting from (e.g.) retinal stimulation—in this scenario, there is no retinal stimulation. The idea is that the Matrix bypasses those stages of perceptual processing entirely, and directly stimulates the brain in a way that brings about neural activity that those earlier stages of perceptual processing would cause in the case of a veridical perception of that sort of banana. In turn, this neural activity causes further neural activity that constitutes early post-perceptual mental states (such as the judgment that there is a yellow, crescent-shaped banana before one). Or consider a 'swamp brain' hallucination, in which a cosmic fluke results in the formation of a brain that happens to be in the last stage of neural activity that constitutes a veridical perception of some kind. It doesn't involve the neural activity immediately resulting from (e.g.) retinal stimulation—in this scenario, the brain didn't exist before it was in its current neural state. The idea is that the brain just randomly pops into existence in the relevant neural state.

Now let's unpack the notion of a total hallucination more fully. As we noted in the introduction, the common conception of a total hallucination is a perceptual experience in which the subject doesn't perceive any mind-independent things, but is nevertheless subjectively indistinguishable from an experience in which they do perceive mind-independent things. A mental state m is *subjectively indistinguishable* from kind of mental state K if and only if the subject of m would not be in a position to know that m is not of kind K if they were in ideal conditions for reflecting on their mental states.[9] Articulating exactly what the 'ideal conditions' are is a delicate matter, but for our purposes it will suffice to give an indicative list: the subject isn't drunk, distracted, or otherwise cognitively impaired, and they have the required concepts (such as the concept of the relevant kind of mental state).[10] For example, a Matrix-generated perceptual experience as of a banana is subjectively indistinguishable from a specific kind of veridical perception of a banana (instantiating a relatively determinate shade of yellow, and a relatively determinate crescent shape, etc.) if and only if the subject of that experience would not be in a position to know that it is not a veridical perception of that kind if they were in ideal conditions for reflecting on that experience.

---

[8] If the boundary between perception and cognition is fuzzy (see, e.g., Logue 2013), it will be difficult to pinpoint exactly which neural activity counts as the last stage that constitutes a veridical perception. But this doesn't affect the challenge to the Relationalist—the point is that wherever we draw the boundary, there will be a neurally-matching hallucination that the Relationalist will struggle to account for. Also, note that including more of the subject's neural activity in what counts as "proximal" perceptual stimulation doesn't substantially alter the dialectic; as long as it falls short of perceptual contact with the subject's environment, the NWR can make all of the same moves we're about to explain.

[9] Note that the subject need not be occurrently recalling or imagining an experience of kind K; the point is that, if they were to do so, they wouldn't be in a position to know that m is not of kind K.

[10] Note that what the ideal conditions are might depend on what the correct metaphysics of perceptual experience is. For example, if naïve realism is true, knowledge of one's own experience doesn't just require looking 'within', which may well have implications for what the ideal epistemic conditions are (see Logue ms. (a)).

Now let's return to the assumption mentioned at the beginning of this section (now setting aside 'causally matching' hallucinations for the reasons given above): are neurally-matching hallucinations a subclass of total hallucinations, as they're typically construed? That is, are neurally-matching hallucinations perceptual experiences in which the subject doesn't perceive any mind-independent things, but which are nevertheless subjectively indistinguishable from an experience in which they do perceive mind-independent things?

Where the Disjunctivist answers this question in the affirmative, the distinctive thesis of NWR is to answer 'NO', allowing for a non-disjunctive response to the argument from hallucination. NWR denies the initial assumption of the argument that there can be hallucinatory experiences that are both subjectively indistinguishable to normal perceptual experiences and yet not essentially relational in nature. There are then two broad strategies that NWR can pursue.

The first strategy is one we'll call 'Object-Supplying'—it holds that the candidate experience does in fact have a relational, object-involving nature. On this strategy, neurally-matching hallucinations are subjectively indistinguishable from a possible perception, but they involve *perceiving* mind-independent things (or some relation akin to perceiving) after all—and so are not total hallucinations as they are typically construed.[11] The second strategy is one we'll call 'Indistinguishability-Denying'—it holds that the candidate experience does not in fact have phenomenology that is subjectively indistinguishable from normal perceptual experience. On this strategy, neurally-matching hallucinations do not involve perceiving mind-independent things, but they are subjectively *distinguishable* from ordinary perceptions of such things.

To briefly illustrate these strategies, consider (yet again) the BIV and Matrix scenarios. Many philosophers might naturally think of these as cases in which the subject, whilst conscious, is not consciously aware of her external surroundings at all – the vat/Matrix produces a total hallucination. A NWR theorist adopting the Object-Supplying strategy will want to describe the case as one in which the subject is in fact having a relational experience, and so will need to identify something external to the subject as the object of awareness – for example, some part or process of the vat/computer. By contrast, a NWR theorist adopting the Indistinguishability-Denying strategy will accept that the case is one in which the subject does not have a relational experience of her external surroundings, and will insist that the subject's experience is thereby subjectively distinguishable from an ordinary perception.

The Indistinguishability-Denying strategy is in one respect less 'radical' than the Object-Supplying strategy, insofar as it does not require accepting that the subject of a neurally-matching hallucination perceives anything. Many find this claim extremely counterintuitive;

[11] If one uses the word 'hallucination' such that hallucinations are by definition non-relational (e.g., as we've defined 'total hallucination' above), then according to the Object-Supplying strategy neurally-matching 'hallucinations' are not genuine hallucinations in that sense of the word at all (they are either illusions or perceptions). But this is merely a terminological matter: a different terminological choice can allow that it is still correctly called a 'hallucination' despite being relational in nature (see Masrour's distinction between d-hallucinations vs m-hallucinations—2020: 740-1).

however, in the next section, we will explore ways of dislodging such intuitions. The Object-Supplying strategy is in one respect less 'radical' than the Indistinguishability-Denying strategy, insofar as it does not require denying that two subjects—one in a perceptual scenario, the other in a neurally-matching hallucinatory scenario—will have the same phenomenology. The Object-Supplying NWR theorist can simply maintain that this common phenomenology is partially constituted by one kind of object in the normal perceptual case and by a different kind of external object in the (so-called) hallucinatory case. By contrast, the Indistinguishability-Denying strategy requires the NWR theorist to deny that the subject in the hallucinatory scenario has an experience with the same phenomenology as the subject in the normal perceptual scenario, *even when the two subjects are internal/neural duplicates with respect to proximal perceptual stimulation*. Of course, any variety of Relational theorist will already be committed to the idea that perceptual consciousness constitutively involves external stuff outside the subject's body. So the bare idea that such duplicates can have different conscious experiences (experiences with different 'phenomenal natures', as Martin (2002: 187) would put it) is not, by the Relational theorist's standards, especially radical. What is more radical is the idea that such duplicates can have experiences with *qualitatively different phenomenology* (different phenomenal characters, as Martin (2002: 187) would put it). For, at least prima facie, it can seem that in the case of such duplicates there is, by hypothesis, no scope for such an alleged phenomenal difference to make any kind of causal-functional difference. And so, again prima facie, one might think that such an alleged phenomenal difference would have to be causally epiphenomenal – i.e., unnoticeable and undetectable by the subject herself. However, in section 5, we explore how the Indistinguishability-Denying theorist can avoid this epiphenomenalist consequence by adopting a particular (heterodox) metaphysics of mind and the metaphysics of causation entailed by it. Again, this illustrates how evaluating NWR approaches depends on large background commitments and on one's dialectical starting point. Notice finally that it is a logical option to pursue a *mixed* NWR strategy, which adopts the Object-Supplying strategy for some purported hallucinatory experiences and the Indistinguishability-Denying strategy for other purported hallucinatory experiences.[12] For reasons of space we will not discuss such mixed strategies here.

4. The Object-Supplying strategy

There are by now quite a variety of views in the literature that can be thought of as pursuing some version of an Object-Supplying strategy in response to hallucinatory or 'Brain in a Vat'/'Matrix' style scenarios. But it is worth noting that not all of these Object-Supplying theorists are interested in defending a Naïve Realist account of experience – so not all of these Object-Supplying theorists should be counted as 'New Wave Relationalists'.

There is a tradition, arguably starting with Bouwsma (1949) and including Putnam (1981), Davidson (1986) and Chalmers (2005, 2022), which maintains that in a supposedly sceptical

[12] The options are not in principle mutually exclusive with respect to any particular hallucinatory experience: one could consistently say that a given neurally-matching hallucination involves *perceiving* mind-independent things, and that it is subjectively *distinguishable* from ordinary perceptions of such things. However, it's not clear what the motivation for occupying this region of logical space would be.

BIV style scenario the subject would in fact have largely true beliefs that refer to the vat-generated 'world' that it experiences. Whilst the focus for these theorists was the accuracy (and reference) of the envatted subject's beliefs, they are all presumably committed to the idea that an envatted subject would enjoy *perceptions* of her vat-environment, which would allow her to form (largely) true beliefs that successfully refer to objects and features in her vat-world. This can be thought of as a kind of 'Object-Supplying' strategy insofar as it denies that the envatted subject would have purely internal, hallucinatory experiences that leave her entirely 'out of contact' with her surroundings. However, none of these theorists were pursuing this Object-Supplying strategy as a way of defending Naïve Realism about the metaphysics of perceptual experience – they were concerned rather with its anti-skeptical and/or reference-securing potential. Likewise, an important recent paper by Byrne and Manzotti (2022) argues that in hallucinatory experiences there is always an external, physical object of experience, though this object may be 'gerrymandered' and consist of spatio-temporally scattered parts. Byrne and Manzotti argue that their view best accounts for the 'palpable particularity of hallucination' – i.e. they want to maintain that the phenomenal character of hallucination is object-dependent in broadly the same way that the phenomenal character of successful perceptions is, supposedly, object-dependent. But again, Byrne and Manzotti are not themselves interested in defending Naïve Realism, and they frame their view in terms of a *proposition* determining how things seem for the subject. (However, Byrne and Manzotti do explicitly note that their view 'should be congenial to the naïve realist dissatisfied with the current menu of theories of hallucination' (2022: 330, fn. 8). They also take their own preferred propositional formulation to be one way of cashing out the intuitively appealing and Naïve-Realist-flavoured idea that in successful perception a 'portion or tract of reality is revealed' (2022:329).) The fact that neither Byrne and Manzotti nor the theorists in the Bouwsma-Putnam tradition were motivated by defending Naïve Realism might nevertheless be construed as helpful to the NWR cause. For it shows that there can be *independent motivations* to endorse an Object-Supplying account of hallucinations that have nothing to do with defending Naïve Realism. The Object-Supplying strategy thus should not be seen as an ad-hoc or desperate manoeuvre for defending one specific view about the metaphysics of perception, but as an approach to hallucinations that has been found independently plausible by philosophers outside of the Naïve Realist camp.

Another version of the Object-Supplying strategy – though here the word 'object' is somewhat awkward – which *is* employed in defence of a Relational theory of experience can be found in the work of Mark Johnston (2004). The core idea here is that in cases of hallucination the phenomenal character of the subject's experience *does* still consist in a relation to something external and mind-independent, but the relation is to *uninstantiated sensible qualities*. According to Johnston then, there is a common factor that the subject is aware of across both perception and hallucination: a complex of sensible properties, or 'sensible profile'. In the case of a genuine perceptual experience these sensible properties are instantiated by the physical objects in the subject's immediate environment. In the case of hallucination, the subject is aware of the same sensible profile, but here the properties are uninstantiated. Johnston thus wants to avoid the idea that in hallucination we are aware of mental objects, such as sense-data, and also the idea that we are aware of any distinctively mental qualities, such as 'qualia'. He would also, of course, deny that we are aware of any external physical objects – what we are aware of are familiar properties like

shapes and colours that just happen to be uninstantiated, hence the slight awkwardness with labelling Johnston's view 'Object-Supplying'. However, Johnston does explicitly allow that 'sensible profiles…are themselves *objects of awareness*' (Johnston 2004: 149, emphasis added). He also takes a key virtue of his view to be that: 'It is ordinary qualities… that account for the so-called subjective character of [hallucinatory] experience' (2004: 146). It is thus fair to treat Johnston as effectively offering an Object-Supplying, Relational approach to hallucinations, though the alleged 'object' is of a somewhat metaphysically exotic kind. In a similar vein Butchvarov (1998) argued that in hallucination we have awareness of Meinongian, non-existent objects[13]. So, Byrne and Manzotti, Johnston and Butchvarov all provide versions of the Object-Supplying strategy that rely on appealing to somewhat metaphysically exotic kinds of objects of awareness for hallucinations – gerrymandered 'scattered' objects, uninstantiated property-complexes, or Meinongian non-existent objects. Though we will not here attempt to offer further support for any of these 'exotic' Object-Supplying strategies, they do serve to illustrate one of the main themes of this chapter: which is that evaluation of Naïve Realism and of NWR in particular, will turn on evaluating various wider metaphysical theses about the nature of objects and properties in general.

Raleigh (2014), Ali (2016) and Masrour (2020) all propose an Object-Supplying strategy for BIV style scenarios that has similarities with the Bouwsma-Putnam approach, but which is explicitly committed to a Relational account of perceptual experience.[14] The proposed object of hallucination on these views is some part of the Vat or Matrix computer, which is hypothesized to be creating the pattern of stimulations for the subject's brain. Whilst this is hardly an everyday, familiar object of perception, it is not something that would be at all *metaphysically exotic*. Moreover, it is an object that has been hypothesized by the Relational theorist's *opponents*, so it is fair game for the Relational theorist to make use of. Both Raleigh (2014) and Masrour (2020) emphasize how there must presumably be some kind of internal data-structures or mechanisms in the hypothesized device that are playing a very similar causal role (vis-à-vis the subject's perceptual systems) as the familiar external objects and features in the normal perceptual case. Raleigh and Masrour also both point out that these internal structures will presumably need to exhibit the right kind of counterfactual dependencies concerning what the subject *would* experience were she to perceptually explore and attend differently to the virtual scene, so that as the subject's visual system tries to saccade and scan the experienced (virtual) scene the (virtual) objects and features display the right kinds of perceptual constancies (etc.) so that it is indistinguishable from the normal perceptual case. Likewise, if a different conscious subject

[13] One might also think of the view of Umrao Sethi (2020) here. Sethi's view is that the familiar, ordinary perceptible properties of objects – such as macroscopic shapes and colours – can exist either as mind-independent instances that 'inhere' in objects or as mind-dependent instances. As with Johnston's view, Sethi holds that in both perception and hallucination we are aware of the very same sensible properties – e.g. shapes and colours – though only in the perceptual case are these properties instantiated by familiar external objects. What is distinctive about Sethi's view is that she allows that these properties can be 'ontologically over-determined', so that in the perceptual case the one token property instance exists *both* as a quality of the external object *and* as a mind-dependent instance. Whilst there are affinities with Johnston's position, Sethi's view is perhaps not so naturally thought of as a version of the Object-Supplying strategy, since in the hallucinatory case the familiar properties of shape and colour (which are supposed to fix the phenomenal character of the experience) are held to be mind-dependent.

[14] See also Thompson and Cosmelli 2011, although strictly speaking they are defending Enactivism (which is very similar in spirit to Naïve Realism).

were hooked up in the right way to these same data-structures within the Vat/Matrix computer, they would enjoy the same kinds of experiences and could perform the same kinds of perceptual explorations of the virtual, computer-generated scene. Thus a plausible case can be made that there would have to be standing structures within the hypothesized machine/computer which would be importantly isomorphic to the external objects and features in the normal perceptual scenario so as to mimic the causal roles played by those familiar objects and features in bringing about the subject's experiences. And so the NWR theorist can plausibly maintain that envatted/Matrix subject would be having relational, object-involving experiences of these structures within the Vat/Matrix device. This Object-Supplying NWR strategy with BIV/Matrix scenarios relies on the idea that being connected to the Vat/Matrix machine in the right way can be a way of gaining perceptual awareness of some internal parts or features of the machine.

It is worth noting in this regard that all Naïve Realists will presumably *already* accept that certain kinds of external features are not visible to normal unaided human vision, but could become visible to humans via alterations/additions to the human visual system. We know that various birds and insects are visually sensitive to light in the ultraviolet spectrum and so can see colourful patterns on the petals of flowers that are invisible to us humans. But we humans could come to see these external colourful features of the world if our visual systems were altered in the right way. Likewise, there are features which we cannot see using unaided vision but which we can see using a microscope or telescope or various other kinds of visual prostheses[15]. And so a proponent of the Object-Supplying strategy can plausibly maintain that the internal parts of a Vat/Matrix super-computer can have various potentially visible features that are invisible if we look at them using normal unaided human vision and normal physical light reflected from them, but which become visible when the visual system is hooked up in the right way. Since all Naïve Realists must presumably already accept that if we were to modify or augment the human perceptual systems in the right way, sensible qualities of the external environment would be revealed that are not revealed to normal, unaided human consciousness, it should be no great stretch for Naïve Realists to accept that when the human visual system is hooked up in the right specific way to the super-computer, sensible qualities of the computer are revealed that would be invisible if we were to look at the innards of the computer with normal unaided human vision. One way of developing this would be to treat the Vat-computer's parts as having more than one colour, a position defended by various 'primitivists' about the metaphysics of colour (see e.g. Kalderon 2007, Allen 2016). Alternatively, one could maintain the more orthodox view that a physical surface only has one colour – i.e. one *physical colour* – but allow that in realising a virtual object, components inside the computer can also realise a *virtual colour*.

How exactly to think of the role of the Vat/Matrix computer, *qua* object of experience, will depend on where one is positioned on the spectrum between Edenic and phenomenal-internalist views. If one is a NWR theorist at the Phenomenal Internalist end of the spectrum, then all one has to make plausible on this Object-Supplying strategy is the idea that some part or process in the Vat/Matrix machinery is the object of relational consciousness, where this does not require that the object do any work constituting or

[15] Masrour (2020) discusses night vision goggles, whilst Raleigh (2014) discusses the technique of de-focusing the eyes in just the right way which must be learned in order to see the picture in a stereogram.

determining the specific phenomenal character of the experience. If one is a NWR theorist located at least some way towards the phenomenally externalist Edenic end of the spectrum, then one would also have to make plausible the idea that the relevant parts or processes in the Vat/Matrix machinery can at least partially constitute the phenomenal character of the subject's experience. And whilst it may seem relatively plausible that some part or process in the computer can play the same causal and counterfactual roles vis-à-vis the subject's perceptual system as a normal perceptual object (such as a lemon) it may, prima facie, seem less obvious that a part or process inside a super-computer could be partially constituting *this* specific kind of yellow-ish and round-ish phenomenal character I enjoy as I look at a lemon. So what can be said by NWR theorists in support of the idea that data-structures and processes in a computer, which are not themselves yellow or lemon-shaped, could partially constitute the phenomenal character of a relational experience as of a yellow lemon?

Firstly, according to the majority of Relationalists located in between the extremes of Eden and Phenomenal Internalism—who accept that the external objects of perception only *partially* constitute the phenomenal character of experience—for any specific phenomenal character/property (way of appearing in consciousness), there will be a range of intrinsically different external properties that can partially constitute that specific character when combined with the right mode/way/manner of perception[16]. For example, there is a whole range of different (3-D) shapes that can partially constitute a specific phenomenal, spatial way of looking if they are viewed in the right, specific way (from the right angle and distance, etc.). Likewise, there is a whole range of different surface reflectance properties that can partially constitute a specific phenomenal colourful way of looking if viewed in the right specific way/manner (i.e. lighting, surrounding context and shadows, also perhaps facts about the visual system). All Naïve Realists must already accept the existence of metamers, and so accept that a range of intrinsically quite different things can all look yellow in normal perceptual experiences. And of course all Naïve Realists accept the role of spatial perspective and so accept that a range of differently shaped things can all look, say, ellipse-shaped in normal perceptual experiences. Conversely, a specific perceptible property/feature can partially constitute a whole range of different phenomenal ways of looking depending on the way/manner/subject side of the relation. The point here is that for any Naïve Realists who accept the idea that external features only partially constitute phenomenal character, they must *already accept* that a whole range of intrinsically different external features can all end up appearing the same, indistinguishable way in experience – e.g. this specific yellow-ish way of looking – given the right specific ways of being perceptually related to them. So there would not be any extra theoretical commitments in accepting that yet another kind of external feature – such as a specific kind of data structure within a super-computer – can also appear looking that specific way, given the right, very specific way of perceptually relating to it (i.e., being hooked up to the super-computer in just the right specific way).

[16] Naïve realist talk of 'ways' or 'manners' or perceptually experiencing an object might be thought of as something on the subject side of the conscious perceptual relation, or alternatively as characterizing the relation itself. For further discussion see Raleigh 2021 and the contributions by Pautz and by French & Phillips in this volume.

Secondly, just as familiar physical objects have physical shape and physical colour properties, so *virtual* objects – i.e. objects realised by computational structures – can have *virtual shapes* and *virtual colours*. And so, just as physical surfaces that reflect physical light in a specific way will look yellow to normal human observers, so also virtual surfaces which reflect virtual light in a specific way can look yellow to envatted human observers. In both cases the human's visual system is operating normally and successfully revealing a real, standing mind-independent feature that other observers could also be perceptually related to if they occupy the right 'standpoint' or 'perspective' (where in the case of the BIV/Matrix this 'perspective' requires being hooked-up correctly to the computer). Notice that this allows even a Naive Realist located more towards the Edenic extreme of the spectrum to accept that a Vat-computer could constitute or determine the yellow-ish phenomenal character of an experience. For even though the computer may not contain any *physically-realised* instances of yellow for the subject to be perceptually related to, it contains some *virtually-realised* instances of yellow for the subject to perceive. Here the role of the subject's brain and visual system being hooked up to the computer in the right way would be conceived as merely *enabling* this external virtual colour property to be revealed, without playing any role partially constituting the phenomenal yellow-ish-ness.

Thirdly, the Object-Supplying strategy need not treat the subject's experience of the Vat/Matrix as *veridical*; they can allow that the subject's experience of the internal parts of the Vat/Matrix computer is illusory. How exactly Naïve Realists are to think of illusions is a large and controversial subject. But (as mentioned above in fn. 6) these days it is fairly common for Naïve Realists to hold that in the case of (say) the Muller-Lyer illusion, it is the lines' actual shapes and actually-equal-lengths which are partially constituting the phenomenal character of the experience – even though the lines end up appearing unequal in length in the subject's consciousness. Likewise, with the straight stick in water: it is still the stick's actual straight shape that is doing work partially constituting the phenomenal character of experience despite the fact that it looks bent. And so when it comes to the internal parts of the hypothesized computer in a BIV/Matrix scenario – once we assume that the subject's experience is illusory, these parts need not be actually yellow in colour in order for them to be doing constitutive work determining the yellow-ish phenomenal character of the subject's experience[17].

In summary: accepting that parts and processes inside a hypothesized Vat/Matrix computer could at least partially constitute the phenomenal character of experience does not require any significant extra theoretical commitments from Naïve Realists. A component of the computer that would not look phenomenally yellow-ish to normal unaided human vision, could look yellow-ish when the visual system is connected to the computer in precisely the right way – allowing the subject to be acquainted with this component of the computer in the required way. After all, such component processes in the computer would presumably have been ingeniously designed so as to precisely match the causal-counterfactual profile of a familiar physical object of perception, such as a yellow lemon, so long as the subject is connected to the computer in the required way. So for a Naïve Realist there should be no intuitive resistance to the idea that the computer's components can also play the same sort

[17] See Ali (2016) for extended discussion of how 'hallucinatory' scenarios of this kind could be treated as cases of illusions.

of role in constituting phenomenal character that, say, a yellow lemon can play – again, given that the subject is 'viewing' those computational components in just the right way. This could be thought of as a case of successful perception or as a case of illusory but still relational experience. It could be thought of in terms of virtual colours, or in terms of the physical computer having multiple physical colours. And depending on where one stands on the spectrum from Eden to Phenomenal Internalism, the component of the computer could play (almost) no role constituting the phenomenal character, or it could partially constitute it together with contributions from the subject's side. (And perhaps even at the Edenic end of the spectrum, the computer could be fully determining the phenomenal character in virtue of its realisation of virtual objects with virtually-realised shapes and colours.)

In our view then, the idea that one can be perceptually acquainted with parts/processes inside the Vat/Matrix device seems to be the most promising version of the Object-Supplying strategy for NWR to deal with BIV/Matrix style scenarios. It also does not require the subject to be acquainted with anything as exotic and potentially controversial as uninstantiated properties, spatio-temporally scattered objects or Meinongian non-existent objects. It just requires that the parts/processes in question can, when one is perceptually related to them in just the right way, look/appear just like familiar everyday objects. But of course, this approach will not apply to 'Cosmic Swamp-Brain' style scenarios in which it is stipulated that there is no such external object for the subject to be consciously related to. To the extent that one thinks of these Cosmic Swamp-Brain scenarios as genuine possibilities that demand a response, any NWR theorist who wishes to adopt our recommended Object-Supplying strategy for classic BIV/Matrix cases could only do so as part of a mixed strategy that has something else to say about Cosmic Swamp-Brains.

5. The Indistinguishability-Denying strategy

In contrast to the Object-Supplying strategy, the Indistinguishability-Denying strategy holds that neurally-matching hallucinations do not involve the subject perceiving mind-independent things—and *for this reason* they are phenomenally different from ordinary perception. Moreover, this strategy claims that they phenomenally differ from ordinary perception to the extent that the subject is able to distinguish them from ordinary perception just on the basis of reflection on their experience. As is frequently observed, perception has a distinctively 'presentational' phenomenal character—a kind of phenomenal character that is far more rich, vivid, and immersive than that associated with other kinds of mental states (e.g., judgment). Arguably, talk of 'presentation' fails to do justice to the way in which perceptual phenomenal character shoves the world 'in your face' (and all over the rest of your body). For this reason, in what follows we will refer to this kind of phenomenal character as full-on, in your face phenomenal character. According to the Indistinguishability-Denying strategy, whatever phenomenal character comes with mere neural stimulation, it is something that falls noticeably short of that.

This view is a radical departure from contemporary orthodoxy in philosophy of perception, and so getting people to take it seriously is no mean feat. One way of warming up to it is to reflect on what would happen if we built a BIV or Matrix apparatus and stimulated a brain in the same way it's stimulated in the course of ordinary perception. Arguably, we don't know

for sure what would happen. One possibility—the one pretty much everyone assumes would in fact be the case—is that the experiences had by such a subject would be phenomenally exactly like experiences had by subjects of ordinary perceptions, and hence subjectively indistinguishable from them. The other—arguably underappreciated—possibility is the one put forward by the Indistinguishability-Denying strategy. On this possibility, the experiences had by such a subject would be phenomenally *different* from experiences had by subjects of ordinary perceptions, and hence subjectively *distinguishable* from them.

At this point in the dialectic, the key point to note is that *both* possibilities have the epistemic status of *empirical conjectures*. Of course, there are reasons that have led practically everyone to bet on the orthodox possibility, which we will come to in a moment. But do these reasons constitute *conclusive* reasons in favour of believing that this possibility would turn out to be actual if we managed to construct a BIV or Matrix apparatus? As we will argue shortly, these reasons certainly have considerable force— but they fall well short of being conclusive. This is because their conclusiveness depends on heavy-duty background methodological and metaphysical commitments about which reasonable people can disagree.

Let us now turn to these reasons. This is not the place to provide an exhaustive catalogue of all the possible reasons and detailed responses to them (for that, see Logue ms. (b)); here, we will limit our focus to the ones we suspect are most likely behind resistance to the unorthodox possibility.

One thought that seems to be behind much of the resistance is the idea that *neural activity alone is sufficient for full-on, in your face perceptual phenomenal character*. But why should we accept this claim? Arguably, the strongest case would be some actual conclusive (or near enough) empirical evidence for it.[18] In the face of such evidence, it would be difficult to deny that a neurally-matching hallucination would be phenomenally exactly like an ordinary perception of some kind and hence subjectively indistinguishable from one. However, Bill Fish (2009) has persuasively argued that the kind of empirical evidence typically cited shows—at best—that neural activity alone is sufficient for *some* kinds of perceptual phenomenal character (e.g., phosphenes or perceptual imagery). This evidence does not support the further claim that the kind of phenomenal character generated by neural

[18] Alternatively, one might support something in the vicinity of claim in question by way of a conceivability argument: it's conceivable that neural activity generates full-on, in your face perceptual phenomenal character on its own, so it's possible, and this possibility is one the naïve realist must account for. Or one might leave neural activity out of it altogether and argue that since full-on perceptual phenomenal character without perception is conceivable, it's possible—which would also cause trouble for the Indistinguishability-Denying theorist (and indeed the Object-Supplying theorist). The link between conceivability and possibility is of course a controversial issue in modal epistemology that we don't have the space to discuss fully here, so we'll just sketch a line the New Wave Relationalist can take. Even if conceivability entails possibility, it's *ideal* conceivability that entails metaphysical possibility (Chalmers 2002). And it's far from clear that the scenarios at issue are ideally conceivable. Ideal conceivability is something along the lines of conceivability upon ideal rational reflection. And who's to say that ideal rational reflection couldn't reveal that "in your face" perceptual phenomenal character requires perception? The brute assertion that it couldn't is arguably tantamount to begging the question against naïve realism. (See also Fish 2009: section 5.2, and Masrour 2020: section 6.)

activity alone can be scaled up into full-on, in your face phenomenal character.[19] The empirical evidence is consistent with either possibility regarding what would happen if we managed to construct a BIV or Matrix apparatus, and so the Indistinguishability-Denying strategy is a live option for the Relationalist.

For the sake of argument, let's suppose that the experiences had by a BIV or Matrix subject would be phenomenally different from experiences had by subjects of ordinary perceptions. The question then arises: how phenomenally different would they be? Officially, a proponent of the Indistinguishability-Denying strategy ought to regard this as an open empirical question. Again, it is not clear what would entitle us to the claim that we know for sure what would happen if we built a BIV or Matrix apparatus just based on armchair philosophical speculation.

But a Naïve Realist's background commitments will shape their conjecture about what would happen. For example, consider a proponent of the Indistinguishability-Denying strategy who is keen to defend a view as far from Phenomenal Internalism as possible, and so wants to minimise the contribution of features of the subject to perceptual phenomenal character. Their conjecture would be that the relevant sort of neural stimulation alone would result in extremely attenuated perceptual phenomenal character, or perhaps none at all. By contrast, one who is open to a more 'middle ground' version of naïve realism might be inclined to bet that neural stimulation alone would result in more substantial perceptual phenomenology—something that still falls noticeably short of full-on, in your face phenomenal character, but is nevertheless somewhat vivid (something akin to Humean ideas as opposed to impressions). The same considerations apply in the case of the Cosmic Swamp-Brain scenario: what phenomenology would the newly formed swamp subject enjoy? The options for a proponent of the Indistinguishability-Denying strategy range from total 'darkness within,' at one end, to imagery-like phenomenology that is subjectively *dis*tinguishable from the full-on, in your face phenomenal character typical of perception, on the other.

In short, the reasons given in support of the claim that neural stimulation of the relevant sort is sufficient for full-on, in your face perceptual phenomenal character are inconclusive. It is still an open empirical question what would happen if we built a BIV or Matrix apparatus

[19] Instead of appealing to the results of direct brain stimulation, one might appeal to relatively mundane hallucinations (such as an exhausted new parent's hallucination as of their baby crying) in order to support the claim that neural activity is sufficient for full-on, in your face perceptual phenomenal character (see Beck 2023). Let's grant for the sake of argument that "noise" in the perceptual system is sufficient for at least some aspects of the phenomenology of a mundane hallucination, such as the auditory phenomenology as of a baby's cry (as suggested by Beck 2023: section 2.2). The key question is whether the phenomenology attributable to the perceptual noise alone can be scaled up into the full-on, in your face perceptual phenomenal character typical of ordinary perception. And is far from clear that it can. One of the author's recollections of her own new-parent-crying-baby hallucinations is that they were quite far from having in full-on, in your face phenomenal character—the phenomenology was much fainter, and prompted doubt (and wishful thinking) concerning whether the crying was actually happening. In short, the mundane hallucinations Beck appeals to are arguably consistent with the following hypothesis. When they have full-on, in your face phenomenology, that is attributable to the subject (mis)perceiving some mind-independent entity in their environment, and any phenomenology attributable to perceptual noise alone falls short of that kind of phenomenology.

and stimulated the subject's brain in the relevant ways, and the possibility that the subject would be in states phenomenally different and subjectively indistinguishable from ordinary perception has not been ruled out.

However, there is another, much more powerful reason underlying resistance to the Indistinguishability-Denying strategy. This reason stems from a commitment to neural determinism: something along the lines of the idea that the neural states that a subject is in at a given time determine the neural states that the subject is in immediately afterwards (aside from the neural states that are determined by non-neural inputs). In the case of an ordinary veridical perception of a yellow, crescent-shaped banana, the subject (S_P) is in a neural state that (partially, according to the Naïve Realist) constitutes their perceptual experience. Given neural determinism, this neural state determines further neural states, and some of these constitute certain doxastic dispositions—in particular, a disposition to believe that one is perceiving a yellow, crescent-shaped banana. Now consider a subject in a BIV or Matrix apparatus (S_H), where their visual cortex is being stimulated so as to exactly replicate the neural state that partly constitutes S_P's ordinary veridical perception of a yellow, crescent-shaped banana. Given neural determinism, this neural state would determine further neural states, and some of these would constitute certain doxastic dispositions—including the disposition to believe that one is perceiving a yellow-crescent-shaped banana.

It is not clear how the Indistinguishability-Denying strategy can make sense of this. The strategy holds that the BIV/Matrix subject *can tell their experience apart* from an ordinary veridical perception of a yellow, crescent-shaped banana, and this ability would presumably be embodied by a disposition to believe that they are *not* veridically perceiving a yellow, crescent-shaped banana (on the basis of the lack of full-on, in your face perceptual phenomenology). But neural determinism entails that this subject *would* be disposed to believe that they are veridically perceiving a yellow, crescent-shaped banana. So how could the subject possibly be in a position to tell that they are not veridically perceiving, as the Indistinguishability-Denying strategy claims?[20]

To avoid this result, a proponent of the Indistinguishability-Denying strategy can reject the presupposition of neural determinism that leads to it.[21] That is, they can deny that the perceptual neural state involved in the BIV/Matrix scenario would cause the same doxastic neural states as it would in the case of ordinary veridical perception. And they can do this if they help themselves to *top-down causation*. This would allow them to say that the fact that the BIV/Matrix experience lacks full-on, in your face perceptual phenomenal character

[20] Perhaps it is possible for a subject to harbour dispositions to believe propositions that contradict each other, but it would certainly be weird. The Indistinguishability-Denying strategy would be on far stronger ground if it could avoid this result.

[21] Another possibility is to accept neural determinism, and so accept that S_P and S_H go into the same kind of post-perceptual neural state, but reject that their neural states constitute the same kind of doxastic state (it's a disposition to believe that one is perceiving a yellow crescent-shaped banana in the former case, and a disposition to believe that one is not perceiving a yellow crescent-shaped banana in the latter case). We won't pursue this possibility here, but it's worth noting that it's not obviously an easier row for the Indistinguishability-Denying theorist to hoe: they'll have to defend a theory of doxastic content on which this is possible, and they'll have to give up on local supervenience of the mental on the physical.

causally affects which neural state the subject goes into next—the subject goes into a *different* neural state than they would have gone into had they been perceiving, and this neural state constitutes something *other* than the disposition to believe that they are perceiving. To be sure, the notion of top-down causation is typically regarded with suspicion these days, and so this route is not for metaphysical conformists. In the remainder of this section, we will briefly sketch a broader metaphysical picture that vindicates top-down causation, and some reasons one might find it attractive.[22]

David Charles (2021) has recently articulated and defended a metaphysics of mind that he calls neo-Aristotelian *inextricabilism*. According to Charles' interpretation of Aristotle, mental phenomena are 'inextricably psycho-physical'. This claim breaks down into two parts:

> [A] You can't give a complete theory of the mind without mentioning specific physical entities (e.g., neurons).[23]
>
> [B] You can't give a complete theory of those specific physical entities without mentioning the mental states and processes they're involved in (e.g., perception).[24]

[A] goes against the current conventional wisdom that we can isolate 'purely mental' subjects, states, and processes, such that we can give a complete theory of them without mentioning any physical entities at all, and *then* ask how they're related to the physical domain. Similarly, [B] goes against the current conventional wisdom that we can isolate 'purely physical' entities that are involved in mental states and processes, such that we can give a complete theory of them (in terms of the kinds of entities postulated by physicists, say), and *then* ask how they're related to the mental domain. We can of course theorise about physical and mental entities independently of each other, but this involves *abstracting away* from unified, inextricably psycho-physical entities. We can focus on just the matter/physical entities (e.g., neural states/surface spectral reflectances/chemical properties) or just the mental/psychological entities (e.g., phenomenal character) by abstracting away from the psycho-physical state in thought, but we shouldn't get carried away into concluding that they are *really* separable.

We can see how top-down causation is involved in this metaphysics by focusing a key idea underlying [B], which distinguishes inextricabilism from 'standard' identity theories of mind. Standard identity theories hold that all causal and metaphysical explanation proceeds from

---

[22] One might be able to vindicate top-down causation by appeal to a different metaphysical picture, but we'll focus on the option that strikes us as the most plausible.

[23] By 'complete theory of the mind', I take it that Charles means a theory of the mind that answers all questions about mental phenomena (other than questions about their relation to physical phenomena). What Charles denies is that all questions about mental phenomena (other than questions about their relation to physical phenomena) can be answered solely in mental terms.

[24] By 'complete theory of the specific physical entities', I take it that Charles means a theory of those physical entities that answers all questions about them (other than questions about their relation to mental phenomena). What Charles denies is that all questions about the relevant physical entities (other than questions about their relation to mental phenomena) can be answered solely in terms of other physical entities.

the 'bottom up': everything about a mental phenomenon (e.g., what it's like to see yellow, the belief that one is seeing a yellow thing) can be fully explained in terms of the physical entities that cause and constitute it. By contrast, inextricabilism holds that with respect to some questions, causal and metaphysical explanation goes from the top down. There are facts about those physical entities that are explained by facts about the mental state, rather than the other way around. And this is why you can't give a complete theory of the physical entities involved in a mental state without mentioning that mental state (as [B] claims).

Why should we think that there are facts about physical entities that are explained by facts about mental states, rather than the other way around? The reason is that there are certain facts we can't properly explain without appeal to top-down mental causation. Let's warm up with a relatively simple example discussed by Helen Steward (2012). Given that some molecules are arranged into a wheel, we can explain (e.g.) the wheel's trajectory in terms of the interactions of these molecules. But *why* are these molecules arranged wheelwise? What caused *that* to be the case? It might be a coincidence, but the point is that if it's not, we need an explanation. Steward observes that "…from the point of view of low-level physics (say), it is just not possible to gain any understanding of how the co-occurrence of [the] different phenomena required for the production of a wheel has been provided for by the universe." (2012: 237) If we're just looking at quarks, protons, atoms, molecules, and so forth, we're not going to be able to find an answer the questions of why they came to be arranged as they are in fact arranged; all we're going to find is more arrangements of microphysical entities. And those further arrangements of microphysical entities aren't sufficient to explain why the molecules are arranged wheelwise, because we can ask the same kind of 'why' question about them—that would just amount to pushing the bump around under the explanatory rug. If we want a complete explanation why the molecules are arranged wheelwise, we need to appeal to the fact *someone wanted a wheel*, and so someone with the required skill planned to make one, and the relevant molecules happened to be in their vicinity and suitable for being shaped into a wheel. This is a plausible instance of top-down causal explanation: we explain why the molecules are arranged wheelwise in terms of facts about someone's desire for a wheel and their intention to make one.

According to Charles's interpretation of Aristotle, broadly the same kind of thing is going on with arrangements of neurons. We can ask: *why* are these neurons arranged as they are? Again, "…from the point of view of low-level physics (say), it is just not possible to gain any understanding of how the co-occurrence of [the] different phenomena required for the production of [a certain arrangement of neurons] has been provided for by the universe." (adapted from Steward 2012: 237). If we're just looking at quarks, protons, atoms, molecules, and so forth, we're not going to be able to find an answer to the questions of why those neurons are arranged as they are in fact arranged; all we're going to find is more arrangements of microphysical entities (about which we can ask the same kind of 'why' question).

The kind of top-down causation this version of the Indistinguishability-Denying strategy requires is an instance of this general phenomenon. The idea is that facts about a subject's perceptual phenomenal states can causally explain which neural states the subject goes into next. For example, let us return to S_P, who is perceiving a yellow, crescent-shaped banana with all of the full-on, in your face phenomenology that involves. In this case, a causal

consequence of this *phenomenal* fact is that S_P goes into a *neural* state that constitutes a disposition to believe that they are seeing a yellow, crescent-shaped banana. Now consider again S_H, whose BIV or Matrix-induced experience lacks full-on, in your face phenomenology (according to the Indistinguishability-Denying strategy). If we can appeal to top-down causation, we can say that the causal consequence of *this phenomenal* fact would be that the subject goes into a *different neural* state—one that *doesn't* constitute a disposition to believe that they are seeing a yellow, crescent-shaped banana.

Which exact neural states S_H would go into depends on what kind of perceptual phenomenology (if any) the neural stimulation would be sufficient for. For illustration, suppose that S_H isn't getting any visual input via normal channels (i.e., their eyes are closed if they're an embodied subject in the Matrix), and that they are in ideal conditions for reflecting on their mental state (i.e., they aren't cognitively impaired, they're able to attend to their experience, have the required concepts, and so on). If the neural stimulation is not sufficient for any perceptual phenomenology at all, S_H would go into a neural state which constitutes a disposition to believe that they're not seeing anything. But if the neural stimulation is sufficient for vivid visual imagery that nonetheless falls short of full-on, in your face phenomenal character, S_H would go into a neural state which constitutes a disposition to believe that they are enjoying detailed visual imagery. Again, which of these possibilities would actually come to pass should be regarded as an open empirical question.

In short, the Indistinguishability-Denying theorist can (i) address the challenge posed by neural determinism by rejecting that thesis, (ii) reject that thesis by appealing to top-down causation, and (iii) justify their appeal to top-down causation by embracing inextricabilism. However, this raises the question: is inextricabilism worth embracing? It would be rather ad hoc for the Indistinguishability-Denying theorist to embrace it only because it rescues their view from an unwelcome consequence. Fortunately, the potential advantages of inextricabilism are legion—for reasons of space, we'll offer a brief but tantalising sketch before concluding.

Once one adopts an inextricabilist metaphysics, longstanding philosophical problems start to look a lot more tractable. The *hard problem of consciousness* can be framed in terms of the question: how does consciousness arise from physical states and processes? The *problem of free will* (or, if you like, the *hard problem of agency*) can be framed in terms of the question: how does agency arise from physical states and processes? The hard problem of consciousness and the problem of free will can be thought of as specific instances of the *mind/body problem*, which can be framed in terms of the question: how does the mind arise from bodily states and processes? The mind/body problem (and its sub-problems) can be thought of as specific instances of the problem of *reconciling the manifest and scientific images*, which can be framed in terms of the question: how does the manifest image of the world arise from the scientific image of it? The inextricabilist's answer to all these questions is: these phenomena don't arise from physical entities at all. These questions falsely presuppose that the manifest/mind/consciousness/agency can be fully explained in terms of the microphysical entities discovered through scientific investigation. Inextricabilism gives us a way to reject the presuppositions that generate the problems while still maintaining that the manifest/mind/consciousness/agency *just are* physical phenomena.

In summary: the Indistinguishability-Denying strategy is viable in the context of a broader inextricabilist metaphysics, one which aims to dissolve problems of explaining the manifest image in terms of the scientific image by denying that causal and metaphysical explanation only goes in one direction (from the scientific to the manifest). However, that seems like a metaphysics worth seriously considering.

6. Conclusion

We have argued that there are plausible, defensible versions of both the Object-Supplying and the Indistinguishability-Denying strategies for dealing with classic BIV/Matrix style scenarios.

The specific version of the Object–Supplying strategy that we find most plausible for BIV/Matrix cases relies on the idea that elements in the Vat/Matrix computer can at least partially constitute the same kind of phenomenal character that is produced when we perceive familiar everyday objects. This may not be appealing for those Naïve Realists located at the Edenic extreme of the Spectrum. But we argued that in adopting this strategy a Naïve Realist need make no large further assumptions beyond the idea that phenomenal character is at least partially constituted/determined by the subject and the specific manner in which she is perceptually related to the object, something that *most* Naïve Realists already accept. Moreover, the general idea that the subject would be perceiving the Vat/Matrix generated environment can be motivated independently of defending Naïve Realism. However, to repeat, this strategy will not be applicable for Cosmic Swamp-Brain-style scenarios. (Though different Object-Supplying strategies – which appeal to more exotic kinds of objects – would presumably be applicable.)

The version of the Indistinguishability-Denying strategy that we find most plausible does involve embracing some heterodox metaphysical ideas concerning top-down causation. No doubt this might seem a high price to pay for those of a more orthodox persuasion. But again, the kind of inextricabilist metaphysics that we have appealed to in developing the Indistinguishability-Denying strategy can be independently motivated, quite apart from its potential benefits for NWR. And of course, the Indistinguishability-Denying strategy can apply equally well to Cosmic Swamp-Brain cases as to the classic BIV/Matrix scenarios.

Finally, it is worth emphasising that we regard it as an *open empirical question* whether or not a subject in a BIV/Matrix scenario would have an experience with full-blown perceptual phenomenal character, indistinguishable from a normal case of perception. So of course we likewise regard it as an open empirical question which of these strategies an NWR theorist should ultimately adopt. But the existence of a defensible line of response for either eventuality means that the general prospects for a non-disjunctive, 'new wave' form of Naïve Realism look, to our eyes, healthy[25].

[25] Material from this chapter was presented at the 2023 Philosophy of Mind Workshop at the University of Luxembourg. We are grateful to the audience on that occasion for their helpful questions. Special thanks to the editors of this volume for their many insightful comments, which significantly improved this paper.

References

Ali, R. 2018. 'Does Hallucinating Involve Perceiving?' *Philosophical Studies* 175: 601-627.

Allen, K. 2016. A Naïve Realist View of Colour. Oxford: Oxford University Press.

Antony, L. 2011. 'The Openness of Illusions', *Philosophical Issues* 21: 25-44.

Beck, J. 2023. 'Mundane Hallucinations and New Wave Relationalism', *Nous* 57 (2): 391-413.

Beck, O. 2019. 'Rethinking Naïve Realism', *Philosophical Studies* 176 (3): 607-33.

Bouwsma, O.K. 1949. 'Descartes' Evil Genius', *Philosophical Review* 58 (2): 141-51.

Brewer, B. 2008. 'How to Account for Illusion', in A. Haddock and F. Macpherson (eds.) Disjunctivism: Perception, Action, Knowledge. Oxford: Oxford University Press.

Butchvarov, P. 1998. Skepticism About the External World. Oxford: Oxford University Press.

Byrne, A. and Logue, H. 2008. 'Either/Or', in A. Haddock and F. Macpherson (eds.) Disjunctivism: Perception, Action, Knowledge. Oxford: Oxford University Press.

Byrne, A. and Manzotti, R. 2022. 'Hallucination and Its Objects', *Philosophical Review* 131 (3): 327-59.

Campbell, J. 1993. 'A Simple View of Color', in J. Haldane and C. Wright (eds.) Reality: Representation and Projection. Oxford: Oxford University Press.

Campbell, J. 2002. Reference and Consciousness. Oxford: Oxford University Press.

Chalmers, D. 2005. 'The Matrix as Metaphysics', in C. Grau (ed.) Philosophers Explore the Matrix. Oxford: Oxford University Press.

Chalmers, D. 2002. 'Does Conceivability Entail Possibility?', in T. Gendler & J. Hawthorne (eds.) Conceivability and Possibility. Oxford: Oxford University Press.

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

Chalmers, D. 2022. Reality+: Virtual Worlds and the Problems of Philosophy. New York: W.W. Norton.

Charles, D. 2021. The Undivided Self: Aristotle and the 'Mind-Body' Problem. Oxford: Oxford University Press.

Davidson, D. 1986. 'A Coherence Theory of Truth and Knowledge', in E. Lepore (ed.) Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson. Cambridge: Blackwell.

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

French, C. 2014. 'Naïve Realist Perspectives on Seeing Blurrily', *Ratio* 27 (4): 393-413.

French, C. 2016. 'Idiosyncratic Perception', *Philosophical Quarterly* 66 (263): 391-99.

French, C. 2018. 'Naïve Realism and Diaphaneity', *Proceedings of the Aristotelian Society* 118 (2): 149-75.

French, C. and Phillips, I. 2020. 'Austerity and Illusion', *Philosophers' Imprint* 20: 1-19.

Haddock, A. and Macpherson, F. (eds.) 2008. Disjunctivism: Perception, Action, Knowledge. Oxford: Oxford University Press.

Johnston, M. 2004. 'The Obscure Object of Hallucination', *Philosophical Studies* 120: 113-83.

Johnston, M. 2007. 'Objective Mind and the Objectivity of Our Minds', *Philosophy and Phenomenological Research* 75 (2): 233-68.

Kalderon, M.E. 2007. 'Color Pluralism', *Philosophical Review* 116 (4): 563-601.

Kalderon, M.E. 2011. 'Color Illusion', *Nous* 45: 751-75.

Logue, H. 2012. 'Why Naive Realism?', *Proceedings of the Aristotelian Society* 112: 211–237.

Logue, H. 2013. 'Experience of Natural Kind Properties: Is There Any Fact of the Matter?'

Philosophical Studies, 162: 1-12.
Logue, H. ms (a). 'An Anti-Sceptical Metaphysics of Perception'.
Logue, H. ms (b). World in Mind.
Macpherson, F. 2023. 'A New Theory of Illusion and Hallucination and Its Application to Dreams', Plenary Talk at The Science of Consciousness Conference, Taormina 2023. Available at https://www.youtube.com/watch?v=4yg2Le9szKQ
Martin, M.G.F. 2004. 'The Limits of Self-Awareness', *Philosophical Studies* 120: 37-89.
Masrour, F. 2020. 'On the Possibility of Hallucinations', *Mind* 129: 737-768.
Noë, A. 2004. Action in Perception. Cambridge MA: MIT Press.
Penfield, W. and Perot, P. 1963. 'The Brain's Record of Auditory and Visual Experience: A Final Summary and Discussion,' *Brain* 86: 595-696.
Putnam, H. 1981. Reason, Truth and History. Cambridge: Cambridge University Press.
Raleigh, T. 2014. 'A New Approach to "Perfect" Hallucinations', *Journal of Consciousness Studies* 21: 81-110.
Raleigh, T. 2021. 'Visual Acquaintance, Action and The Explanatory Gap', *Synthese* 199: 5551–5569.
Sethi, U. 2020. 'Sensible Over-Determination', *Philosophical Quarterly* 70 (280) 588-616.
Sellars, W. 1962. 'Philosophy and the Scientific Image of Man', in R. Colodny (ed.) Frontiers of Science and Philosophy. Pittsburgh, PA: University of Pittsburgh Press.
Steward, H. 2012. A Metaphysics for Freedom. Oxford: Oxford University Press.
Stoneham, T. 2008. 'A Neglected Account of Perception', *Dialectica* 62 (3): 307-22.
Thompson, E. and Cosmelli, D. 2011. 'Brain in a Vat or Body in a World? Brainbound versus Enactive Views of Experience', *Philosophical Topics* 39 (1): 163-80.