IN DEFENSE OF THE ARMCHAIR: AGAINST EMPIRICAL ARGUMENTS IN THE PHILOSOPHY OF PERCEPTION

Peter Fisher Epstein
Brandeis University

ABSTRACT: A recurring theme dominates recent philosophical debates about the nature of conscious perception: naïve realism's opponents claim that the view is directly contradicted by empirical science. I argue that, despite their current popularity, empirical arguments against naïve realism are fundamentally flawed. The non-empirical premises needed to get from empirical scientific findings to substantive philosophical conclusions are ones the naïve realist is known to reject. Even granting the contentious premises, the empirical findings do not undermine the theory, given its overall philosophical commitments. Thus, contemporary empirical research fails to supply any new argumentative force against naïve realism. I conclude that, as philosophers of mind, we would be better served spending a bit less time trying to wield empirical science as a cudgel against our opponents, and a bit more time working through the implications of each other's views – something we can accomplish perfectly well from the comfort of our armchairs.

In contemporary philosophical debates about the nature of conscious perception, there is one theme that seems to return again and again: opponents of naïve realism claim that the view is directly contradicted by empirical science. These arguments are often complex, but, looked at another way, they boil down to a simple battle cry: "Naïve realism is false because SCIENCE."

I will lay my cards on the table. I am a member of the Anti-Naïve Realism Team. I think naïve realism's externalist account of conscious perception—according to which the character of perceptual experience is determined by the objects and properties in the external world that subjects perceive, rather than by the states of those subjects' brains—is entirely wrongheaded. But, I must confess, I am also no fan of the style of argument those on my Team most often employ. Results from empirical science may subtly shift the philosophical ground on which debates about the nature of conscious experience take place. But, all too often, the steps lying between a given empirical finding and such subtle philosophical impacts are ignored or glossed over; the actual commitments of the other side are misunderstood or misrepresented; and the issue is framed as though naïve realists cling to their view only because they have somehow remained ignorant of an entire body of relevant empirical research – not because, as I believe to be closer to the truth, the two sides simply interpret such research differently, or apply a different calculus to the philosophical question of how to weigh up the costs and benefits of each theory, in light of what we (all) know empirically.

In particular, I think much more attention needs to be paid to two questions. First: what is actually *empirically* demonstrated by the scientific studies philosophers have pointed to in attacking naïve realism? And second: do the empirical results genuinely tell against naïve realism, given its overall philosophical commitments?

On the first question: Naïve realism's opponents frequently employ premises in their arguments that they characterize as empirically-demonstrated facts, when the claims in question are really the joint upshot of (a) empirical observations; and (b) philosophically-loaded principles of interpretation. In resisting these claims, naïve realists are portrayed as rejecting science—as denying, or being unaware of, (a). But naïve realists' resistance is far more often a matter of disputing (b), on non-empirical grounds. So, the elaborate empirical data is beside the point; naïve realists can happily accept everything their opponents point to with respect to (a), but reject their conclusions because

they reject the principles assumed at step (b). The real question, then, is whether those principles are philosophically justified. If they are not, the empirical findings are irrelevant, because they do not on their own support the key claims. But, as I will explain below, even if the principles are philosophically justified, the empirical data is still irrelevant, as it would then be otiose: with such principles onboard, naïve realism could be straightforwardly refuted, without any need for contemporary scientific findings.

On the second question: Naïve realism's account of perception is typically embedded within a broader philosophical picture; and, I will argue, given naïve realists' antecedent philosophical commitments—in particular, their claim that sensible qualities, like colors, are objective, irreducible, causally-efficacious properties of external objects—empirical findings that might otherwise seem problematic turn out to have ready explanations, and to fit neatly within an independently-motivated metaphysical framework. Such findings, then, do not represent novel reasons to reject naïve realism; they add no real argumentative weight to the case against the view.

Below, I will illustrate these points by exploring in detail some recent work by Adam Pautz, who claims that contemporary empirical science provides a definitive, novel argument against naïve realism.[1] The problems I will highlight are not unique to Pautz. To the contrary, Pautz is actually more careful than most in charting the philosophical logic of his arguments; he lays out explicitly how the empirical results he cites are supposed to bear on the truth or falsity of his target. But, I will suggest, Pautz's arguments still fall prey to the difficulties I sketched above.

In working through Pautz's empirical arguments in detail, I am taking up a challenge implicitly issued by Tyler Burge, who asserts that "[n]o proponent of [naïve realism] has, in a careful and informed manner, confronted the facts from psychology"[2] and that "a good deal of philosophy has proceeded with insufficient reflection on the science, or has offered unconvincing rationales for taking it to be irrelevant to philosophical problems."[3] Though I am, myself, no proponent of naïve realism, my analysis of the facts from psychology will (I hope) be careful, informed, and sufficiently reflective about the science; whether the rationales I supply are convincing is a question best left to the reader.

1. The State of Play

Naïve realism is, according to its proponents (and even some of its detractors), the default, intuitive theory of perception. But the view has long been dismissed: most philosophers have agreed with Hume's famous assessment that "the slightest philosophy" reveals naïve realism to be untenable. Traditional arguments against naïve realism, like Hume's, tended to focus on cases of illusion and hallucination, the so-called "bad cases" of perceptual experience, and they were largely of the armchair variety – they didn't appeal to cutting-edge findings in the empirical sciences in building the case against naïve realism.

The argumentative terrain looks much different today. Naïve realism has been enjoying something of a renaissance; its contemporary advocates insist that the traditional arguments against it are irredeemably flawed. In the face of this pushback, many opponents of the view have argued that naïve realism should be abandoned not because of fusty old armchair philosophy, but because cutting-edge perceptual psychology has shown it to be false. Some of these contemporary opponents agree with naïve realists' assessment that the older arguments are not decisive; instead, they assert, it is empirical science that tells definitively against the view.

---

[1] See Pautz (2017 and 2021).

[2] Burge (2005, 29).

[3] Burge (2005, 9).

Examples of this argumentative approach are numerous; I will mention a few.[4] Burge (2005) claims that empirical psychology types illusions and hallucinations together with veridical perceptions; he argues that naïve realism is incompatible with this scientifically-established taxonomy, since it is committed to the "disjunctivist" claim that the delusive "bad cases" share no fundamental kind with ordinary perception. Block (2010) suggests that empirical science has demonstrated that, holding external stimuli constant, shifts in attention, driven entirely by changes in neural processing, can alter the phenomenology of perception; he argues that naïve realism cannot account for this effect, since it holds that the phenomenology of perception is determined by external stimuli. Other features of perception that philosophers have argued are both (a) established by empirical science and (b) incompatible with naïve realism include: cases of multiple, incompatible visual representations of the same scene;[5] perceptual processing's Bayesian structure;[6] the success of "prediction error theory" as an empirical account of perception;[7] and the "cognitive penetration" of perception by subjects' belief states.[8]

In reading these arguments, one comes away with the impression that the refutation of naïve realism can simply be read off the science, and that naïve realists must be woefully, or willfully, ignorant of the basic, empirically-established facts of the very subject matter about which they theorize. Pautz, for example, writes that naïve realism "flies in the face of decades of research in psychophysics and neuroscience"[9] and that "[n]aïve realists like Campbell, Fish, and Martin neglect the scientific facts."[10] The overall narrative suggests that contemporary naïve realism is a kind of Luddite rearguard movement. This line of thought is articulated most explicitly (and most vitriolically) by Burge, who proclaims that:

> One can no longer pronounce from the armchair on the form and nature of human perception. Such issues are to be determined by empirical investigation, not by armchair pronouncements uninformed by understanding the relevant science. Human perception is the subject matter of a science.[11]

This admonition has allegedly been ignored by naïve realists, whose view, Burge tells us, "exhibits ignorance of the most elementary aims, claims, and methodology of the science of perceptual psychology."[12] Burge sets the naïve realist's waywardness in historical perspective:

> It is fairly unusual, at least since the days of Descartes and Newton, for philosophical views to be as directly at odds with scientific knowledge…. Hegel's claim that there are seven planets comes to mind.[13]

---

[4] Several of these arguments are canvassed by French and Philips (forthcoming), who argue that their version of naïve realism is in fact compatible with the relevant empirical facts.
[5] Nanay (2014).
[6] Rescorla (2015).
[7] Noordhof (2021).
[8] Cavedon-Taylor (2018).
[9] Pautz (2013, 241).
[10] Pautz (2017, 25).
[11] Burge (2011, 71).
[12] Burge (2005, 66).
[13] Burge (2005, 29). Perhaps it is apt, given that naïve realists are not in fact committed to the science-denying positions Burge attributes to them, that Hegel himself never in fact made any such claim about the planets. See Craig and Hoskin (1992).

Faced with this narrative, my posture in this paper may appear (much to my chagrin) to be a bit like that described by William F. Buckley, in his famous mission statement of the conservative: I find myself standing athwart (philosophical) history, yelling "STOP." Not only opponents of naïve realism like Burge, but even some of the view's contemporary advocates, hold that real progress on questions about the nature of perception can only come from contemporary empirical science. The trend is such that one would be hard-pressed to find a recent philosophical paper on perception that did not center on experimental psychology. But, to my mind, the real philosophical action is to be found not in the details of contemporary science, but in the to-and-fro of old-fashioned armchair debate. Like any other opponent of naïve realism, I would be thrilled to find myself supplied with firm empirical grounds for rejecting the view. Upon careful examination, however, it seems to me that the empirical facts sit perfectly well with naïve realists' philosophical picture, and thus provide no rebuttal to their armchair arguments from the comforting authority of science.

2. Basic Naive Realism

The central target of Pautz's empirical arguments[14] is a view he calls *Basic Naïve Realism* (*BNR*), which he attributes to Bill Brewer, John Campbell, and Bill Fish (among others).[15] Pautz characterizes BNR as involving three central claims:[16]

(1) A *simple or primitivist account of qualitative properties* (like colors and smells), according to which qualitative properties are real, "high-level" properties of external-world objects that are distinct from, and do not reduce to, underlying physical properties.
(2) An *externalist account of the phenomenal character of perception*, according to which the character of a given perceptual experience is constituted by the character of the external-world objects being perceived, rather than by features of the subject's brain.
(3) A simple "causal" or "selectionist" account of the brain's contribution to experience, according to which the brain's role in perception is not to serve as the grounds of perceptual phenomenology, but instead to act as a kind of "enabling condition," allowing the subject to be "acquainted" with certain features of the external world—namely, the simple, qualitative properties mentioned in (1)—which do ground the subject's phenomenology.[17]

Pautz centers his case against BNR on a series of empirical studies suggesting that the structure of conscious experiences—for instance, that the experience of blue is more similar to the experience of purple than it is to the experience of green—correlates more closely with the structure of the brain processes associated with such experiences than with the structure of the properties—such as

[14] While I focus on Pautz's specific arguments in much of the paper, I take many of the points I raise to be applicable quite broadly.
[15] See Pautz (2017 and 2021).
[16] Pautz (2017 and 2021) contrasts BNR with more complex versions of naïve realism (put forth by Keith Allen, Ori Beck, Craig French, Ian Phillips, and Heather Logue), which have been developed (at least in part) to respond to empirically-grounded arguments against the basic version of the view. One upshot of my arguments will be that, to the extent these more complex versions of naïve realism represent a retreat in the face of alleged conflicts between the original "basic" version and empirical science, such a retreat is unwarranted. I do not address Pautz's arguments against the more complex versions of naïve realism because those arguments are, by and large, non-empirical in nature.
[17] How exactly to understand the brain's "enabling role" will be explored in detail in Section 6.

spectral reflectance profiles—of the external-world objects that are consciously perceived. We can summarize the two key empirical premises of Pautz's arguments as follows:

Good Internal Correlation (GIC): Experience is closely correlated with neural states.

Bad External Correlation (BEC): Experience is not closely correlated with the properties of the external objects perceived.

Pautz builds on these findings to develop a series of specific arguments against BNR, which I consider below. But it is already apparent how the empirical data might seem to undermine the view: given GIC and BEC, it seems reasonable to suppose that it is the well-correlated brain processes, and not the poorly-correlated external-world properties, that ground the character of experience.

I want to explore two questions about the empirical findings Pautz cites as establishing the crucial premises of his arguments. First, do those findings really support GIC and BEC? I will suggest that this is far from clear. The findings are, on their own, insufficient to establish GIC and BEC; an implicit, non-empirical premise—a "bridge principle" linking empirically observable behavior to unobservable conscious experience—is needed to reach those conclusions. But this bridge principle is one that naïve realists, in responding to more traditional arguments against their view, are already committed to rejecting; and so, it can hardly be utilized as a premise in a new argument against naïve realism. Furthermore, the data Pautz points to, even granting the controversial premise, do not establish GIC and BEC nearly as conclusively as Pautz suggests.

The second question I want to explore is whether the findings of GIC and BEC (assuming, now, for the sake of argument, that they are empirically sound) really tell against naïve realism. I will argue that, when we consider the overall shape of the naïve realist's metaphysical picture and the antecedent philosophical commitments of the view, the empirical findings Pautz points to can be accommodated straightforwardly, and so they provide no real argumentative force against Pautz's targets.

3. A Hidden, Non-Empirical Premise

The empirical evidence Pautz points to in support of GIC and BEC comes from a variety of studies in the psychology and neuroscience of perception, across visual, olfactory, and auditory modalities. The various examples all conform to the same general pattern. We take a set of what we might call "perceptual observation events" ($\{O_1, O_2,\ldots, O_n\}$)—e.g., a subject's perception of a series of colors—where each observation event involves a corresponding neural property ($\{N_1, N_2,\ldots, N_n\}$); an external-world property ($\{W_1, W_2,\ldots, W_n\}$); and a property of conscious experience ($\{C_1, C_2,\ldots, C_n\}$). We then discover, through contemporary scientific methods, that the similarity ordering, or some other structural feature, of the C-variables matches that of the N-variables but differs from that of the W-variables.

Pautz (2017 and 2021) provides an illustrative example from the domain of color vision:

O_1: viewing a blue ball
O_2: viewing a purple grape
O_3: viewing a green leaf

N_1: neural processing while perceiving the ball
N_2: neural processing while perceiving the grape
N_3: neural processing while perceiving the leaf

W_1: spectral reflectance of the ball
W_2: spectral reflectance of the grape
W_3: spectral reflectance of the leaf

C_1: experience of blue
C_2: experience of purple
C_3: experience of green

According to Pautz, what we discover empirically is that, for the consciousness-variables, C_1 is more similar to C_2 than it is to C_3; for the neural-variables, in line with the C-variables, N_1 is more similar to N_2 than to N_3; but, by contrast, for the external-world-variables, W_1 is no more similar to W_2 than it is to W_3. That is: the structure of the C-variables is well-correlated with that of the N-variables, but not with that of the W-variables; we have GIC and BEC.

As I noted, Pautz characterizes his claims about the correlations among these variables as deriving from empirical science. This suggests that the claims are based on experimental observation of the relevant phenomena. In particular, it suggests that we experimentally observed, in each of O_1, O_2, and O_3, the N-variables (neural processing of subjects); the W-variables (external-world properties of the objects perceived); and the C-variables (conscious experiences of subjects). We then noted which were correlated and which were not.

The relevant empirical observations pertaining to N-variables and W-variables are reasonably straightforward: Pautz includes images of fMRI scans and spectral reflectance curves, representing standard methods of making empirical observations of N- and W-variables. But when it comes to C-variables, there is a notorious difficulty in studying the phenomenon in question—conscious experience—by means of empirical observation. For the conscious experience of subjects is not directly observable in a lab; it is not obvious how, in principle, it could be empirically observed at all. So, as Chalmers highlighted long ago, what we in practice end up doing, in trying to investigate consciousness scientifically, is to use an observable proxy.[18] Typically, this proxy will be a set of behaviors that we take to be tightly linked to experience; most commonly, we rely on subjects' introspective reports. For instance, if subjects report that their experience of blue is more similar to their experience of purple than to their experience of green, we take that be a reliable indication that their experiences themselves in fact have such a similarity structure.

This may seem pedantic. But, in evaluating philosophers' claims that naïve realism is refuted by empirical science, it's important to be clear about what we have evidence of, if we stick to the *empirical* data. The reason this is so important in the present context is that, as I am about to explain, moving beyond what is strictly empirical in many allegedly empirical arguments against naïve realism requires a non-empirical premise that naïve realists are antecedently committed to rejecting. And so, these arguments, absent further elaboration, beg important questions against their target.

Given the impossibility of empirically observing consciousness itself, the phenomena of GIC and BEC, if they are to be directly supported by empirical science, can only concern correlations between subjects' behavior and their brain processes. And so, when citing empirical evidence, we should replace claims about correlations involving C-variables with claims about correlations involving *behavioral variables* (call these *B-variables*).

Interestingly, one empirical study that Pautz cites in support of the claim of GIC for color vision is quite explicit on this score. Here's how that study reports its central finding:

[18] See Chalmers (1996).

> *Behavioral judgments* of the similarity between colors closely match the similarities between the neural responses to these colors.[19]

Pautz, however, frames such empirical findings in terms of correlations involving C-variables:

> Neuroscience has shown that neural similarity is the only accurate predictor of *qualitative similarity*.[20]

So, there is a gap in the argument. This is not a point about Pautz's specific methodology; it applies to any attempt to employ empirical data in an argument against naïve realism. In order to get from empirically-observed correlations between behavior and other phenomena (brain processes, or external-world properties) to claims about what is well correlated with *experience*, we need to take a step beyond the empirical data. We need a principle linking experience to observed behavior—what Chalmers (1996) calls a "bridge principle." Such bridge principles are not themselves empirically derived: we don't (and, for principled reasons, seemingly can't) conduct an experiment to determine empirically whether consciousness is associated with various behavioral (or otherwise functional) phenomena, for the simple reason that consciousness is not observable, and any principle we appeal to in order to make inferences from empirical observations would be precisely the kind of principle at issue.[21]

Thus, the move from experimental observations to conclusions about consciousness that is typical of empirical arguments against naïve realism is best understood as revealing those arguments' reliance on a hidden premise—a bridge principle—of the following kind:

> (BP) Subjects' behaviors (including their introspective reports) are reliable indicators of the character of their experiences.

Without this kind of bridge principle, the empirical data—which, again, only concern observable behaviors, like verbal reports—would have no bearing on naïve realism, which is a theory of conscious experience. BP is needed to bridge the logical gap in empirical arguments against naïve realism.

But it is well known that naïve realists typically reject this non-empirical hidden premise; denying such a link between introspective reports and the character of experience is central to their standard "disjunctivist" response to the traditional armchair argument from hallucination. In a hallucination, a subject's behaviors, including their introspective reports about the character of their experiences, will match the behaviors they display when they are perceiving veridically.[22] And so, the standard anti-

---

[19] Bohon et al. (2016, 18 (emphasis added)), cited in Pautz (2021, 158).

[20] Pautz (2017, 25 (emphasis added)).

[21] This point, originally highlighted by Chalmers (1996), about the reliance of empirical arguments in the philosophy of mind on a hidden, non-empirical premise, has recently been emphasized by Wayne Wu (2020), who has come to think that some of his own earlier empirically-grounded arguments implicitly relied on a non-empirical premise concerning the reliability of introspection.

[22] At a minimum, this will be true of what Martin (2004) calls "causally-matching hallucinations": hallucinations induced by artificially putting a subject's brain into the very same state it is in when the subject has a veridical perception. Recently, some naïve realists have attempted to escape the traditional argument, without rejecting BP, by denying that such causally-matching hallucinations are even so much as possible (see, e.g., Masrour (2020)). Such a move does not seem very promising to me—it seems all too clear that causally-matching hallucinations are indeed possible—but I will not pursue the issue here.

naïve realist argument goes, we can conclude that the subject's experience in the case of hallucination—when no external-world object is being perceived—is the same as in the case of veridical perception. But then we have two cases involving very different external-world objects—in one case, there is no relevant object at all—which nonetheless have the same character; and this would seem to undermine the naïve realist's claim that the character of experience is determined by the objects perceived.

This traditional argument against naïve realism requires no knowledge of the details of contemporary empirical psychology. It does, however, rely crucially on a premise like BP; this is what allows us to move from a claim about a subject's introspective reports in a case of hallucination to a claim about the character of that subject's experience. And it is this crucial premise that naïve realists standardly reject: they hold that, in such cases, subjects' introspective reports do not reflect the true character of their experiences. Instead, according to the naïve realist, such reports are the expression of a kind of *epistemic* failure: subjects in such cases are unable to recognize that their experiences do not have the same character as their veridical experiences.

The point of rehearsing this well-worn dialectic is that, if we help ourselves to a bridge principle like BP, we can employ a very traditional argument from hallucination to refute naïve realism, without the need for the kinds of contemporary scientific findings that Pautz and others claim to be the real problem for the view. But naïve realists respond by simply rejecting BP: we can't (they insist) just read off conclusions about unobservable conscious experience from empirical observations of behavior (including introspective reports).

So here is how things stand. Pautz cites various empirical results as demonstrating that there is good correlation between experience and brain processes, and bad correlation between experience and external-world properties. But, upon more careful reflection, we see that what these studies in fact demonstrate, if we stick to empirical observations, is that brain processes are (while external-world properties are not) well-correlated with subjects' behaviors. We thus need an additional, non-empirical premise—a bridge principle linking behavior to experience—in order to draw any conclusions about experience from these empirical studies. But the kind of bridge principle that seems needed—something like BP—is exactly the kind of principle that the naïve realist is already committed to rejecting, in formulating a response to more traditional arguments against the view. Thus, the empirical data fail to supply any new argumentative force against the naïve realist.[23]

But perhaps this is too quick. For there is one key difference between some contemporary empirical arguments against naïve realism, including Pautz's, and the more traditional argument I

[23] Though I cannot here address all of the many empirical arguments directed against naïve realism in the literature, it seems to me that most of the arguments mentioned above (Section 2) are likewise undermined by reliance on an implicit bridge principle—one naïve realists are antecedently committed to rejecting—which is required to get from empirical findings to conclusions about conscious experience. Burge's argument, for example, centers on perceptual psychologists' categorizing hallucinations along with veridical experiences as of a single type – a categorization that, to the extent it is empirically-grounded, can only be based on observable criteria. Burge's conclusion—namely, that naïve realists are wrong to insist that hallucinations and veridical experiences are distinct types of *conscious* states—must, then, rely on a bridge principle of the kind described above. Curiously, Burge himself sometimes acknowledges that the empirical work he cites is essentially silent on questions about consciousness. He writes, "The psychological theories that I have discussed do not attempt to explain consciousness. There is, currently, no scientific theory of consciousness" (Burge 2005, 47). And yet, Burge insists that the empirical work he points to definitively undermines naïve realism – a proposed account of consciousness. Assessing whether there is a way to reconcile Burge's seemingly conflicting claims falls outside the scope of the present paper; Campbell (2010) argues that Burge's arguments are indeed undermined by the logical gap between empirical work in cognitive science and conclusions about consciousness.

sketched above. These contemporary arguments focus exclusively on cases of normal perception, whereas older arguments against naïve realism tended to focus on cases of illusion or hallucination. And, one might suggest, naïve realism's rejection of BP is not totally implausible in cases of hallucination (or other "bad cases"); in such cases, it does not seem absurd to suggest that subjects suffer a kind of epistemic failure in reporting on their experiences.[24] But the same move is much less palatable if the cases in question are cases of ordinary perception. So, consider a modified version of BP:

> (BP_O): *In cases of ordinary perception*, subjects' behaviors (including their introspective reports) are reliable indicators of the character of their experiences.

Surely, one might think, if we restrict our attention to central cases of ordinary perception, we can assume that subjects report accurately on the character and structure of their own experiences. Consider, again, the case of color vision. Here, the observed behavior is simply subjects' judgments of the similarity ordering of their color experiences. If naïve realism can only be salvaged by denying that, e.g., the experience of blue really is more like the experience of purple than the experience of green, then the view has paid a heavy price indeed.

So, let us grant that, at least in cases like ordinary human color perception, the similarity structure of the relevant experiences corresponds to that reported by subjects: that is, for such cases, everyone, naïve realists included, should accept (BP_O), and so we can safely move from empirical findings of correlations involving behavioral variables to conclusions about experience. Having granted this point, how does the empirical argument against naïve realism fare?

4. Do the empirical data support GIC and BEC?

The first question to ask at this stage is whether empirical science actually presents us with any data that are potentially problematic for naïve realism. In the case of Pautz's argument, this amounts to the question of whether the data support the twin claims of GIC and BEC. These claims hinge crucially on how we evaluate the similarity structures of the relevant properties – the C-, N-, and W-variables. Let us begin with the alleged similarity structure of the W-variables. In summing up his key claim for the case of color perception, Pautz writes that empirical science has revealed that:

> the reflectance of the [blue] sphere is *not* more like the reflectance of the [purple] grape than the reflectance of the [green] leaf.[25]

What is Pautz's justification for this claim? He points us to three charts, depicting the spectral reflectance curves of the three objects in question (see Fig. 1).[26]

[24] Wu (2020) argues that we should carefully assess reasons for thinking that introspection is likely to be (un)reliable in specific kinds of experimental paradigms. One natural application of Wu's suggestion would be not to take for granted the reliability of introspective reports in cases of hallucination.
[25] Pautz (2021, 157).
[26] Pautz (2021, 158).

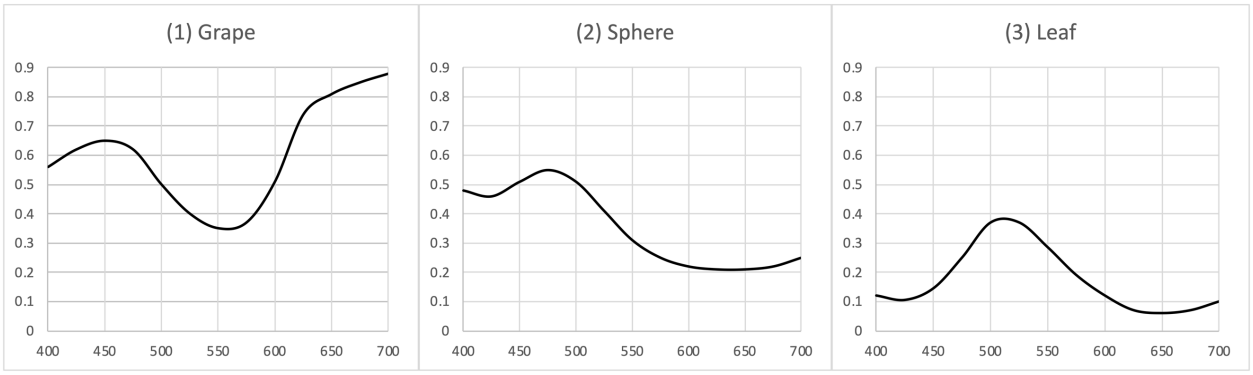

FIG. 1: Spectral reflectance curves associated with (1) purple grape, (2) blue sphere, and (3) green leaf. Reflectance proportion (y-axis) is shown across 400-700nm range (x-axis). Adapted from Pautz (2021, 158).

Pautz's point, I take it, is that it should simply be obvious from a visual inspection of these charts that the similarity structure is as he describes: (1) and (2) are no more similar to each other than they are to (3).

Perhaps this claim is based on the overall shapes of the curves, with (2) and (3) sharing a "dip-hump-dip" structure that (1) lacks; or perhaps the thought is that (1) and (3) seem to be inverted versions of each other, leaving (2) as the less similar curve. But I must confess that it seems equally plausible to me to hold that (1) and (2) are indeed more similar to each other than either is to (3). (1) and (2) both reach a peak before 500nm, while (3) does not; in addition, (1) and (2) both include substantial curve-segments above 0.4 on the y-axis (reflectance proportion), while (3) has none. Looked at in these ways, the structure of the external stimuli would indeed match the structure of the associated experiences: the physical properties of the blue ball are, by these measures, more similar to those of the purple grape than they are to those of the green leaf; and so, we actually get *good* external correlation (GEC) in the case of color vision.

My point is not that these latter measures of similarity are more legitimate than the ones Pautz seems to have in mind when he claims that the similarity structure of the curves fails to match that of experience (nor am I suggesting that Pautz's argument is built on these three charts alone). My real point is that similarity, in general, is always similarity *along some dimension*, and it is not always straightforward to determine what the relevant dimension of similarity is. Even when we are just "eyeballing" curves, we can gauge similarity in many ways. And so, we might think that the external-world stimuli in Pautz's empirical examples only seem to lack the right similarity structure because we have failed to evaluate the stimuli along the right dimension of similarity – the dimension that, the naïve realist will say, explains the similarity structure of the associated experiences.

One way to restrict the potentially infinite search for an appropriate dimension of similarity would be to insist that we only count as having found a case of good correlation if the dimension along which that correlation is measured is a *natural* one. Pautz seems to have something like this in mind when he writes that, "[b]y any natural measure, it is not the case that the reflectance of the ball objectively resembles the reflectance of the grape more than the reflectance of the leaf."[27] But I take the types of measures I mentioned above, on which the reflectance of the blue ball is indeed more similar to that of the purple grape than to that of the green leaf, to show that this is just false – those measures seem perfectly natural.

[27] Pautz (2017, 25).

This point about needing to pin down the relevant dimension of similarity is not merely an idle concern. For consider the kinds of findings Pautz points to in support of his claim about GIC, this time with respect to olfaction:

> The group-averaged imaging and perceptual datasets were each projected onto a common three-dimensional space using MDS [multidimensional scaling], indicating that the imaging maps of PPC [posterior parietal cortex] linear correlations closely overlapped with the perceptual maps of odor quality similarity.[28]

The dimension of variation along which these studies assess the similarity of subjects' neural properties is, as can be seen in the above passage, extremely complex. Here is a visual representation of these kinds of data sets (where the data has already been run through a series of algorithms to produce a color-coded image):[29]

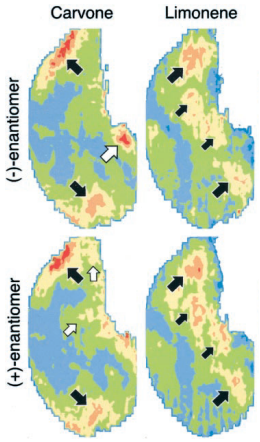

FIG. 2: Neural response to enantiomers of carvone and limonene in rat brains. From Linster et al. (2001). (Copyright 2001 Society for Neuroscience.)

This graphic depicts neural activity in rat brains while four different stimuli are presented. The two images on the left depict neural responses to enantiomers (subtly chemically different versions of a single kind of molecule) of carvone; the two on the right depict neural responses to enantiomers of limonene. The study reports that, *when a complex multidimensional scaling analysis is performed*, the neural responses to the enantiomers of carvone register as very different, while the neural responses to the enantiomers of limonene look quite similar. But suppose we simply "eyeball" the images, as we did in the case of the spectral reflectance curves, and assess the similarity structure visually. Would we think it obvious that the two neural responses on the left were very different from each other, while the two on the right were nearly identical? Clearly not. The similarity structure the study finds in the neural activity here—the one that leads Pautz to claim that there is GIC—is only detectable when the data are coded using an extremely elaborate set of algorithms. That is, we needed to do some very fancy footwork to find a dimension of similarity along which the neural activity correlates well with the structure of the associated experiences.

Above, I noted that Pautz claims that no natural measure of similarity will reveal GEC; but, given the kinds of extremely complex processing of empirical data employed in the neuroscientific research just cited, it seems questionable whether the dimensions of similarity along which we find

[28] From Howard et al. (2009), cited in Pautz (2021).
[29] From Linster et al. (2001), cited in Pautz (2011).

GIC are themselves natural. Indeed, one study of olfaction cited by Pautz notes that "conventional, univariate analysis of the fMRI signal in PPC averaged across the set of voxels for each subject failed to reveal substantial differences between the odorants."[30] So, a simple, natural way of assessing the similarity of the neural responses to different odorants failed to reveal a similarity structure correlated with experience. The study goes on to note that, once more sophisticated techniques were used to generate a more complex, multivariable measure of similarity, "the analysis revealed distributed, odor-specific overlapping ensemble activity patterns in PPC and orbitofrontal cortex" that did show good correlation to olfactory experience. But whether this latter analysis counts as comparing neural responses along a *natural* dimension of similarity is quite dubious: at the physical level, a pattern that is revealed only by applying complex multivariable analysis and lumping together the activity of many spatially non-continuous cortical areas is not an especially natural one.[31]

By parity of analysis, we should only conclude that the similarity structure of the external stimuli fails to match that of the associated experiences if we somehow determine that no similarly elaborate multidimensional scaling analysis will reveal a suitable similarity structure in the W-variables. But there is little reason to think this is the case. Neuroscientists themselves are unlikely to perform an exhaustive search for correlations between experience and highly-complex, non-natural external variables; they do not typically subject external variables to the kinds of multidimensional-scaling analysis that they apply to neural data. One study of olfaction cited by Pautz does claim that "no systematic description of how these [external molecular] characteristics relate to particular odor qualities has been developed."[32] But this claim does not even seem consistent with the current state of olfactory science: Keller et al. (2017) report that they were able to use machine learning to generate models of odor qualities that "predict the perceptual qualities of virtually any molecule with high accuracy and also reverse-engineer the smell of a molecule" based on its structure. These machine-learning models are relatively new and still far from perfect, but they suggest that there is indeed a way to analyze external odor stimuli that will reveal a (highly-complex) dimension of similarity along which those stimuli correlate with the structure of experience. And, in a sense, we already knew that there must be at least one such dimension of similarity: namely, those stimuli's dispositions to elicit certain responses in human subjects. In the case of color vision, for instance, we know that humans react to blue things and purple things more similarly than to green things; this is the result of highly complex processing of the spectral reflectances of those objects by our visual systems; and that highly complex processing could, in principle, be codified into a complex dimension of similarity along which we'd see the expected similarity structure in the reflectance data. This dimension of similarity might appear very non-natural; but what I flagged above about the analyses used by neuroscientists to detect GIC is that those analyses likewise involve non-natural dimensions of similarity.[33]

---

[30] Margot (2009), cited in Pautz (2021).

[31] Another study cited by Pautz (Brouwer and Heeger 2009, cited in Pautz 2021) notes that the way that spatially non-continuous pockets of neural activity in V4 must be lumped together, in order to detect a pattern that correlates well with color experience, does not even seem to conform to the traditional "columnar" architecture of the brain. This, again, does not seem like a particularly natural dimension of neural similarity.

[32] Cowart and Rosin (2001), cited in Pautz (2021).

[33] Pautz marshals evidence from a wide range of perceptual domains in defending his claim of GIC/BEC. Here, I focus primarily on the domain of olfaction (and, to a lesser extent, color vision). This is for two reason: first, addressing in detail all of the many studies Pautz cites would fall outside the scope of a single paper; and second, olfaction is typically seen as the domain in which claims of GIC/BEC are hardest for externalists to resist – so, if the empirical case for GIC/BEC in that domain is less secure than is typically thought, that would represent a significant challenge to the received understanding of the empirical facts.

The point here, again, is not unique to Pautz; it is, rather, something to keep in mind when evaluating any empirically-based argument in the philosophy of perception. Whether particular scientific studies, or domains of inquiry, are likely to find non-obvious correlations between experience and internal (or external) variables depends in large part on the researchers' motivations and presuppositions. Neuroscientists operate under a broadly internalist assumption that experience is to be explained, somehow, by brain activity; and so, if they cannot find an obvious, natural dimension of similarity in the neural data that correlates with experience, they will cut and splice that data in myriad ways in order to locate a non-natural dimension that does. Machine learning researchers, by contrast, are looking for the aspect of external stimuli that our perceptual systems use to guide our responses, and so they will sift through the external properties and find (perhaps very complex, non-natural) patterns that can predict the target behaviors. Neither group is particularly concerned with conceptual questions about what counts as "natural" similarity, and neither group is evaluating competing philosophical views about whether internal or external factors ground conscious experience.

So, here is where our investigation of the science has brought us. Pautz makes two key empirical claims: GIC and BEC. We can assess this pair of claims by either (a) demanding that the correlations in question be natural ones; or (b) allowing correlation along any dimension of similarity, including non-natural ones, to qualify. If we take option (a), then the data don't support GIC. The only dimensions of similarity along which scientists have empirically detected GIC are extremely complex, non-natural ones; so, if we don't count such non-natural similarities as legitimate, we don't get GIC. But if we instead take option (b), then there is no reason to accept the claim of BEC. All that the empirical data show is that experience does not correlate well with external properties along simple, natural dimensions of similarity. If we are allowed to play around with the data for external-world properties the way neuroscientists do in detecting neural patterns that correlate with experience—if we proceed the way machine-learning researchers do, when building models of odorants that correlate with subjects' responses—there is every reason to think that we will indeed find a suitable dimension of external similarity as well.

In the following sections, I will put aside these concerns about the strength of the empirical evidence for GIC/BEC in order to assess a different kind of question about the empirical case against naïve realism. I will look at a series of arguments Pautz makes on the basis of GIC/BEC and ask: even granting that the empirical support for those claims is sound, does the empirical evidence contribute any additional reasons to reject naïve realism, beyond those supplied by more traditional armchair arguments?

5. The Internal Dependence Argument

I begin by considering Pautz's most well-known argument against BNR, the *Internal Dependence Argument* (*IDA*).[34] The argument centers on a series of hypothetical cases of what Pautz calls "coincidental variation" involving a normal human perceiver and a "twin," who is a member of a hypothetical human-like species with subtle neural differences from actual humans. We can consider a case of coincidental variation involving a human subject, Ap, and his twin, Tap, whose species has evolved visual receptors just like ours but with different post-receptoral processing, such that Tap sometimes goes into different neural states in response to the same external color stimuli. Specifically, we can suppose that, although Ap and Tap are both members of species evolved to track the same external colors, Tap's neural responses to green and purple are inverted relative to

[34] See Pautz (2006, 2011, and 2021).

Ap's (Tap's response to blue, by contrast, matches Ap's). In terms of the N-variables we spelled out earlier:

Ap's neural state while perceiving blue = N_1 = Tap's neural state while perceiving blue
Ap's neural state while perceiving purple = N_2 = Tap's neural state while perceiving green
Ap's neural state while perceiving green = N_3 = Tap's neural state while perceiving purple

The key question Pautz presses is whether, when Tap perceives a green leaf and is in N_2, Tap's experience will be the same kind of experience Ap has when he perceives the same green leaf and is in N_3 (call this verdict "Same Experiences") or a different kind of experience (call this verdict "Different Experiences"). Pautz's IDA against BNR runs as follows:

1) BNR implies Same Experiences.
2) Different Experiences.
3) Therefore, BNR is false.

Premise (1) is a fairly direct consequence of the way Pautz defines BNR. BNR says that the character of ordinary perceptual experiences is determined by the external properties being perceived. Ap and Tap are perceiving the same external property—greenness—in normal circumstances, in the way that their respective species have evolved to (for Ap, by going into N_3, for Tap, by going into N_2). So Ap and Tap must be having the same kind of experience.[35]

In order to resist the argument, then, the BNRist will have to reject (2). Interestingly, Pautz gives two distinct arguments in support of this premise—one an empirical argument based on GIC/BEC, the other a non-empirical argument. My goal in this section is not to argue that (2) should be rejected. Instead, I will suggest that, while the non-empirical argument Pautz gives in defense of (2) puts real pressure on the BNRist, the version involving empirical evidence adds no additional argumentative weight. *A priori* reflection on cases of coincidental variation reveals that BNR is committed to certain claims about bridge principles in such scenarios that might seem implausible. But, once we recognize, from the armchair, that the BNRist must bite those bullets, the empirical evidence supporting GIC/BEC is, dialectically, otiose.

5.1. The Non-Empirical Version of IDA

Here is the non-empirical argument for Different Experiences:

(i) If two subjects have the same kind of experience, then they have the same behavioral dispositions.
(ii) Tap does not have the same behavioral dispositions as Ap when perceiving green.
(iii) Therefore, Tap's experience when perceiving green is different from Ap's.

On (ii): Recall that when Tap perceives green, he is in the same internal state (N_2) that Ap is in when Ap perceives purple. Granting a minimal premise of the locality of physical causation, this implies that Tap's behaviors when he perceives green will be the same as Ap's when Ap perceives purple: the immediate physical cause of Ap's behaviors is the neural state N_2, so Tap, when he is likewise in N_2, will display the same behaviors. For example, when Ap is in N_2, he describes the objects he is perceiving—purple objects—as similar to blue objects (which he perceives by going

[35] Below, I will consider versions of naïve realism that, unlike BNR, deny Same Experiences.

into N_1). That behavior is driven by the physical state of Ap's brain. So Tap, when he is in N2, will describe the objects he is perceiving—green objects—as similar to blue objects (which Tap also perceives by going into N_1). So, Tap's sorting behaviors will be different from Ap's: whereas Ap sorts purple objects with blue objects, Tap will sort green objects with blue objects.

We can thus conclude that Tap and Ap will indeed exhibit different behavioral dispositions when they perceive green. In order to move from this claim about behavioral differences to the conclusion of Different Experiences, we need a way of connecting behavior to experience—that is, we need a bridge principle. In laying out the non-empirical version of IDA, Pautz invokes his favored bridge principle, which he labels "The Behavior-Experience Link":

> If two actual or possible individuals have qualitatively identical color experiences, then they have the same color-related behavioral dispositions. If two actual or possible individuals have suitably different color experiences…, then they have different color-related behavioral dispositions.[36]

Given this bridge principle (expressed in (i)), the fact that Ap and Tap have different color-related behavioral dispositions when perceiving green implies that they also have different color experiences.

Note that this argument does not appeal to any empirical evidence from contemporary brain science. Instead, it relies solely on a minimal locality of (physical) causation principle, along with a bridge principle, which, as explained above, is not itself derived from empirical evidence.

In the face of this *a priori* argument, how might the BNRist who denies Different Experiences respond? Rejecting (ii) seems implausible: all (ii) requires is that we accept that two subjects who are in identical internal neural states will be caused by those states to behave in the same way—e.g., if N_2 causes Ap to utter sounds and move his arms in certain ways (when describing or sorting colored objects), N_2 will cause Ap's twin Tap to utter the same kinds of sounds and make the same kinds of movements. Denying these kinds of claims would require an extremely radical rejection of established principles governing how physical causation works.[37]

So, in order to resist the non-empirical version of IDA, BNRists will have to reject (i): they will reject Pautz's bridge principle, or at least deny that it is applicable to Tap's case. We've already seen that naïve realists typically do reject bridge principles for certain kinds of cases—namely, the so-called "bad cases" of illusion and hallucination. In those cases, we noted, naïve realists are committed to denying that conclusions about experience can be read off of subjects' behaviors. Here, BNRists will have to make a similar move: they will say that, although Tap behaves as though he experiences green things to be similar to blue things, in fact his experience when he perceives green things is the same as Ap's—it is, despite his sorting behaviors and introspective reports to the contrary, actually *less* similar to his experience of blue things than is his experience of purple things.

Above, I suggested that an across-the-board rejection of bridge principles linking behavior to experience was implausible, and that naïve realists should accept BP_O: they should acknowledge that, in cases of ordinary perception, such as when we perceive colors in normal conditions, our introspective judgments genuinely track the character of our experiences. What the non-empirical version of Pautz's IDA shows is that the BNRist will have to say that BP_O does not apply to hypothetical subjects like Tap. This is, it seems to me, a genuine cost, over and above the cost naïve

---

[36] Pautz (2006, 219).

[37] Such principles are not strictly *a priori*, but they do not rely on the kinds of empirical work Pautz and others claim to be the real problem for BNR. We might follow Pautz (2006, 220) in calling this version of IDA "relatively *a priori*."

realists already accept in denying the intuitive idea that behaviors in bad cases track subjects' underlying experiences. Tap's perception of green objects is not, in any obvious sense, defective: as Pautz emphasizes, Tap goes into N_2 because that is the neural state his species has evolved to go into when perceiving green objects.[38] So, even more than in the case of an electrode-induced hallucination or a dream, it seems surprising that Tap's behaviors would diverge from the character of his experience. Accepting that Tap's case is not, in the relevant sense, a case of "ordinary" perception is a real bullet that the non-empirical version of Pautz's IDA forces the BNRist to bite.

How might the BNRist defend biting this bullet? Here is the kind of line they would likely take:

> "Tap's case is a merely hypothetical one, involving a subject with different cognitive structures from any we've actually encountered. Note that we had to go to non-actual cases to find this kind of thing; we haven't empirically observed any *actual* cases of coincidental variation.[39] So the fact that we endorse a link between behavior and experience in our own case—i.e., we accept BP_O—doesn't mean we can assume the same holds for Tap. Perhaps the result of the non-standard wiring we've saddled Tap with is to make the true character of his experience cognitively inaccessible to him; we've disrupted the link between the character of experience and the behaviors normally associated with experience, like introspective report."

I do not wish to defend this line of thought (indeed, as a member of the Anti-Naïve Realism Team, I find it quite implausible). I simply want to emphasize that this is the kind of position BNRists are committed to in light of the *non-empirical* version of Pautz's IDA. They will have to hold that the perceptual experiences of hypothetical coincidentally-varied subjects like Tap, even in conditions that are normal for those subjects, do not count as "good cases": in such cases, just as in hallucination, there is a radical divergence between subjects' underlying experiences and their associated behaviors.

My claim that the empirical evidence of GIC/BEC is, with respect to Pautz's IDA, dialectically otiose now amounts to the following: such evidence has no argumentative force against someone who is already prepared to bite the bullets involved in the position just described—bullets that, again, were fired from the armchair.

5.1. The Empirical Version of IDA

I turn now to the empirical version of Pautz's IDA, which appeals to the empirical evidence of GIC/BEC to support the claim of Different Experiences. To see how this version works, we can summarize Ap's and Tap's perceptual situations, in a scenario in which each is perceiving a green leaf, as follows:

	Ap	Tap	Same/Different?
N-Variable	N_3	N_2	Different
W-Variable	Reflectance of leaf	Reflectance of leaf	Same
C-Variable	C_3	???	???

[38] In Pautz's terms, N_2 bears the "optimal cause" relation to green for Tap: it is the neural state Tap's species would be caused to go into by green in (perhaps evolutionarily-specified) optimal conditions (Pautz 2006, 221).

[39] This is something Pautz himself acknowledges (see Pautz 2013, 252-253).

The key question concerns the value of Tap's C-variable—the character of his conscious experience—when he sees a green leaf. Will it be the same as Ap's, or different? If C-variables are determined by W-variables, as the BNRist holds, then Tap will have the same kind of experience as Ap; if, by contrast, they are determined by N-variables, Tap will have a different experience.

Pautz argues that we do not simply face a stalemate of competing "intuitions" about which of these verdicts to accept because empirical evidence of GIC/BEC supports Different Experiences. This is where Pautz sees empirical findings doing meaningful philosophical work. He writes that it is an "empirical fact that neural similarity is the only good predictor of similarity in color experiences. So, given the empirical facts, the only reasonable verdict is that [Ap] and [Tap] have different experiences."[40] That is: our empirical evidence suggests that, for humans and other animals we've studied, N-variables correlate better with experience than do W-variables. And so, in Tap's case, we should likewise expect the character of experience to line up with N-variables, rather than W-variables. This gives us the conclusion of Different Experiences.

Crucially, this argument requires applying findings from empirical studies of our own perceptual episodes to reach a conclusion about the character of Tap's experience. But that is precisely what we cannot do in this argumentative context, given the BNRist's response to the non-empirical version of IDA outlined above. As explained in Section 3, empirical studies only count as providing evidence about experience to the extent that we're willing to accept a bridge principle linking observed behaviors to experience. The bridge principle naïve realists are willing to accept is BP_O: they hold that empirical evidence only bears on questions of experience in "ordinary" cases of perception. Furthermore, as just explained, the *a priori* version of IDA forces the BNRist to deny that Tap's experiences are ordinary, in the relevant sense. Against this background, then, the empirical evidence Pautz cites is dialectically powerless: it allows us to reach a verdict about Tap's experience only by presupposing that those experiences are ones to which BP_O is applicable—the very claim the BNRist has already been forced to reject, on *a priori* grounds.

The most the empirical evidence can do, then, is to reveal that the BNRist is committed to treating cases of coincidental variation, like Tap's, as non-ordinary cases that fall outside the scope of BP_O. But, as explained above, the non-empirical version of IDA already shows this. Thus, the empirical evidence is dialectically otiose.

6. Sophisticated Selectionism

In the previous section, I argued that empirical evidence of GIC/BEC cannot be used to support the claim of Different Experiences in Pautz's coincidental variation cases, and so IDA does not provide an empirical case against BNR. But I also suggested that *a priori* reflection on those cases reveals that BNR is committed to some implausible claims: the BNRist must hold that Tap has color experiences just like Ap's, even though his resulting behaviors and introspective reports fail to reflect the character of these experiences. And that, I suggested, is a genuine cost: in accepting the verdict of Same Experiences, the BNRist has to bite some meaningful bullets. So, we might wonder, is there a version of naïve realism that avoids these costs?

Pautz (2011) himself spells out such a view, which he labels *Sophisticated Selectionism* (henceforth, *SOPH*), as a potential way for the naïve realist to respond to his IDA. And, I want to suggest, some of the main targets of Pautz's IDA—philosophers he cites as endorsing BNR—actually hold just this kind of view, and so are immune to the problems IDA raises for BNR. In this section, I explore this version of naïve realism; in subsequent sections, I will consider whether empirical arguments provide reason to reject it.

[40] Pautz (2017, 31).

To begin, it will be helpful to clarify why, precisely, BNR is committed to the implausible verdict of Same Experiences. The source of the problem is that BNR places almost no constraints on what *kinds* of brain states can serve as "enabling conditions" on perceptual acquaintance. As Pautz defines the view, BNR includes a simple "causal-functionalist" account of acquaintance, according to which:

> "If you undergo some internal state or other that (i) causally detects in the biologically normal way the state of an external object being F… and that (ii) plays the right general functional role (e. g. it is… "poised" to be cognitively accessed), then you are consciously acquainted with that external state."[41]

Tap's neural states are different from Ap's, but, because BNR holds such internal factors to be entirely irrelevant to the question of which properties subjects are acquainted with, the fact that Ap and Tap are causally related to the same external properties is enough to ensure that their experiences will be the same.

On this picture, just about any neural state can play the role of enabling acquaintance with a given external-world property, *F*. But, as Pautz notes, naïve realists might reject this permissive account of acquaintance and instead hold that, for any neural state *N*, *N* can only enable perception of *F* if *N*, in addition to being causally related to *F* and being "poised" for cognitive access, "matches" *F* (in some yet-to-be-defined sense). This is Pautz's definition of *SOPH*.[42] On such a picture, the naïve realist could avoid the conclusion that Tap has the same kind of experience as Ap by insisting that the neural state Tap goes into when viewing a green object, N_2, does not match greenness, and so it does not result in an experience, like Ap's, whose character is constituted by acquaintance with that property.

We might now wonder what it takes for a neural state to match an external-world property and thereby enable acquaintance with it. In answering this question, we must remember that, for the naïve realist, the relevant external-world properties will be primitive, irreducible qualitative properties, and not physical properties—even when the physical properties in question serve as the supervenience base of qualitative properties. So, for the naïve realist, merely having neural states that are causally sensitive to certain external physical properties—spectral reflectances, say—may not be sufficient to achieve acquaintance with whatever high-level qualitative properties—like colors—supervene on those physical properties. The crucial question is whether the neural states in question bear the right sort of relation to the high-level properties—whether the two sets of properties match.

One plausible kind of constraint on matching that the SOPHist might posit is a structural one. In Tap's case, for example, there is a set of qualitative properties in the external environment—the colors—that have a specific structure (e.g., purple is more similar to blue than is green). Pautz stipulates that Tap has a set of neural states that are causally sensitive to the physical supervenience bases of these properties.[43] But, the SOPHist might say, Tap's states, unlike Ap's, bear no trace of a connection to the *colors themselves*, in that they do not track the structure of the colors (for example,

---

[41] Pautz (forthcoming, 8).

[42] See Pautz (2011, 412).

[43] Pautz (2013, 33) says that we can stipulate that Ap and Tap "detect the same light involving property L," and makes clear that this means that Tap has neural states that are causally sensitive to the physical properties associated with colors. But, given the naïve realist's anti-reductionism about qualitative properties, it is far from clear that this entails that Tap detects any *color*. Both Campbell (2021) and Fish (2013) suggest that they do not accept any such entailment.

they lead Tap to sort green things, rather than purple things, together with blue things). So, whatever those neural states are doing for Tap, the SOPHist will hold that they are not reacting causally to *the colors* in a way that matches them and would facilitate conscious awareness of them (if they were, surely Tap would be able to recognize that blue things were more like purple things than green things!).

Campbell explicitly suggests just this kind of structural constraint on what is required for a set of neural states to facilitate acquaintance with a set of high-level properties like the colors:

> Consider now what has to happen for there to be perceivers who can see those high-level properties, the colours. These perceivers must have perceptual systems that are causally responsive to those high-level structures…. If, at the levels of computation and algorithm, we are describing representations that are sensitive to the high-level structure of the colours, then anything recognizable as a biological implementation of those representations will have to have a corresponding biological structure.[44]

This implies that, on Campbell's SOPHist picture, the correct verdict about Tap is not Same Experiences. Tap has a set of neural states that are reliably correlated with the physical supervenience bases of the colors. But, because the structure of those neural states doesn't align with that of the high-level properties, it just isn't a way of achieving acquaintance with those properties. So, Tap isn't acquainted with the colors; and so he couldn't be having the same experiences as Ap, whose experiences are constituted by acquaintance with the colors.[45]

What will the SOPHist who rejects Same Experiences in this way say about Tap's experiences? Pautz suggests that SOPH will attribute illusory experiences to Tap:

> [On SOPH, Tap's neural state] is not the "right" adjustment, even though it is the adjustment [his] species evolved to undergo in response to the state of something being [purple]. Instead, [the SOPHist] might say, it matches something being green. So [Tap] does not perceive the state of object (sic) being [purple], though [his] state N_2 is appropriately caused by the state of the object being [purple]. Rather, on [SOPH, Tap] has an illusory experience as of something being green.[46]

But it seems to me the SOPHist is unlikely to ascribe illusory color experiences to Tap. Remember that the SOPHist denies that Tap's neural state is caused by the *purpleness* of the object; that neural state, though it is caused by the *physical supervenience base* of purple, is not part of a perceptual system that is causally sensitive to any colors at all. And, while Tap's neural state might count as the "right adjustment" to facilitate acquaintance with green in a subject capable of perceiving colors, Tap is not such a subject. So it seems dubious that the SOPHist will take Tap to be having illusory color experiences: we might suppose that illusions of color require a background of successful perception of colors, and Tap is not neurally equipped to ever perceive those high-level properties.

Instead, the SOPHist might take Tap to be in a kind of "zombie" state, one lacking in color phenomenology altogether, when his brain goes into N_2. For Tap, N_2 is correlated with a certain

---

[44] Campbell (2021, 411).

[45] Note that the kind of structural constraint this version of SOPH imposes on acquaintance is compatible with a wide (perhaps infinite) variety of specific physical states playing the acquaintance-enabling role. Hypothetical Martian subjects with silicon brains, for example, could achieve acquaintance with the colors, so long as their silicon brain states had a structure that lined up with that of the colors.

[46] Pautz (2011, 413-414).

physical property in the world—a spectral reflectance, say—without serving as a mechanism of acquaintance with the color green. But spectral reflectances aren't the kind of thing that subjects can be perceptually conscious of, on naïve realism. So, on the SOPHist's picture, Tap's being in N_2 isn't associated with his being perceptually conscious of anything at all.

In the previous section, I suggested that it was a significant cost to BNR that it had to treat Tap's case as "non-ordinary," one in which the subject's behavioral dispositions completely fail to reflect his underlying experience. Now, I am suggesting that on SOPH, the alternative version of naïve realism introduced to avoid this cost, Tap turns out to be in an even more bizarre situation: he is a kind of color-zombie, behaving as though he is having experiences of color when he has none at all. And so, one might worry, SOPH isn't much of an improvement on BNR.[47]

But I think that SOPH—which, as noted above, appears to be the version of naïve realism actually endorsed by some of Pautz's central targets—does indeed offer a superior response to Pautz's IDA. The SOPHist holds that acquaintance, which is a relation to high-level qualitative properties, requires more than the kinds of physical-causal connections Pautz stipulates to obtain in Tap's case. It requires that a subject be equipped with a system that "matches" the high-level properties that are potential objects of acquaintance, like colors. This is a principled position, falling out fairly directly from a central element of naïve realism itself—the idea that perception requires a connection to high-level, irreducible properties like colors, rather than a relation to physical properties—not an ad hoc response designed to deal with a problematic case. And, for the SOPHist, it is this antecedent philosophical commitment about the nature of acquaintance and the metaphysics of sensible qualities that motivates the claim that Tap must be a kind of zombie.

7. The Missing Explanation Argument

In the previous section, I laid out what I take to be the version of naïve realism actually endorsed by some central targets of Pautz's empirical arguments, and I suggested that such a view can avoid the costs associated with BNR, the view Pautz attributes to them. In this section, I will consider whether another of Pautz's arguments, which he explicitly formulates as an argument against SOPH, provides empirical reasons to reject the view.

In order to introduce this argument, which Pautz (2011) labels the *Missing Explanation Argument* (*MEA*), I want to return briefly to BNR and to consider a different way of pressing an empirical case against that version of naïve realism. As noted above, BNR is committed to the view that Tap's color experiences do not align with the structure of his neural states. That is, in Tap's case, we have *bad internal correlation* (*BIC*). But, empirically, we observe that, in actual cases of perception, we always find GIC. And, we might think, this presents BNR with an explanatory challenge: given the possibility of BIC in hypothetical cases like Tap's, why is it that we actually find only GIC?

As Pautz notes, SOPH can provide a straightforward answer to this question.[48] According to SOPH, perception is only possible if the structure of a subject's internal states aligns with—and

[47] On an alternative version of naïve realism, Tap fails to be acquainted with the colors, but is instead acquainted with a distinct set of qualitative properties that, though grounded in the same underlying physical properties as colors, have a distinct structure that is matched by Tap's neural states. This kind of view, endorsed by Allen (2017), would not, like SOPH, entail that Tap has zombie states; nor, like BNR, would it entail Same Experiences. But it requires positing an enormous range of qualitative properties, all grounded in the same set of physical properties, in order to accommodate all possible cases of coincidental variation. Like Pautz (2021, 225), I don't find this picture particularly compelling, but I will not address its merits here, as Pautz's reasons for rejecting it are not empirical in nature.

[48] Pautz (2011, 412).

thereby enables acquaintance with—the external-world properties that ground the character of experience. So BIC, on SOPH, is simply impossible: any subject capable of having a perceptual experience of, say, color, must have neural states that match the structure of the colors. We are guaranteed to find GIC in every case of perception.

But the explanatory challenge for BNR—to explain why we don't empirically observe any cases of BIC—will now simply re-emerge in a different form for SOPH. Cases like Tap's, in which a subject's internal neural states fail to match the structure of the external-world properties—like colors—that ground phenomenal character, are just as metaphysically possible on SOPH as on BNR. It's just that SOPH regards such cases not as cases of BIC—cases in which subjects have color experiences whose structure does not align with their neural states—but as "zombie" cases—cases in which subjects, in spite of behavioral indications to the contrary, have no genuine perceptual experiences at all. So, according to the SOPHist, there are two different kinds of neural systems subjects could have: zombie systems (*Z-systems*) that track physical properties like reflectances, but in a way that does not enable perceptual acquaintance with high-level properties like colors; and genuine perception systems (*P-systems*), whose structure matches that of the sensible properties, and thereby allows for acquaintance with those properties. But the SOPHist also claims that, in all *actual* cases, we only ever find P-systems—as Pautz puts it, according to SOPH, there is a "sweeping universal regularity in nature."[49] And this *Sweeping Regularity*—that we only ever find P-systems, never Z-systems—requires explanation. Pautz's MEA can now be summed up as follows: on SOPH, the *Sweeping Regularity* would require explanation; the SOPHist can offer no satisfying explanation of the *Sweeping Regularity*; and so SOPH should be rejected.

Pautz appeals to evolutionary considerations in defending his key "missing explanation" claim:

> [W]hether [species] came to perceive the response-independent sensible properties items had prior to evolution, or came to have [zombie states]…, was dependent on their particular biological make-up [i.e., whether they evolved P-systems or Z-systems]. Further, the evolution of [Z-systems] was a real possibility…. Now the process that led to a creature having a particular biological make-up was "blind" to the true… sensible properties of external items. Instead, it was determined by the unique set of selection pressures operating on the creature's ancestors: their particular habits, dietary needs, predators, and environments. What could make it likely that these factors should conspire to result in *all* creatures (not only *homo sapiens*) normally perceiving the true… sensible properties of *all* external items, [rather than having zombie states]?[50]

We can summarize Pautz's argument here as follows:

(1) On SOPH, the *Sweeping Regularity* requires an evolutionary explanation.
(2) Evolution could explain the *Sweeping Regularity* only if evolution were sensitive to sensible properties.
(3) Evolution is not sensitive to sensible properties.
(4) So, there is no evolutionary explanation of the *Sweeping Regularity*.
(5) So, on SOPH, we can't explain the *Sweeping Regularity*.

But it seems to me that SOPHists will simply reject premise (3) of this argument; that is, they will claim that evolution is indeed sensitive to sensible properties. Pautz defends (3) by stating that evolutionary processes are "blind" to sensible properties, and are instead determined by the

[49] Pautz (2011, 418).
[50] Pautz (2011, 415-416).

"particular habits, dietary needs, predators, and environments" of our evolutionary ancestors. But, for the naïve realist, sensible properties, like colors, were *part* of the environments in which our ancestors evolved. An organism with the right kind of perceptual mechanisms can, according to SOPH, achieve perceptual acquaintance with such properties. And, presumably, being able to perceive those properties would have been adaptive for our ancestors—it would have enabled them to successfully navigate their qualitative-property-laden environments. So there was evolutionary pressure to develop a neural system that could do so. That we now find organisms that have such perception-enabling systems—P-systems—is thus unsurprising.

Pautz's claim that evolution is blind to sensible properties might be based on the idea that the only kinds of environmental properties relevant to evolution are physical ones. Viewed through the lens of physics, there is, even by the SOPHist's lights, no meaningful difference between P-systems and Z-systems: both are causally sensitive (albeit in unsystematic ways) to external physical properties, like spectral reflectances. The real difference between the two kinds of systems, according to SOPH, is only apparent at the level of sensible properties: P-systems underwrite a capacity to perceive irreducible, high-level properties, while Z-systems do not. So the SOPHist's evolutionary explanation of the *Sweeping Regularity* does seem to turn on the idea that high-level properties, like colors, can causally influence evolution.

There are two reasons one might reject this idea. First, Pautz suggests that Z-systems might be "fully adaptive": there could be a Z-system that led to fitness-enhancing behaviors—guiding the organism through its environment by being causally responsive to physical properties like reflectances—even if, as the SOPHist holds, such a system wouldn't allow for genuine perception.[51] That is, on SOPH, species could have evolved to navigate their environments in two different ways: via P-systems, which track sensible properties, or via Z-systems, which instead track physical properties. There would seem to be no obvious adaptive advantage to the former method of navigation over the latter, and so high-level properties, even if they were genuinely part of the evolutionary environment, wouldn't result in any evolutionary pressure towards P-systems over Z-systems. Thus, the SOPHist's proposed evolutionary explanation fails.

Let us grant, on the SOPHist's behalf, that Z-systems could be as fitness-enhancing as P-systems. It seems to me that the SOPHist can still argue that evolution was more likely to produce P-systems than Z-systems. Remember that, at the level of physics, the connection between external properties and neural activity—both for P-systems and Z-systems—is extremely messy and unsystematic. So, the SOPHist might say, evolution had two possible pathways available, if it was to produce a system that allowed our ancestors to navigate their environments: it could either (a) respond to the presence of physical properties in the environment by evolving a massively complex system of neural states correlated, in structurally unsystematic ways, with those features (a Z-system); or (b) respond to the high-level sensible properties in the environment, like colors, by evolving a neural system whose states neatly align with the simple structure of those properties (a P-system). Even if both kinds of navigation systems would, in theory, be equally adaptive for the organism, it seems plausible that evolution was much more likely to "hit on" the massively simpler path leading to a P-system.

A second reason to reject the SOPHist's claim that high-level properties, like colors, influenced evolutionary processes is that such a claim might seem to entail a violation of the nomological completeness of physics.[52] On the picture I sketched above on behalf of the SOPHist, it would seem that the probabilities that certain physical events would occur (such as the mutations leading to P-systems rather than Z-systems) were not fully fixed by the totality of the antecedent physical facts,

[51] See Pautz (2011, 416).

[52] Thanks to an anonymous referee for raising this worry.

but were instead determined by non-physical, high-level properties. That is, the SOPHist would seem to be committed to something like the following counterfactual claim:

> (CF) If the physical facts prior to the evolution of P-systems had been the same, but sensible properties like colors had been absent (or distributed differently), then the probabilities associated with the physical events leading to P-systems would have been different.

The nomological completeness of physics entails that if the antecedent physical facts are held constant, there can be no difference in the probabilities of subsequent physical events. So, we might worry, SOPH, in accepting CF, is committed to a violation of this principle.

But note that CF is not merely a counterfactual claim; it is a counter-possible one. Naïve realists typically hold that high-level properties, though not reducible to physics, do supervene on physical properties as a matter of nomological (or perhaps metaphysical) necessity. So, the antecedent of CF, in describing a world in which the physical facts remain the same while the qualitative facts differ, describes a world that is, metaphysically or nomologically, impossible—it is, at a minimum, a world in which the laws of nature fail to hold. If, in such a world, we find a violation of a nomological principle like the completeness of physics, that is surely unsurprising. The SOPHist's commitment to CF, then, does not amount to a problematic rejection of the nomological completeness of physics; it entails only the unobjectionable (indeed, almost tautological) claim that there could be violations of such nomological principles *if the laws of nature were different*.

So it seems to me that the SOPHist can offer a plausible response to Pautz's MEA by providing an evolutionary explanation of the *Sweeping Regularity* that (a) does not violate the nomological completeness of physics, and (b) is based on the underlying anti-reductionist metaphysics that lies at the heart of naïve realism—that is, on the idea that sensible properties like colors are objective, irreducible, genuinely causally-efficacious features of the external world. Thus, the finding that we always observe GIC instead of BIC—which, on SOPH, amounts to the *Sweeping Regularity* with which we find genuine perception, never zombie states, in the actual world—does not provide empirical reason to reject SOPH.

8. The Simplicity Argument

I now want to consider one final empirical argument of Pautz's, which targets any version of naïve realism—BNR and SOPH included—that is committed to "objective primitivism," the view that sensible properties like colors are mind-independent, irreducible features of the external world. Unlike the arguments considered to this point—IDA and MEA—this argument does, by my lights, reveal a cost for naïve realism stemming from the findings of empirical science. But, I will argue, this cost is not nearly as steep as Pautz suggests; and, moreover, it is the kind of cost naïve realists already accept in myriad domains, as a known implication of their radically antireductionist metaphysics. Thus, the empirical evidence lacks meaningful argumentative force against naïve realism.

8.1. Naïve Realism's "External Laws"

As noted above, most naïve realists hold that qualitative properties, like colors, are not reducible to underlying physical properties, like spectral reflectances. Still, there is undeniably a connection between qualitative properties and physical ones—changes in spectral reflectance, for example, are associated with corresponding changes in color. Naïve realists typically account for this by holding that high-level qualitative properties, though not reducible to physics, are grounded in (or supervene

on) physical properties. So, on naïve realism, there will be lawlike relations between high-level qualitative properties and physical properties in the external world. And, given BEC, these "external laws" will have to be complex and unsystematic.

By contrast, an internalist view that takes experience to be grounded in neural states can eliminate these external laws. On such a view, we need not posit any objective, irreducible qualitative properties in the world, and so we won't need laws linking such properties to underlying physical properties. Instead, we will need a set of "internal laws" linking neural states to experience—N-variables to C-variables. And, given GIC, these internal laws will be relatively systematic.

Putting these ideas together, Pautz formulates his "Simplicity Argument" against naïve realism, and, specifically, its commitment to objective primitivism, as follows:

> (1) Naïve realism requires "external laws"; given BEC, these will be unsystematic and numerous.
> (2) Internalist views only require "internal laws"; given GIC, these will be systematic and simple.
> (3) Thus, simplicity considerations favor internalist views over naïve realism.[53]

How much of a cost, in terms of simplicity, does the need for external laws impose on naïve realism? Pautz suggests that the cost is enormous. He claims that, given BEC, these laws will take the form of an extremely long list, made up of "millions of brute 'external laws' – one for each and every distinct sensible property." For example:

> Necessarily, if there is an odor cloud made up of R-carvone, then it has the distinct primitive smell quality minty.
>
> Necessarily, if there is an odor cloud made up of R-limonene, then it has the distinct primitive smell quality citrus.

According to Pautz, "each and every one of these external laws will have to be taken as basic" because, given BEC, "there is no general, systematic functional dependence" of qualitative properties on physical properties, and thus no way to "predict the specific associations" we find between individual qualitative properties and their physical supervenience bases, in a way that would allow us to reduce the list to a simpler general formula.[54]

But this characterization of the kinds of external laws naïve realism requires is not supported by the empirical evidence. As explained in Section 4, even in the case of olfaction—the domain in which the relation between experience and external physical features is taken to be least systematic—there are now machine learning algorithms that can predict many odorants' sensory properties based on their physical properties. Such algorithms can, for example, predict which other odorants a novel stimulus will be judged to resemble. This shows that, contrary to Pautz's claims, the external laws required on naïve realism would not be a "mere list" of millions of individual qualitative-physical relations; there is enough systematicity in the connections between underlying physical properties and qualitative properties to allow for prediction of "specific associations," something that would be simply impossible if external laws genuinely amounted to mere lists. And so, there is much less of a cost to the naive realist's external laws, in terms of simplicity, than Pautz's "list" characterization suggests.

---

[53] See Pautz (2017, 27 and 2021, 217).
[54] Pautz (2021, 217-219).

To spell this out a bit further, consider how Pautz characterizes the *internal* laws linking neural states (N-variables) and experience (C-variables) on the internalist's picture. Pautz (2021, 219) suggests that the empirical finding of GIC means that, unlike in the case of naïve realism's external laws, there are "systematic functional dependencies" between N- and C-variables, and so we will only need "a small handful of general, systematic laws" to describe them, such as: "If a person undergoes PPC… neural state *N*, then they experience smell quality *g(N)*," where *g* is a "systematic function associating neural states organized in neural 'similarity spaces' (abstract spaces in which neural states are organized by similarity) with sensible properties in congruent 'quality spaces'."

But note that, given the success of the machine-learning algorithms mentioned above in codifying the links between external physical properties and smells, we can likewise characterize the naïve realist's *external* laws via functions associating physical properties organized in similarity spaces (known as "embedding spaces") with high-level odor qualities in congruent "smell spaces." Here is an example:

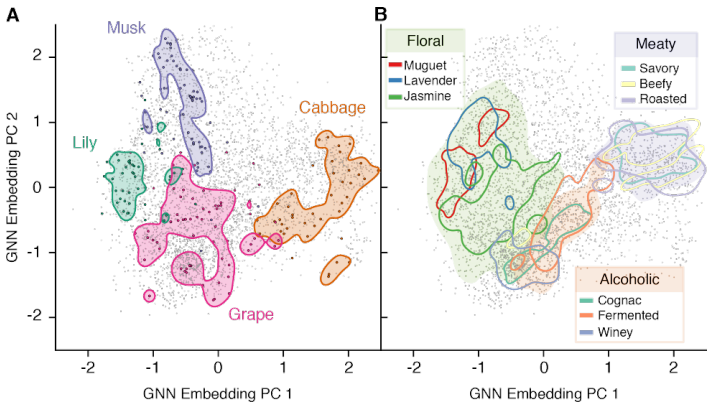

FIG. 3: 2-D embedding space in which odorant molecules are arranged according to a machine-learning analysis of their physical features. The similarity structure of the molecules as measured by the machine-learning model maps closely onto the similarity structure of the associated smell qualities, as expressed by linguistic descriptors (e.g., "floral"). From Sanchez-Lengeling et al. (2019), reproduced with permission of the corresponding author. Copyright 2019, Alexander B. Wiltschko.

In this chart, a machine-learning algorithm has arranged odorants in an embedding space—one determined by the dimensions of physical variation along which the algorithm analyzes molecules—that has the same similarity structure as the associated qualitative properties.[55] So, as with the internal laws mapping neural states to smell quality space, we have here an external law systematically mapping physical features to associated high-level qualitative properties.

The contrast between the naïve realist's external laws and the internalist's internal laws, then, is not, as Pautz claims, that the former amount to an immensely long list of one-off associations, while the latter are a small set of functions mapping congruent similarity spaces onto each other. Both internal and external laws can be captured by such mapping functions, and both can be used to make predictions about the qualitative properties associated with novel stimuli. The difference between external and internal laws, if there is one, must come down to a difference in the characters of the particular functions used to create the mappings in the two cases. Pautz's claim that the need for external laws imposes a cost on naïve realism, in terms of simplicity, amounts to the claim that the function mapping internal neural states to congruent quality spaces is more systematic or natural than the function mapping external physical properties to congruent quality spaces.

[55] See Sanchez-Lengeling et al. (2019).

The algorithms needed to generate embedding spaces of the kind pictured above are notoriously complex; such machine-learning algorithms are often referred to as "black boxes" because it is difficult to spell out the dimensions of similarity along which they analyze their inputs. I will not dispute the claim that there might indeed be some advantage, in terms of simplicity, to the internal laws linking N- and C-variables, when compared to the external laws linking physical properties with associated qualities that such "black box" algorithms generate. I simply want to highlight that this advantage is a rather subtle matter of degree. As noted in Section 4, the dimension of similarity along which internal neural states map onto qualitative character are not themselves perfectly natural—they typically involve lumping together activity in discontinuous brain regions via principal components analyses. So, the conclusion of the Simplicity Argument, taking into account the empirical details, amounts to this: internalist views are to be preferred because they require internal laws that, although not perfectly natural, are reasonably systematic, while the naïve realist's external laws are somewhat less systematic (though not to the point of being mere lists).

8.2. Antireductionism, Inter-Level Laws, and Complexity Costs

Let us grant that the empirical findings of GIC/BEC do reveal this kind of simplicity cost for naïve realism: naïve realism's external laws are, compared to the internal laws needed on alternative accounts, relatively unsystematic and complex. I now want to ask: How serious a cost is this additional complexity for someone who endorses naïve realism, and the philosophical commitments that come with it?

For a philosopher hoping to give a strongly unified metaphysical account of the world—one on which all properties are reducible to physics, say—having to include such unsystematic grounding laws would be a major cost. It would mean introducing a *type* of complexity in nature not otherwise required. But naïve realists typically reject such unified metaphysical accounts, for reasons independent of the empirical findings Pautz highlights. As Campbell puts it, consideration of "the special sciences, such as psychology, zoology, and economics, which cannot really be regarded as merely branches of physics," has led naïve realists to endorse a "pluralist" metaphysics on which high-level properties, though supervenient on physical properties, do not always map neatly onto them.[56] According to Campbell,

> There is, of course, nothing particularly remarkable about the idea that there may be high-level structures without any echo at the level of physics. Consider, for example, the dynamics of a credit squeeze. So far as the high-level economic properties go, there are quantitative models for what is happening: the failure of confidence, the liquidity crises for businesses, and so on. Now all this economic structure presumably supervenes on what is going at the level of gluons. But that's not to say that we should expect there to be quantum-mechanical structures, with independently identifiable quantum-mechanical significance, that correspond to the structure of a credit crunch.[57]

Naïve realists, then, are already committed to a picture on which, across innumerable domains, we find irreducible, high-level properties supervening in unsystematic ways on physics. The number of such laws, and their degree of complexity, is likely to be immense. The additional complexity needed to accommodate the naïve realist's posited qualitative properties—the need for external laws—is thus but one small, not especially noteworthy slice of the complexity naïve realists already accept in

[56] See (Campbell and Cassam 2010, 3).
[57] Campbell (2021, p. 411).

adopting their broadly antireductionist picture of the relation between high-level properties and physics.[58]

The need for such a complex metaphysics, and for innumerable, unsystematic laws linking physics to high-level properties, may indeed be thought a major cost for naïve realism. I certainly take it to be a strong reason to reject the view: naïve realists argue, in various ways, that we need to accept the existence of such irreducible, unsystematically-grounded high-level properties; but the cost of such a complex picture, by my lights, outweighs the force of any such arguments. Crucially, though, this cost is not one unique to the domain of qualitative properties, and it is not one only recently revealed by cutting-edge neuroscience; it is, rather, a general cost associated with the metaphysics naïve realism endorses. Recognizing the specific instance of unsystematic higher-level/lower-level laws required in the qualitative domain on naïve realism—the cost revealed by the empirical findings of BEC—should not meaningfully change our overall evaluation of the merits of the view, whatever that evaluation might be. Given naïve realism's metaphysical commitments, such costs are already baked into the philosophical debate. Put another way: the empirical evidence of GIC/BEC, though it is not entirely inert in the case of Pautz's Simplicity Argument, does not move the needle very much.

5. Conclusion

My purpose in this paper was not to defend naïve realism as the correct theory of perceptual experience (remember, I am a card-carrying member of the Anti-Naïve Realism Team!). Instead, I have been advocating for a radical methodological reorientation of debates about the nature of consciousness: we should abandon the style of empirically-grounded argumentation that dominates contemporary philosophy of mind.

I highlighted two kinds of questions about the application of empirical scientific findings to philosophical debates about consciousness that are too often neglected. The first concerns what the science actually shows, empirically, and what non-empirical premises philosophers (and, occasionally, scientists themselves) might be sneaking in. I suggested that many arguments against naïve realism leap from empirical evidence about observable phenomena, like subjects' verbal reports, to conclusions about experience, by way of implicit, non-empirical bridge principles; and, I noted, these bridge principles are often ones that naïve realists already reject, in formulating responses to more traditional "armchair" arguments against their view. Thus, the empirical data fail to advance the philosophical case against naïve realism. I also emphasized that the science itself is often messier than philosophers who cite it let on. In doing so, I was attempting to take seriously Burge's admonition that philosophers ought to attend to the details of the science. To my mind, that means that we shouldn't just take as authoritative the way empirically-oriented philosophers report *what science says*. Careful assessment of claims like Pautz's—for instance, that GIC and BEC are scientifically proven facts—requires that we consider, for example, whether the scientific studies

[58] Though Campbell is quite explicit in his endorsement of this picture, one could adopt naïve realism's broadly antireductionist metaphysics, while denying that the relation between high-level properties and physics in non-qualitative domains will, in general, be extremely complex. Given the wide range of high-level properties naïve realists typically postulate, Campbell's claim that complex high-level/lower-level relations will be the rule, rather than the exception, seems more plausible to me, and I am not aware of any naïve realists who explicitly claim that the qualitative domain is unique in this regard. But, for a (perhaps merely hypothetical) naïve realist who denies the existence of complex "inter-level" relations in non-qualitative domains, the empirical finding of BEC, in revealing the need for *uniquely* unsystematic external laws linking physics to qualitative properties, would impose a more significant theoretical cost.

cited evaluate internal and external variables according to a uniform standard. Empirical psychologists are not themselves focused on arbitrating between competing internalist and externalist theories of the nature of consciousness; often, they are working within specific paradigms that simply assume a brain-based view of experience. And so, a question that might be crucial philosophically—for example, whether, with enough effort, we could find a non-natural dimension of external similarity along which we'd find GEC—might be one that working neuroscientists simply don't bother considering (while other scientists, in fields like machine learning, might be more likely to carry out the relevant research).

The second question that is often neglected in today's philosophical debates concerns the ways in which the philosophical commitments of naïve realism—in particular, the view's primitivist account of qualitative properties and its radically antireductionist metaphysics—bear on empirical arguments against the view. The SOPHist's claim that we should not count Tap as perceiving colors; the evolutionary explanation the SOPHist offers of the *Sweeping Regularity* with which we find GIC; the naïve realist's contention that the simplicity costs associated with objective primitivism are not especially significant: all of these fall out fairly directly from a principled, independently motivated philosophical picture.

The naive realist's philosophical picture is one that many, myself included, do not endorse; but the real reasons to reject the view do not stem from empirical findings that fit quite comfortably within it. Instead, the real argumentative terrain comprises philosophical questions about what kinds of bridge principles we should accept, and what kinds of theoretical costs are associated with positing irreducible high-level properties. Arriving at this terrain did not require cutting-edge results from perceptual psychology; naïve realism's metaphysical commitments and non-standard bridge principles are explicit features of the view, not hidden costs forced on its advocates by recent empirical findings. The foray into the empirical literature, then, was something of a fruitless detour. As philosophers of mind engaged in debates about the nature of consciousness, we would all be better served spending a bit less time trying to wield empirical science as a cudgel against our opponents, and a bit more time working through the philosophical implications of each other's views – something we can accomplish perfectly well from the comfort of our armchairs.[59]

[59] Versions of this paper were presented at the *New Waves in Relationalism* conference, hosted by Ori Beck and Farid Masrour (May 2021); at a session of Geoff Lee's graduate seminar on *The Metaphysics of Perceptual Experience* at UC Berkeley (October 2021); and at a session of my graduate seminar on *Consciousness and Empirical Science* at Brandeis (December 2021). Thanks to the organizers and participants for their helpful feedback. I would also like to thank Adam Pautz for multiple rounds of extremely in-depth discussion of earlier drafts, and an anonymous referee from *Noûs* for providing very helpful comments. Finally, thanks to Umrao Sethi, whose help throughout the process of writing this paper was, as always, immeasurable.

WORKS CITED

Allen, K. (2016). *A Naïve Realist Theory of Colour*. Oxford: Oxford University Press.

Block, N. (2010). Attention and mental paint. *Philosophical Issues, 20*, 23-63. https://doi.org/10.1111/j.1533-6077.2010.00177.x

Bohon, K. S., Hermann, K. L., Hansen, T., & Conway, B. R. (2016). Representation of perceptual color space in macaque posterior inferior temporal cortex (the V4 complex). *eNeuro*, *3*. https://doi.org/10.1523/ENEURO.0039-16.2016

Brouwer, G. & Heeger, D. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience, 29*, 13992-14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics,* 33, 1-78. https://doi.org/10.5840/philtopics20053311

Burge, T. (2011). Disjunctivism Again. *Philosophical Explorations 14*, 43-80. https://doi.org/10.1080/13869795.2011.544400

Campbell, J. (2010). Demonstrative reference, the relational view of experience, and the proximality principle. In R. Jeshion (ed.), *New Essays on Singular Thought*. Oxford: Oxford University Press.

Campbell, J. (2020). Does that which makes the sensation of blue a mental fact escape us? In D. Brown & F. Macpherson (eds.), *The Routledge Handbook of Philosophy of Colour*. Abingdon: Routledge.

Campbell, J. & Cassam, Q. (2014). *Berkeley's Puzzle: What Does Experience Teach Us?* Oxford: Oxford University Press.

Cavedon-Taylor, D. (2018). Naïve realism and the cognitive penetrability of perception. *Analytic Philosophy, 59*, 391-412. https://doi.org/10.1111/phib.12131

Chalmers, D. (1996). On the search for the neural correlate of consciousness. Retrieved from: http://cogprints.org/227/1/199711001.html.

Cowart, B. & Rawson, N. (2005). Olfaction. In E. B. Goldstein (Ed.), *Blackwell Handbook of Sensation and Perception* (pp. 567–600). Hoboken: Blackwell.

Craig, E. & Hoskin, M. (1992). Hegel and the seven planets. *Journal for the History of Astronomy, 23*, 208-210. https://doi.org/10.1177/002182869202300307

Fish, W. (2013). Perception, hallucination, and illusion: reply to my critics. *Philosophical Studies*, *163*, 57-66. https://doi.org/10.1007/s11098-012-0072-8

French, C. & Phillips, I. (forthcoming). Naïve realism, the slightest philosophy, and the slightest science. In B. McLaughlin & J. Cohen (eds.), *Contemporary Debates in the Philosophy of Mind*. Hoboken: Blackwell.

Howard, J. et al. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nature Neuroscience, 12*, 932–938. https://doi.org/10.1038/nn.2324

Keller, A. et al. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science, 355*, 820-826. https://doi.org/10.1126/science.aal2014

Linster, C. et al. (2001). Perceptual correlates of neural representations evoked by odorant enantiomers. *Journal of Neuroscience, 24*, 9837-9843.

Margot, C. (2009). A noseful of objects. *Nature Neuroscience, 12*, 813-814. https://doi.org/10.1038/nn0709-813

Martin, M. G. F. (2004). The limits of self-awareness. *Philosophical Studies, 120*, 37-89. https://doi.org/10.1023/B:PHIL.0000033751.66949.97

Masrour, F. (2020). On the possibility of hallucinations. *Mind, 129*, 737-768. https://doi.org/10.1093/mind/fzy088

Nanay, Bence (2014). Empirical problems with anti-representationalism. In B. Brogaard (ed.), *Does Perception have Content?* Oxford: Oxford University Press.
Noordhof, Paul (2021). Wading in the Shallows. In H. Logue and L. Richardson (eds.), *Purpose and Procedure in Philosophy of Perception*. Oxford: Oxford University Press.
Pautz, A. (2006). Sensory awareness is not a wide physical relation: an empirical argument against externalist intentionalism. *Noûs, 40*, 205-240. https://doi.org/10.1111/j.0029-4624.2006.00607.x
Pautz, A. (2011). Can disjunctivists explain our access to the sensible world? *Philosophical Issues, 21*, 384-433. https://doi.org/10.1111/j.1533-6077.2011.00209.x
Pautz, A. (2013). The real trouble for phenomenal externalists: new empirical evidence. In R. Brown (ed.), *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience* (237-298). New York: Springer.
Pautz, A. (2017). Experiences are representations: an empirical argument. In B. Nanay (ed.), *Current Controversies in the Philosophy of Perception*. Abingdon: Routledge.
Pautz, A. (2021). *Perception*. Abingdon: Routledge.
Pautz, A. (forthcoming). Naïve realism v representationalism: an argument from science. In J. Cohen & B. McLaughlin (eds.), *Contemporary Debates in Philosophy of Mind*. Hoboken: Blackwell.
Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press.
Sanchez-Lengeling, B.; Wei, J.N.; Lee, B.K.; Gerkin, R.C.; Aspuru-Guzik, A.; & Wiltscko, B.W. (2019). Machine learning for scent: learning generalizable perceptual representations of small molecules. Retrieved from arXiv database (arXiv:1910.10685) at https://arxiv.org/abs/1910.10685
Wu, W. (2020). Is vision for action unconscious? *Journal of Philosophy, 117*, 413-433. https://doi.org/10.5840/jphil2020117826