# Consciousness Meets Lewisian Interpretation Theory
## A Multistage Account of Intentionality

*Adam Pautz*

All thinking has to start from acquaintance; but it succeeds in thinking about many things with which we have no acquaintance.
– Bertrand Russell

Acquaintance is a condition on the possibility of thought and justification.
– David Chalmers

Karl experiences a tomato of a round and somewhat bulgy shape. He believes that rabbits are getting into his garden. He worries that democracy is in trouble. He is confident that 68 plus 57 will always make 125.

In "Radical Interpretation" (1974), David Lewis asked: by what constraints, and to what extent, do the non-intentional facts about Karl determine such intentional facts? There are two popular approaches. First, the *reductive externalist program*. The austere physical facts about Karl are the only facts. Original intentionality reduces to informational-teleological relations between Karl's brain and the world. I will use "tracking relations" as neutral term for this kind of relations. Second, there is the totally different *phenomenal intentionality program*. According to it, the intentional contents of Karl's conscious experiences are determined by his internal brain states, not tracking relations to the environment. And his internal conscious experiences play a crucial role in pinning down the contents of his mental states.

I will argue against both approaches. I will agree with friends of phenomenal intentionality that reductive externalists neglect the role of our internally-determined conscious experiences in grounding intentionality. But I will fault them for not adequately explaining intentionality. They cannot just say "conscious experience explains it" and leave it at that. However, I will sketch an alternative multistage account incorporating ideas from both camps. In particular, by appealing to Lewisian ideas, we can explain how Karl's conscious experiences help to ground the contents of his other mental states. The result is a "consciousness first" approach to intentionality.

My plan is as follows. In §1 and §2 I will catalogue problems for the reductive externalist program and the phenomenal intentionality program. Along the way, I will lay down desiderata for a theory of intentionality. In §3, I will sketch my alternative multistage theory and show how it might satisfy those desiderata.

## 1. PROBLEMS WITH THE REDUCTIVE EXTERNALIST PROGRAM

In §1.1 and §1.2 I catalogue problems for specific ideas within the reductive externalist program. In §1.3 I raise a more general problem. All the problems concern the connection between intentionality and conscious experience.

### 1.1 The problem of experiential indeterminacy

My first problem for the reductive externalist program concerns how to determine Karl's sensory-perceptual experiences. I will assume intentionalism (Chalmers 2010, Dretske 1995, Horgan 2014, Tye 2019). Experiential phenomenology and intentionality are inseparable. The phenomenology Karl's experiences is a matter of what perceptible properties he is conscious *of* (in other words, "experientially represents"). So determining his experiences is a special case of the problem of intentionality.

Roughly, a reductive externalist account of experiential intentionality goes as follows. The perceptible properties ("qualia") are response-independent physical properties (reflectance-types, chemical-types, etc.). The conscious-of relation is a complex tracking relation between subjects and such physical properties. For instance, for Karl to be conscious of the quality red (a certain reflectance-type), and have a "reddish" experience, is for Karl to undergo a subpersonal brain state that has the biological function of tracking (being produced by) the occurrence of red and that is poised to influence the cognitive system. The result is *reductive externalist intentionalism* about phenomenology (Dretske 1995, Tye 2019).

But indeterminacy worries undermine this view. Let me summarize two illustrations that I have developed in much greater detail elsewhere (Pautz 2017).

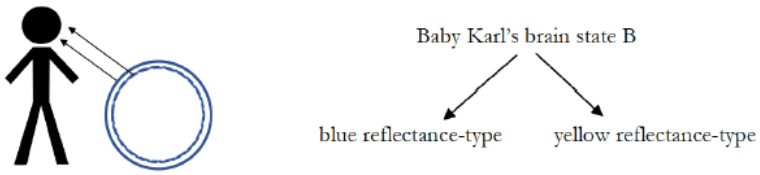


Figure 1: Black-and-white earth (left) and middle earth (right)

First, imagine that Karl and his kind evolved on *black-and-white earth*. On black-and-white earth, the following things are true. First, *surfaces* of all object are either black or white – this is why I call it "black-and-white earth". Second, every object contains a smaller object. In particular, black outer objects contain red inner objects and white outer objects contain green inner objects. But the objects are impenetrable. Third, the color of the inner object and that of the outer object are causally yoked together by way of a natural, super-fast chemical process.

Now suppose Karl views a black object containing a red inner object (Figure 1, l*eft*). Does Karl have a "blackish" experience or a "reddish" experience? I intentionality described the example in physical terms, leaving open the character of his experience. Reductive externalists about phenomenology might say that Karl appropriately tracks, and thereby is conscious of ("experientially represents"), the outer black only. In that case, he has a blackish experience. Alternatively, reductive externalists might say that Karl appropriately tracks, and thereby is conscious of ("experientially represents"), the more distal *inner red*. On this account, although the outer black is part of the causal process, Karl isn't conscious of it – anymore than he is conscious of his retinas or the light. It is just part of the causal process that enables him to be conscious of the inner red. In that case, he has a reddish experience. If the austere physical facts are all the facts, it's hard to see what could make one of these accounts determinately correct and the other incorrect.

My second example (Figure 1, *right*) is arrived at in two stages. First, *Harry* is on earth and *Sally* is on inverted earth. Harry looks at the sky. His brain state *B* has the function of tracking the *blue* of the sky. So, according to reductive externalism about experience, *B* enables them to be conscious of the blue of the sky and to have an experience with a "bluish" phenomenology. On inverted earth, the sky is yellow. Even though the sky is yellow, it puts Sally into the same brain state *B* that Harry is in. In her population, this brain state *B* has the function of tracking *yellow*. So, according to reductive externalism, she is conscious of yellow and has a "yellowish" experience.

In the second stage, Harry and Sally leave their planets and wind up on *middle earth*. This is when Harry met Sally. Even though they evolved separately and belong to different species, they are able to have a baby, Karl. Baby Karl is born without eyes. But he does have a complete visual cortex. One day he undergoes brain state *B* and has a hallucinatory color experience.

On reductive externalism, does Baby Karl's have a bluish or a yellowish experience? Does his brain state *B* have a biological function of tracking *blue* or *yellow* (Figure 1, *left*)? There is no clear answer. In his dad Harry's population, *B* has a history of tracking blue. But, in his mom Sally's population, *B* has a history of tracking yellow.

Such cases make a dilemma for externalists (described in detail in Pautz 2017). One option is *radical experiential indeterminacy*. In the black-and-white earth case, as Karl views the object, it is determinate that he *either* has a blackish *or* a reddish experience, but it is indeterminate *which one*, because it is indeterminate whether he "experientially represents" the outer black reflectance or the inner red reflectance. Likewise, on middle earth, it is determinate that Karl either has a bluish or a yellowish hallucination, but it is indeterminate which one.

But, whatever we may think of radical indeterminacy in thought and language, radical indeterminacy in experiential character is incoherent.

Another option is *arbitrary identities*. Presumably, since relations are abundant, there is a tracking relation – call it tracking$_{17}$ – that Karl on black-and-white earth bears to the outer black but not the inner red; and there is another tracking relation call it - tracking$_{18}$ - that Karl bears to the inner red but not the outer black. Now, maybe it is just a "surd metaphysical fact" (Putnam 1981: 46-48) that the conscious-of ("experiential representation") relation is determinately identical with tracking$_{17}$ instead of tracking$_{18}$, so that it is determinate that Karl is consciously acquainted with the outer black rather than the inner red. And maybe this same tracking relation is one that Karl on Middle Earth bears to (say) yellow rather than blue when having his hallucination. So he has a yellowish experience rather a bluish one.

But the arbitrary identities view flouts the plausible idea that the conscious-of relation is a "stand out" relation. That is, when Karl is consciously acquainted with property *P* (e. g. a certain color), but not at all consciously acquainted with property *P**, there is a massive difference in his relation to *P* and *P**. The arbitrary identifies view flouts this because, while Karl stands in the tracking$_{17}$ relation the outer black (which on this view is identical with the relation of conscious acquaintance), he stands in the barely different tracking$_{18}$ relation to the inner red. On this option, then, there is no way in which the allegedly correct interpretation (*viz.* Karl is conscious of the outer black, and misses by a hair acquaintance with the inner red) "stands out" from the allegedly incorrect one (viz. Karl is conscious of the inner red, and the outer black is just part of the mediating causal process).

In short, the reductive externalist program has difficulty with:

> *Experiential determinacy.* There cannot be radical indeterminacy in the intentional contents of Karl's experiences. Moreover, the correct assignment of contents to his experiences and experience-based thoughts "stands out" (significantly differs) from alternative, incorrect interpretations.

### 1.2 Problems with the inner sentence theory of belief and desire

Now suppose that Karl is prelinguistic hominid with simple beliefs and desires. How might reductive externalists account for this?

Many reductive externalists accept Jerry Fodor's (1990, 2010) inner sentence approach ("the representational theory of the mind"). Here is the version I will focus on. First, although he lacks a language that he can experience, Karl has an "inner" subpersonal language that he cannot experience. The sentences of this inner language somehow get their contents (*that is red*, *that is round*, *there is a friend*) by way of tracking relations to external items. Once the content of an inner sentence is initially fixed, it tends to retain that content, even when it is temporarily severed from its normal connections to perceptual inputs and behavioral outputs. (This is required to explain false and irrational beliefs.) Second, there is a belief-box and a desire-box. The sufficient conditions for "box-inclusion" are functional. I will assume that *one* sufficient condition is this: *if* subpersonal inner sentences $b_1$, $b_2$, . . . and $d_1$, $d_2$, . . . typically interact to cause the actions which, according to $b_1$, $b_2$, . . ., will satisfy $d_1$, $d_2$, . . ., *then* inner sentences $b_1$, $b_2$, . . . are "in the belief-box" and inner sentences $d_1$, $d_2$, . . . are "in the desire-box". Call this *means-ends*. Finally, to believe that $p$ is to have a subpersonal inner sentence in one's "belief-box" that means that $p$, and to desire that $q$ is to have a subpersonal inner sentence is one's "desire-box" that means that $q$.

In my view, this popular reductive externalist approach to belief and desire misses some deep connections between Karl's beliefs and desires and his conscious experiences.

First, the inner sentence theory violates:

> *Conscious-life constraint.* The beliefs and desires of individuals with conscious experiences cannot "radically change" if there is no change in *either* (i) their conscious experiences and *or* (ii) their dispositions to consciously act (including inner or outer speech dispositions, for individuals with language).

Here's an example showing why the inner sentence theory violates this. Karl the prelinguistic hominid is starving, and he is given vanilla ice-cream for the first time. He believes that this white stuff tastes *sweet and good* and wants this sweet, good-tasting stuff in his mouth. He devours it for five minutes. However, in the middle of his chow-down, while his experiences and behavior remain the same, his inner sentences are temporally scrambled. In particular, the subpersonal sentence "this tastes *horribly bitter and disgusting*" is tokened in his belief-box and "I will have this *specific bitter*, *disgusting stuff in my mouth*" is tokened in his desire-box. The inner sentence theory implies that, for this short 10-second interval, Karl suddenly, and for no reason, secretly acquired a new, irrational and totally false belief about the ice-cream (it tastes horribly bitter and disgusting) as well as a crazy desire (to have this specific disgusting stuff in your mouth). Call this *secret scram-*

*bling*. Against this, *throughout* Karl evidently believes it tastes *good* (not disgusting), and wants this *good* stuff in his mouth, even if these mental states are differently realized by different, intrinsically-meaningless symbols at the hidden neural level.

The inner sentence theory also violates the following:

> *Constitutive experience-belief connection.* Experiences do not merely cause beliefs. They are *necessarily* apt to cause beliefs. That is, they are necessarily *compelling*. Necessarily, if Karl experiences clearly different colors, he is at least *disposed to* believe that they are different. If Karl is conscious of a red and round item, he *is disposed to* believe that such an item is present. Necessarily, if Karl has striking taste sensation, or a searing pain, he is disposed to believe he is in that state.[1]

The inner sentence theory violates this because it holds that Karl's conscious experience of a red and round thing is realized by a (iconic) neural representation in one area of his brain, while belief that a red and round thing is there is realized by a (discursive, sentence-like) representation in another area of this brain. Further, the connection between the two is *utterly contingent*, and subject to radical and regular malfunction, like the connection between having joint pain and believing the weather will change, or between fire and a fire alarm. Against this, the experience-belief connection is stronger than that.

Finally, the inner sentence theory of belief and desire violates:

> *Prelinguistic limits.* (i) If Karl (like a prelinguistic hominid) lacks an *outer* language and is limited to having the usual range of human experiences, then there are rough limits on what he can believe and desire. He can have beliefs about perceptible properties, the kinds of things in his environment, other people, the near past and future, and so on. But he cannot believe propositions about specific large numbers, the laws of quantum mechanics, abstract philosophical doctrines, and so on. (ii) He can only form such sophisticated beliefs if he has an *outer* language.[2]

Prelinguistic limits is supported by pretheoretical reflection. Just try to describe possible circumstances where prelinguistic Karl clearly has such sophisticated beliefs: you cannot do it. It also fits the facts: humans came to have sophisticated beliefs only by inventing a sophisticated outer language. I will provide examples in §§3.2-3.4.

But prelinguistic limits is puzzling. What *explains* the necessary restriction on prelinguistic Karl's beliefs and desires? The inner sentence theory doesn't explain prelinguistic limits; in fact, it violates prelinguistic limits because it holds that Karl's beliefs and desires are fixed by his *inner* language, and in principle there are no limits on how sophisticated it

---

[1] For defenses of a constitutive experience-belief connection, see Byrne 2018: sect. 6.2.10; Hawthorne 2006: 249-250; Shoemaker 1996: chap. 3; Lewis 1999: 6.

[2] For defences of prelinguistic limits, see Bennett 1976: 96; Bermudez 2003: 150ff; Blackburn 1984; 137-140; Dehaene 1999: chap. 4; Dennett 1987: 201; Hurford 2014: 124; Sacks 1989: chap. 2; Speaks 2010: 234ff; Spelke 2003; Pinker and Jackendoff 2005: 206; and Wittgenstein 1953: 174ff. I note in passing that the individuals studied by Varley *et al.* 2005 who have aphasia but who are capable of mathematical thought are not a counterexample to prelinguistic limits because they can understand and accept mathematical sentences.

might be. On the inner sentence theory, lacking an *outer* language should be no bar to having arbitrarily complex beliefs and desires.

For instance, pretend that Karl does indeed have something like an inner language. Suppose that Karl has a magical subpersonal detector that only does one thing: when there is a collection of exactly 167 things just behind his head (where he cannot see anything), it causes the symbol "167 things are behind me" to be tokened in his head. Suppose that this inner sentence then combines with sentences in his desire box to lead to rudimentary behavior (for instance, walking forward), so that it satisfies the means-end condition for being in the belief-box. The inner sentence model implies that on such occasions Karl *believes* that there are exactly 167 things behind him. Likewise, there is no reason why prelinguistic Karl could not have within his inner language terms that track *democracies* and *electrons* (Fodor 1990: 111). So inner sentence theory implies that, when simple-minded Karl engages in other rudimentary behaviors, he might count as having beliefs about *democracies* or *electrons*.

But Karl the prelinguistic hominid evidently does not and (in the circumstances) cannot have such beliefs. For instance, he doesn't *believe* that there are exactly 167 things behind him - he has *no idea* what is behind him and no available way of thinking about large exact numbers.

### 1.3 A general problem: internalism about experiential intentionality

In any form, the reductive externalist program holds that all intentionality is grounded in "tracking" relations between the brain and the world. In their essay "The Intentionality of Phenomenology and the Phenomenology of Intentionality" (2002), Horgan and Tienson used an internalist thesis to argue against the reductive externalist program and for their alternative phenomenal intentionality program:

> *Internalism about experiential intentionality.* The phenomenology of experience is not determined by tracking relations to the environment; it is internally-determined. And much intentionality is inseparable from phenomenology.

Horgan and Tienson, like many others, hold experiential internalism to be "self-evident" from the armchair, not requiring argument (2002: n.23). I disagree. However, decades of research in psychophysics and neuroscience support experiential internalism (Pautz 2010, 2019). What pains, smells, color qualities, and so on, we experience are fixed by internal neural processing, not what external physical properties (types of damage, chemical-types, reflectance-types) our sensory systems have the function of tracking. And this does indeed rule out the reductive externalist program. Here are some illustrations.

First, consider a *coincidental variation case* (Pautz 2010). Karl and Twin Karl's sensory systems have the function of tracking the *same* types of damage, chemical-types, reflectance-types, and so on, but their internal sensory processing is very different. Given the empirically-determined role of the brain, they have radically different experiences. Given intentionalism, their experiences, and their experience-based beliefs, *differ* in content. So standard tracking theories (Dretske 1995, Tye 2019, Neander 2017, Williams 2020) fail for experiential intentionality.

The *brain-in-the-void* (BIV) undermines all reductive externalist theories. Given experiential internalism, an accidental, life-long brain in the void (e. g. a "Blotzmann brain")

undergoing all the same (actual and counterfactual) brain states as Karl would also have all the same experiences as Karl. Given intentionalism, the experiences of Karl-the-brain would have rich intentional contents, for instance *there is a round thing there*. And, although BIV-Karl doesn't have the same "wide" beliefs as Karl, it is common-sense that BIV-Karl would share many beliefs with Karl - nearly all false in his case (Lewis 1994: 425). But BIV-Karl would not bear any tracking relations to external states of affairs – for instance, the state of affairs of there being a round thing before him (Pautz 2019).

## 2. PROBLEMS WITH THE PHENOMENAL INTENTIONAITY PROGRAM

The reductive externalist program, then, fails to satisfy several plausible desiderata concerning links between intentionality and phenomenal experience. So let us turn to the phenomenal intentionality program.

Proponents of the phenomenal intentionality program typically accept "internalist intentionalism" about Karl's sensory-perceptual experiences (Chalmers 2010, Horgan 2014, Pautz 2010). The contents of Karl's experiences are determined by his brain states rather than by tracking relations to the environment. Unlike reductive externalism, this view satisfies internalism about experiential intentionality. For instance, BIV-Karl will share many of Karl's intentional states. And it may satisfy experiential determinacy. For example, on middle earth and black-and-white-earth, the contents of Karl's experiences, and so their phenomenal characters, are determinately pinned down by his brain states.

As for Karl's thoughts, friends of phenomenal intentionality accept the *cognitive experience theory*. Instead of explaining Karl's thoughts in terms of hidden inner sentences that track external states-of-affairs, they explain his thoughts in terms of special "cognitive experiences". Cognitive intentionality is "phenomenal intentionality" (Horgan and Tienson 2002).

In my view, the phenomenal intentionality program is along the right lines. In fact, I will incorporate their "internalist intentionalism" about sensory-perceptual experience into my own account in §3. More generally, I think that they are right to emphasize the role of conscious experience in determining intentionality. However, in this section, I will argue against their simple "cognitive experience theory" of thought. Here we need a more complex, multistage story (§3).

I will first describe the cognitive experience theory in greater detail (§2.1). Then I will argue that it violates some important desiderata on a theory of intentionality**.** It fails to adequately explain *thought content* (§2.2), the *holistic character of thought*, and *prelinguistic limits* (§2.3).

### 2.1 The Cognitive Experience Theory of Thought

In giving an account of how Karl has thoughts with certain contents, we must address well-known underdetermination worries due to Quine (1960) and Kripkenstein (1982). For example, let the *quus$_g$-function* be an arithmetical function like the plus-function except that it gives weird results for some specific numbers too large for Karl to compute (e. g. some specific numbers in the googolplex range). Or suppose that extension of *friend** is like that of *friend* in nearly all worlds but that it has a somewhat twisted extension in very remote worlds (so that Karl's finite dispositions are neutral between *friend* and *friend**).

How do the physical facts determine that Karl thinks that that 68 plus 57 equals 125 rather than 68 *quus*$_g$ 57 equals 125, or that Friedrich is a friend rather than a friend*?

The cognitive experience theory holds that Karl's thought-contents are constituted by special "cognitive experiences". So the question "how do the physical facts determine the contents of Karl's thoughts?" becomes the question "how do physical facts determine his cognitive experiences?" Here are some representative passages:

> Something is happening to you experientially as you read. Obviously, there is the visual or auditory experience [and] perhaps a rapid and silent process of forming acoustic mental images. But there is something else - a certain complex modification of the quality of one's course of experience, and not just of one's dispositional set. There is understanding-experience. [Its] existence is sometimes doubted, perhaps because it has no striking experiential feel in the way in which experience in any of the sensory modalities usually does. (Strawson 2010: 8)

> [Cognitive experience] makes it the case that I can think determinately about the number 2 although there is no relevant causal context. *Pfff*! This is the correct account of how it is that content can be determinate in spite of all the problems raised for this idea by Kripke in his book *Wittgenstein on Rules and Private Language*. (Strawson 2010: 354)

> We have no explanation of how the systems of the brain that underlie or realize thought give rise to, or involve, conscious thought experience in the way in which they do. (Strawson 2010: 255, fn.54)

> Certain conscious states are intrinsically such as to ground thought or understanding. There is a conscious state which is intrinsically such as to ground the thought that two plus two equals four. (Goff 2012: 223)

> If consciousness is inside the head, then, in explaining [cognitive] phenomenology, we must confine ourselves to facts about what's going on inside the head; to what me and my brain in a vat twin have in common. (Goff 2012: 232)

> Physically and apart from phenomenology, there is no "one, determinate, right answer" to the question of what is the content of an intentional state. Content identity or determinacy is fixed phenomenally. For example, the what-it's-like of thinking "Lo, a rabbit" is different from the what-it's-like of thinking "Lo, a collection of undetached rabbit parts" (Graham, Horgan and Tienson 2007: 476)

> [My view] maintains that the intentional content of a thought is determined by its *intrinsic* phenomenal properties, *not its relational properties*. My teachers will be very disappointed in me. (Pitt 2009: fn.5, my italics)

> The part of what is thought that is fully determined by [cognitive] phenomenal character [is] a kind of thought content. (Siewert 2011: 264)

> If there is irreducible cognitive phenomenology, it individuates as finely as content. (Kriegel 2015: 62)

In sum, the *cognitive experience theory* holds that, in addition to having sensory-perceptual experiences, Karl has special *cognitive experiences* with certain built-in ("narrow") contents. In fact, to have these experiences just is to grasp certain contents (see Pautz 2013: 209 and Chudnoff 2015: 135). So Karl's phenomenal life is extremely rich. He has a special "democracy" cognitive phenomenology, and a special "plus-function" cognitive phenomenology, and so on. Such cognitive experiences are not reducible to sensory-perceptual experiences. For instance, suppose that Karl has a deaf twin who speaks *sign-language* instead of English, and that Karl and his twin are talking about arithmetic or the state of democracy. Then they have very different *sensory-perceptual experiences.* But, if cognitive experiences are irreducible, they presumably might have the very same *cognitive experiences*.

The cognitive experience theory has some initial appeal. And it may help to avoid a problem we pressed against the inner sentence theory. The inner sentence theory violates the conscious-life constraint on belief (§1.2). The cognitive experience theory might accommodate it by holding that to believe that *p* is to be disposed to the person-level cognitive experience of judging that *p* (Kriegel 2015, Smithies 2019). And maybe a similar story could be given for desire.

The cognitive experience theory is underdeveloped. What's the relationship between Karl's cognitive experiences and his physical-functional states? In the case of *sensory* experiences, two reductive theories have been tried. Behaviorists and functionalists identify sensory experiences with *functional-dispositional states*. Type-type identity theorists identify them with *categorical brain states*. Might either reductive theory work for *cognitive* experiences?

There is reason to think not. To illustrate, suppose that Karl has the cognitive experience with the built-in content that 68 plus 57 equals 125.

First, cognitive experience theorists would not reduce it to a functional-dispositional state involving his dispositions to use "68", "57" and "plus" in certain ways. For one thing, they hold that such dispositions *underdetermine* whether Karl thinks that 68 plus 57 equals 125 or 68 $quus_g$ 57 equals 125. So Karl's cognitive experience must be a "further state" that picks up the slack when it comes to fixing content. For another thing, Karl's cognitive experience is supposed to be a *categorical* state that *explains* Karl's dispositions to use language.

So perhaps Karl's cognitive experience that 68 plus 57 equals 125 is simply identical with his categorical brain state – type-type identity theory. But a simple Leibniz's law argument rules this out too. On the cognitive experience theory, it is part of the essence of this cognitive experience that it is true iff 68 plus 57 equals 125.[3] By contrast, this is *not* part of the essence of any brain state. For any brain state can be fully characterized in

[3] This doesn't presuppose that this intentional state is a "relation to a proposition". So it is compatible with the nonrelational theory of intentionality defended by Prior (1968: 93ff) and Kriegel (2011: chap.3).

terms of *types of neurons* and the *times*, *directions* and *intensities* at which they fire, without mentioning numbers or the plus-function. Therefore, the property of having a cognitive experience with certain built-in truth-conditions is distinct from (though it might be realized by) the property of undergoing a neural pattern.

So the cognitive experience theory naturally leads to *the further fact view* of thought. Some of Karl's thoughts, with their built-in contents, are not reducible in other terms. They are "further facts". This answers Quine and Kripkenstein. Of course, since everything depends on the physical, such facts *depend on* the physical facts about Karl (e. g. his brain states). But it's dependence without reduction. On a dualist version of the further fact view, the dependence is underwritten by contingent psychophysical laws. On a "physicalist" version, it is underwritten by metaphysically necessary "grounding laws" (more on this in the next subsection). I will assume that the cognitive experience theory is a further fact view.

Here is another question. On the cognitive experience theory, *some* thoughts and episodes of understanding are irreducible cognitive experiences. But which ones? Call this the *scope question*. Maybe it applies only to primitive thoughts with very simple contents closely related to perception (Mendelovici 2018). Then a different story is required for more sophisticated thoughts. But, in fact, proponents of the cognitive experience theory typically apply it to quite sophisticated thoughts: thinking that 68 plus 57 equals 125, that he's a friend, that democracy is in trouble, and so on. This "rich view" fits with the above quotations. It will be my target.

### 2.2 Against the Cognitive Experience Theory: Danglers

I don't think that there really are cognitive experiences with built-in phenomenal contents like *68 plus 57 equals 125* or *democracy is in trouble*. My first argument against the cognitive experience theory is that it leads to an incredible account of content.

For instance, on the cognitive experience theory, Karl's "cognitive experience" constitutes his thinking (and understanding the sentence as meaning) that 68 plus 57 equals 125 rather than that 68 $\text{quus}_g$ 57 equals 125. We saw in the previous section that Karl's cognitive experience must be *distinct from* his standard physical-functional states. Still, it is certainly somehow *dependent on* them: mental changes always depend on underlying purely physical changes. But which ones? One odd view would be that Karl's having the specific cognitive experience that 68 *plus* 57 equals 125 at this time, rather than some other cognitive experience, is somehow dependent on his set of *dispositions* to use "68", "57" and "plus" at that time. But, again, proponents of cognitive experience hold that such functional-dispositional states underdetermine content.

The only option remaining is the *brain-based explanation*: Karl's cognitive experiences depend on his *brain states*, even though they are distinct from those brain states. Strawson and Goff endorse something like it in the quotations above.

The brain-based theory requires "intentional laws" linking brain states and thought-contents, for instance:

> If Karl undergoes so-and-so brain state, then he has the cognitive experience that 68 plus 57 equals 125 (rather a cognitive experience that 68 $\text{quus}_g$ 57 equals 125).

> If Karl undergoes such-and-such brain state, then he has the cognitive experience that someone is his friend (rather than a cognitive experience that someone is his friend*).

These special laws are what "solve" the underdetermination worries due to Quine and Kripkenstein. Given the rich variety of possible cognitive experiences, they must be extremely numerous. Given the further fact view, they are brute "necessary connections between distinct existences". On a dualist version, they are contingent (Graham, Horgan and Tienson 2007: 476). On a physicalist version, they are metaphysically necessary "grounding laws" (Rosen 2010: 132).

The brain-based theory is unorthodox. A standard explanation of why Karl thinks that 68 plus 57 equals 125, rather than that 68 $quus_g$ 57 equals 125, appeals to (i) how he's disposed to use certain (inner or outer) symbols together with (ii) considerations of "naturalness" (e. g. Lewis 1992). By contrast, the brain-based theory holds that Karl simply has a categorical cognitive experience with the built-in content *68 plus 57 equals 125*. And this in turn is directly explained by nothing but his *here-and-now brain state* (embedded in a network of such states) together with a special "intentional law".

Still, the brain-based theory is not unprecedented. John Searle has endorsed such a view:

> Intrinsic intentional phenomena are caused by neurophysiological processes going on in the brain, and they occur in and are realized in the structure of the brain [although] we do not know much about the details of how such things as neuron firings at synapses cause [intentional phenomena] (Searle 1984: 5-6)

Although they neglect the issue, I have argued that friends of cognitive experiences (Kriegel, Siewert, Pitt, etc.) are led to the same theory, requiring special "intentional laws". But I will now argue that this theory is incredible, for a few reasons.

First, J. C. Smart (1959) objected to the complexity of brute "dangling laws" connecting brain states with distinct sensory-perceptual experiences. The brain-based cognitive experience theory is even more complex, requiring a slew of additional "danglers": special (nomic or grounding) laws connecting brain states with distinct cognitive experiences with certain built-in contents.

Second, these intentional laws will be arbitrary. What matters to experience are patterns of neural activity. But the connection between undergoing any pattern of neural activity and the thought ("cognitive experience") that 68 plus (rather than $quus_g$) 57 equals 125 is bound to be arbitrary.

Third, while some of Karl's thoughts have quite determinate contents, others have indeterminate contents. For instance, when Karl thinks that 68 plus 57 equals 125, the content is perfectly precise and determinate (at least if numbers are unique Platonic objects rather than set-theoretic constructions). By contrast, when he thinks democracy is in trouble, the content is quite indeterminate. We need to explain indeterminacy no less than determinacy. How can the brain-based cognitive experience theory explain it? One idea (suggested to me by Philip Goff) is that, while some brain states produce cognitive experiences that have built-in *determinate* contents, other brain states produce cognitive experiences with built-in *indeterminate* contents. To use Siewert's (2011: 264) language, "the part

of what is thought that is fully determined by phenomenal character" is determinate in the one case, and indeterminate in the other. There is nothing more to say. However, I cannot accept that. Surely there is a more illuminating explanation – for example, one appealing to differences in "use plus naturalness" (§3.4).

In sum, the cognitive experience theory leads to the brain-based theory of cognitive intentionality, but that theory is incredible. It violates a plausible desideratum:

> *Minimize danglers.* In explaining the determinacy (and the indeterminacy) of Karl's thoughts, we should minimize brute "necessary connections" between Karl's physical states and his distinct intentional states.

### 2.3 Against the Cognitive Experience Theory: Holism

My next problem for the cognitive experience theory concerns the generally accepted thesis of *holism* about thought:

> *Holism.* There are rough, metaphysically necessary connections between Karl's having a certain thought (e. g. the thought that there is a giant red cube there, the thought that someone is a bachelor, or the thought that 68 plus 57 equals 125) and other things, including perhaps: having the capacity for certain sensory-perceptual experiences, having certain inferential dispositions, having certain dispositions to try do certain things (given certain desires), having certain background linguistic or conceptual abilities, and so on.

Elsewhere I posed a dilemma (Pautz 2010: 366 and 2013: 214ff). Cognitive experience theorists can either reject or accept holism. Either way, I argued, they face serious problems.

In response, Philip Goff (2018) and Uriah Kriegel (2015) reject holism. Instead, they accept "atomism" or "modal independence" for cognitive experiences and so also for thoughts. By contrast, Michelle Montague (2019) and Charles Siewert (2016: sect. 6) accept holism. I want to look at their responses.

Let us start with *rejecting* holism. In fact, cognitive experience theorists are under pressure to take this horn. Typically, distinct existences are freely recombinable. So if Karl's thoughts are really special experiential states that are distinct from all other mental states, and if they are not to be explained in terms of his language-use or dispositions, shouldn't they be modally independent from all these things, contrary to holism?

The problem with an atomistic cognitive experience theory is that it implies the possibility "thought scrambling" and "punctate minds". Let us take these in turn.

*Thought Scrambling*. Karl is a prelinguistic hominid who has a perceptual experience of a rock flying towards him which an enemy tribesman has thrown at him. On the cognitive experience theory, Karl presumably undergoes a *second*, quite different cognitive experience $t_{26}$ which constitutes his *judging* that something is moving towards him. That is, there is a peculiar redundancy. Now suppose later he has an identical perceptual experience of another rock flying towards him (he's having a bad day). Given atomism, on this occasion, the cognitive experience $t_{26}$ might be replaced by another cognitive experience $t_{81}$ which constitutes judging *that there is nothing but a stationary giant cube sphere in front of me*. Still, everything else might be the same: he has nothing but a vivid, clear-as-day visual experi-

ence of a rock moving towards his head, he is afraid, he wants to avoid being hit, and he moves away as a result. There is also no change in his inner speech, since he is a prelinguistic hominid who lacks language and inner speech. Indeed, since cognitive experience is supposed be subtle (otherwise it would be uncontroversial), Karl might not even *notice* that $t_{26}$ has been replaced by $t_{81}$. Despite all this, proponents of an atomistic cognitive experience theory must say that, on the second occasion, when Karl has a vivid, clear-as-day experience of a giant rock headed towards him and so quickly moves away, he is "really" secretly judging *that there is nothing but a stationary giant red cube in front of me.*

Given atomism, the cognitive experience theory also implies that the following could happen. Karl is a modern human with language. Whenever he sees a female fox and says "That is a vixen", he "really" has the cognitive experience that it is *bachelor*. But, because he is screwed up in another way, this causes him to have the cognitive experience that it is a fox and the cognitive experience that it is female. And if someone asks him what kind of thing he thinks is in front of him, he *points to* other female foxes around. So his visual experience, his speech, his inferential dispositions, his pointing behavior, and his behavioral dispositions are exactly as if they would be for someone who believes that it is female fox. Still, an atomistic cognitive experience theory absurdly implies Karl has a deeply secret and super irrational belief that the perceived fox is a *bachelor*.

Or suppose that, on one solitary occasion, when Karl says "2 plus 2 equals 4", instead of having cognitive experience that 68 plus 57 is 125, he has the different cognitive experience that 68 $quus_g$ 57 is 125. (Why couldn't there be such a cognitive experience?) However, suppose that there is no difference in his *other* "cognitive experiences", in his use of the mathematical term "plus" in any possible circumstances, and so on. On the cognitive experience ("further fact") theory, on this occasion Karl secretly thinks 68 $quus_g$ 57 is 125 rather than 68 plus 57 equals 125.

Goff (2018: 103-104) holds that these "thought scrambling" cases are indeed metaphysically possible. He says that the only reason we might think otherwise is that they are nomically impossible and difficult to imagine. I disagree. Thoughts cannot float free from everything else in this way. And I'm not really thinking that these cases are only nomically impossible and confusing this for metaphysical impossibility.[4]

How do settle the issue? Here are a few points in my favor. First, you cannot conceive of such thought-scrambling from the first person. Second, the atomistic cognitive experience theory leads to an absurd form of skepticism. Suppose you say "68+57=125". Maybe you can know that you are having THIS cognitive experience; but how can you be so sure that THIS cognitive experience is a one that determines a plus-content rather one that determines a $quus_g$-content, as in the above "scrambling" case? Third, if we follow Goff

[4] Uriah Kriegel has suggested to me a response that goes beyond Goff's. In line with atomism, he accepts the metaphysical possibility of the "thought scrambling" cases described in the text. For instance, there is a possible case in which, thanks to the momentary insertion of an aberrant "cognitive experience", Karl the prelinguistic hominid, while experiencing a rock flying towards him, has a secret and causally isolated cognitive experience *that there is nothing but a giant red sphere before me*, which leaves absolutely no trace on the rest of his mental life (visual experiences, imagery, inner speech) or on his dispositions to act. But he adds that in such a case there is *also* a sense in which Karl believes the obvious - that there is a rock headed towards him - where that sense is given by something like the holistic "interpretation theory" I will propose in §3.2. This mixed view is interesting. But it still allows for atomistic "thought scrambling" – arbitrary thought insertions that leave no traces. And I find that impossible.

and take the permissive line that such cases are possible but just difficult to imagine, we are led to a general modal skepticism. For instance, why not say that round squares are possible but just difficult to imagine?

*Punctate Minds.* Suppose now that Karl is a disembodied brain-in-the-void (BIV). BIV only has the brain state sufficient for a *single* cognitive experience, namely, the that allegedly constitutes thinking that 68 plus 57 equals 125. BIV just has a little experience-nugget. BIV has no other neural machinery. So BIV is not having, and indeed *could not have*, experiences of a number of things (e. g. two marbles or three musical notes, etc.). BIV also has no *language* or *inner speech* (including arithmetical language).

Goff (2018: 103-104) and Kriegel (2015: 54ff) hold that such a punctate mind is possible, as required by their atomism. But I find it *a priori* impossible.[5] Here is a simple argument for my take (Pautz 2013: 213). To think that 68 plus 57 equals 125, you need concepts of *numbers*. And, to have concepts of numbers, you need to be at least *capable of having* experiences of a number of things (they could be non-veridical experiences or even imagistic experiences). But presumably Goff and Kriegel don't think that the cognitive experience that 68 plus 57 equals 125 is *itself* such an experience (e. g. it is not a visual image of 125 things). And, by stipulation, the BIV has *only* this single experience, and is not even *capable* of having any other experiences. So, it cannot have concepts like *68*, *57*, *plus*, and *125*. Therefore, it cannot have such a thought.[6]

Let us then turn to the second option for friends of cognitive experiences. Instead of rejecting holism, they might try to accommodate it within their theory.

Montague (2019: 195-199) opts for this horn. On her view, thoughts are cognitive experiences that are wholly distinct from each other, from dispositions to try to act, and so on. Nevertheless, they are necessarily connected with each other, and with dispositions to try to act, and so on. So, like me, she thinks that the above-described "scrambling cases" are impossible. And maybe the bad-off BIV couldn't have a solitary cognitive experience that 68 plus 57 equal 125 because it doesn't satisfy certain complex pre-conditions for having this experience, as Siewert (2016: section 6 and 1998: 285) suggests.

Now Montague and Siewert are right to accept holism.[7] But they neglect to address a few arguments against combining holism with their cognitive experience theory.

First, the *missing explanation argument* (Pautz 2013: 215-216; Chudnoff 2015: 120). Given the cognitive experience theory, we cannot *explain* holism. We simply must say, as Montague and Siewert do, that, while some experiences (e. g. color experiences) are *not* necessarily connected with dispositions to act or infer in certain ways, other experiences

---

[5] The atomistic cognitive experience theory also makes a false empirical prediction. The theory says that the same cognitive experiences that we have – for instance the cognitive experience that 68 plus 57 equals 125 – could occur in splendid isolation in the absence of capacities for inner or outer speech (as happens in the BIV). But, presumably, since I'm in control of my thought, I'm in control of my cognitive experiences. So this view predicts that right now I could now close my eyes and choose to have the isolated cognitive experience (and hence thought) that 68 plus 57 equals 125, in the *absence* of inner or outer speech ("68 plus 57 equals 125"). But I cannot do that.

[6] Kriegel (2015: 57) responds to this argument by retreating to a different case in which the punctate mind acquired arithmetical concepts by having had sensory-perceptual experiences of numbers of things *in the past*. But his atomism implies the possibility of the more extreme case described in the text in which a BIV system has a single arithmetical thought without *ever having had* for experiences of numbers of things and indeed without even having the *capacity* for such experiences. And I argue that *this* case is impossible.

[7] Horgan and Tienson (2002: 526; and unpublished ms) also take the holist horn.

("cognitive experiences") *are* in necessarily connected with rich dispositions of this kind. Moreover, all these specific necessary connections have no deeper explanation.[8] This is unexplanatory and complicated.

Second, Montague and Siewert's combination of the cognitive experience theory and holism faces the *argument from very distinct existences*. Take the thought 68 plus 57 makes 125. Given the cognitive experience theory, this is constituted by a special cognitive experience *E*. Further, *E* is not itself an (e. g. an imagistic) experience *N* of any number of things (much less 125 things). It is *totally different* from any experience of a number of things. Now, given a plausible form of holism, it's metaphysically necessary that, if a subject has thought 68 plus 57 makes 125, they are capable of having experiences of numbers of things (otherwise the subject cannot have arithmetical concepts). Thus, the cognitive experience theory and holism imply the following: it's metaphysically necessary that, if a subject purely cognitive experience *E*, the subject must have had or be capable of having a totally different experience *N* of a number of things. But it is *a priori* implausible that there is a metaphysically necessary connection between such very different types of experiences. Holism sits poorly with the cognitive experience theory of thought.

Third, the *argument from prelinguistic limits*. In §1, we noted that Karl is only capable of a limited range of thought without language. Thought beyond that range requires having an outer language – a compositional system of representation. This is a holistic connection, broadly understood. It cries out for explanation. We saw that the inner sentence theory of thought fails to explain it (§1.2). The cognitive experience theory also fails to explain it. In general, experience doesn't require language. So if thoughts are special experiences, shouldn't we be capable of any thought without language? For instance, couldn't a neuroscientist manipulate the brain states of prelinguistic Karl, so that he momentarily has cognitive experience that 68 plus 57 equals 125, or the cognitive experiences you in fact have as you read the Declaration of Independence, without language being involved?

One response is that cognitive experience is "perceiving as", in particular, perceiving *a sentence as meaning that p*. So Karl cannot have the relevant cognitive experiences without language.

But there are two problems with this proposal. First, *some* thought is possible without language. Think of prelinguistic children, or Karl as a prelinguistic hominid (§1.2). So *some* limited range of cognitive experiences must be possible without language (Siewert 1998: 277-278). So the proposal needs revision: some cognitive experiences are possible without language, but for some reason *more sophisticated* cognitive experiences (advanced math, the Declaration of Independence, complex physics, etc.) essentially involve language. But this doesn't *explain* prelinguistic limits.

Second, if we have a cognitive experience with the complex content *that the sentence "68 plus 57 equals 125" means that 68 plus 57 equals 125*, we should in principle be able to have a cognitive experience with the *simpler* content *that 68 plus 57 equals 125*. (Compare: if you experience *that there is red thing next to green thing*, you can experience the simpler

[8] *Pace* Chudnoff (2015: 121), the general thesis of phenomenal holism ("no two partial phenomenal states can be the same if they belong to different total phenomenal states") is logically two weak to entail and explain all the *specific* and *varied* holistic-inferential connections that must obtain between cognitive experiences ("there is a bachelor", "68 plus 57 equals 125"), other cognitive experiences, sensory experiences, and behavioral dispositions.

content *that there is a green thing*.) In that case, prelinguistic Karl should in principle be capable of having a cognitive experience in which he directly "grasps" this complex mathematical content without linguistic mediation, against prelinguistic limits.

### 2.4 Is Cognitive Experience Theory Supported by Introspection?

In sum, the cognitive experience leads to brute laws ("danglers") connecting Karl's brain states with sensible rather than twisted intentional contents (§2.2). And it doesn't fit well with the basic "data" about thought, namely, holism and prelinguistic limits (§2.3).

However, it might be replied that we are stuck with the cognitive experience theory because it is introspectively evident. For instance, suppose you read "2 plus 2 equals 4" or "democracy is in trouble". You understand what is said in a flash. Intuitively, no one could be in the same total phenomenal state as you without understanding the words as meaning that *2 plus 2 equals 4* and *democracy is in trouble*. The experiential state of grasping the meanings of these familiar words is the categorical ground of your disposition to provide certain answers to questions about their meaning and use. Since your sensory-experiential experiences are insufficient for grasping these contents, experiences of grasping them must be special, further experiences. If we didn't have such experiences, reading would be boring.

But there are two problems with the appeal to introspection. First, consider the following *continuum argument*. You read mathematical sentences of increasing complexity ("2 is greater than 1", "2+2=4", etc.). Eventually you get to ones involving very large numbers, imaginary numbers, more sophisticated mathematical functions, and so on. Let us stipulate that, by ordinary standards, you count as "understanding" all the sentences in the series: after all, you understand the constituent expressions, you understand the Arabic number notation, you understand the mathematical functions, and so on. That is, you are disposed to give correct answers (maybe with some effort) when asked to explain what they mean. Now, cognitive experience theorists hold that early in the sequence (e. g. "2 is greater than 1", "2+2=4") you have, over and above your sensory-perceptual experiences of the sentences and your dispositions to explain them, categorical cognitive experiences which consist in "grasping" or "seeing" the precise mathematical contents of the sentences. But, presumably, they will say that, eventually, when the sentences become longer and more abstract, you do *not* have such categorical cognitive experiences which consist in grasping the precise mathematical contents of the sentences. For surely your cognitive experiences are just not that rich and fine-grained. If these sentences had had slightly different meanings in English, you're here-and-now experience of reading them would have been the same. In these more abstract cases, you only have sensory-perceptual experiences of the words and certain dispositions to use and explain them. (That is, their view of such cases resembles the view opponents of cognitive experiences like myself would apply to *all* sentences in the series, even the initial basic ones.) If so, there must be an answer to the "scope question" (§2.1): *where* in the series of sentences did you stop having cognitive experiences in which you grasped the precise mathematical content of the sentence? No answer stands out as clearly correct (including "it's indeterminate but hereabouts"). This is very odd if we can know by introspection when our experiential state determines that we understand certain contents.

Second, I suggest that we can explain away the introspective appeal of the cognitive experience theory.

To see how, let us start with another case. Suppose you enter a room and see a few familiar friends. At first blush, you could not be in the same total experiential state and yet fail to know *who they are*. But, on second thought, this is not the case. In principle, you could have the very same total experiential state (with the same sense of familiarity) but without really having any idea who they are. The reason you might think otherwise is that in the actual situation information about who they are is *easily available*. You can easily open up a dossier of information about any one of them. So you might mistakenly think that all that information is somehow already "there", part of your total experience.

Likewise, when you read "democracy is in trouble", you could easily unpack what these words mean. I suggest that this explains away the appeal of cognitive experience theory. It explains why you might think that all that information is somehow already *there* and part of your experience. But this is an illusion. Seeing familiar words is like seeing old friends. And just as you could have the same total experience of your friends but not really know who they are, so you could have the same total experience of the words "democracy is in trouble" but not really know what they mean. Indeed, you might fail to understand them as meaning *anything at all* and have no idea how to use or define them. For you might have the same total experience of the words but utterly fail to satisfy the functional-dispositional requirements on understanding them as meaning anything (Putnam 1981: 4ff; Pautz 2013: 213; Chudnoff 2015: 147). Of course, you typically "just know" in the moment that you understand the words as meaning something. But, contrary to the cognitive experience theory of thinking and understanding, your total *experience* in the moment itself doesn't entail that you understand them as meaning something.

## 3. OUTLINE OF A MULTISTAGE THEORY OF INTENTIONALITY

In the course of criticizing the reductive externalist program and the phenomenal intentionality program, I laid out several desiderata. For instance, the contents of Karl's sensory-perceptual experiences are internally-determined. His beliefs and desires are connected to his conscious-life. We need to explain the determinacy (and indeterminacy) of thought-content. There are necessary limits to his prelinguistic thought and holistic constraints on his thoughts. Neither the reductive externalist program nor the phenomenal intentionality program adequately accommodates all these desiderata.

I will now outline a multistage theory of intentionality satisfying all the desiderata. Like friends of phenomenal intentionality, I will defend internalist intentionalism for Karl's sensory-perceptual experiences. And I will suggest that his conscious experiences play a crucial role in grounding determinate intentionality. But, like reductive externalists, I will suggest we need a real explanation of the contents of Karl's thoughts. We cannot just say "cognitive experiences do it". My explanation will co-opt some elements of Lewis's account of Karl: his "interpretationism" about Karl's beliefs and desires (1974) and his appeal to "naturalness" (1992).

In outlining my theory, I will pretend that Karl's life spans human history. Diagrammatically, the theory goes as follows:

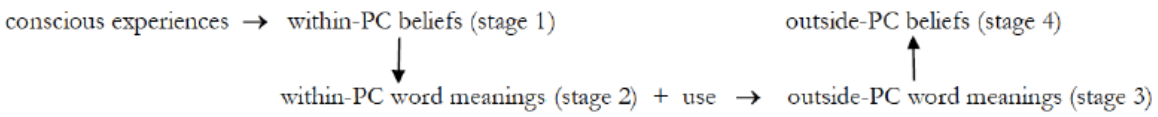


Figure 2: a multistage theory of intentionality

### 3.1 Stage One: Karl's Conscious Experiences with Thin Contents

My multistage theory begins with a view associated with the phenomenal intentionality program: an internalist and nonreductive form of intentionalism about Karl's sensory-perceptual experiences, as opposed to the reductive externalism form of intentionalism criticized in §1.

The quickest way to get a hold of this view is to compare it to the sense datum view defended by Russell in *The Problems of Philosophy* (1912). Suppose Karl views a tomato. On Russell's view, the physical tomato is intrinsically colorless. Karl's brain generates a red-dish and round sense datum, and Karl bears a special relation of conscious acquaintance to this sense datum. The internalist form of intentionalism I favor is similar, but without sense data. It replaces sense data with states of affairs that may or may not obtain. Because of Karl's neural processing, Karl stand in a special, irreducible relation – the *conscious-of* relation – to the ostensible state of affairs of something being reddish and round. In illusion and hallucination, the state of affairs doesn't obtain. It seems to Karl that there is a sense datum but there really isn't one.

Here is a quick argument for this view (Pautz 2019). Given experiential internalism, BIV-Karl (§1.3) might have the same tomato-like experience as Karl. Thus, BIV-Karl might be conscious of ("experientially represent") the (uninstantiated) property of being round. But BIV-Karl's brain state does not have the function of tracking (being produced by) that spatial property. Indeed, BIV-Karl bears no interesting physical relation whatever to the property. So the conscious-of relation (the "experiential representation relation") is not identical with any physical relation. As Ned Block (2019: 426) says, it appears that "we internalists should acknowledge an irreducible representation relation".

So the picture is one of grounding without reduction. Somehow, Karl's brain states ground his being conscious of various states of affairs, but the conscious-of relation is not reducible to any tracking or other physical relation. On a dualist version (Levine 2019), these "grounding" connections are contingent. On a physicalist version (Rosen 2010: 132), they are metaphysically necessary. I'm neutral here.

Internalist intentionalism is consistent with both "illusionism" about the traditional secondary qualities and also with "realism". On a realist version (McGinn 1996), physical things *acquired* colors-as-we-see-them when they came to habitually cause us to have experiences of those colors; the colors of things co-evolved with color experiences. On an illusionist form (Chalmers 2010, Horgan 2014, Pautz 2010), this is not so. Physical surfaces don't have colors-as-we-see them.

Karl also has experiences of acting. These, too, have built-in contents. These contents might have the form: *I'm making so-and-so happen* (Bayne and Levy 2006).

There is a debate about whether the built-in "phenomenal contents" of our experiences "thin" or "rich". In §2, I denied that Karl has special "cognitive experiences" with

built in rich contents involving democracy and large exact numbers. For reasons I cannot go into here (but see Byrne and Siegel 2017), I also reject a "rich" view of Karl's sensory-perceptual experiences on which their phenomenal contents involve high-level properties like *being a tomato*, *being edible*, *expressing fear*, or *being wrong*. Rather, they only involve colors, shapes, movement, and gestalts (abstract complexes of shapes-sizes-colors). Likewise the phenomenal contents of his *pains*, *pleasures*, and *emotional experiences* only concern bodily qualities.

Karl's conscious experiences are a source of reasons. On Pryor's "dogmatism" (2000), if Karl is conscious of ostensible state of affairs *p*, then he thereby has a basic *prima facie* reason to believe that *p*. Karl's experience-based reasons extend beyond the thin contents of his experiences. For instance, his history of experiences provides a reason to think all emeralds are green (rather than grue). Karl's conscious experiences are also source of reasons for desire. For instance, if Karl has a severe pain, he has a basic reason to desire that it go away.

In sum, Karl starts with experiences with relatively thin contents. The next stages of my account propose "extension mechanisms" whereby Karl might move to beliefs, desires, and other intentional states with richer contents.

### 3.2 Stage Two: Karl's Beliefs-Desires within the Perceptual Circle

Imagine that Karl still lacks an outer language. Nevertheless, he has certain basic beliefs and desires. The second stage is a theory of them:

> *Best systems theory*: *If*, given his history of *conscious experiences* and consequent *dispositions to act*, all the *best interpretations* assign to Karl the belief that *p* or the desire that *q*, *then* this grounds Karl having the belief that *p* or the desire that *q*.

To a first approximation, the best systems do the best job overall of maximizing Karl's rationality given his dispositions to act and his conscious experiences.

This account is inspired by Lewis's account in "Radical Interpretation" (1974). It is often called "interpretationist". But this term suggests instrumentalism. So I prefer "best systems theory".[9]

Let us consider an example. Return to the example where a rock is flying towards Karl and so he intentionally moves away (§2.3). Infinitely-many perverse interpretations fit Karl's behavior. One of them is that he believes (despite experiencing otherwise) that the rock is moving *away*, that he wants to be hit on the head, and he believes that by moving away he will magically cause the rock to reverse direction and hit his head. What makes such perverse interpretations incorrect? This is an "underdetermination worry" not unlike those of Quine (1960) and Kripkenstein (1982).

Lewis (e. g. 1986: 38ff; 1994: 427ff) provides an elegant solution. Given his experience as of rock headed towards him, Karl's has a reason to believe that a rock is headed *towards* him. In addition, the desire to be hit on the head is unreasonable. So the above perverse interpretation gratuitously portrays Karl as massively departing from rationality. The interpretation that has Karl departing least from rationality assigns to him the belief that a rock is headed towards his head and a desire that it not hit his head. This is what

[9] For further defense of development of this view, see also Braddon-Mitchell and Jackson 2007: chap. 11.

singles it out as the *correct* interpretation. Call this the *reasons-based solution* to the underdetermination problem.

In general, if we are to avoid underdetermination, we cannot say that the best systems are just a matter of best "fitting the subject's behavior", as on behaviorism. Rather, we must define them as the systems that achieve an optimal balance of maximizing Karl's *substantive rationality* (responding to *experience-based reasons*) as well as his *structural rationality* (behaving so as to maximize expected utility, coherence, etc.). That is, prelinguistic Karl's beliefs and desires fixed jointly by his behavioral dispositions on the output side *and* by the reasons provided by his sensory-perceptual experiences on the input side.[10] Karl's conscious experiences (the first stage) are explanatorily (but not temporally) prior to his basic beliefs and desires.[11]

Since I am no skeptic, I think that Karl's experiences provide him with basic reasons to believe things that somewhat extend beyond the thin contents of his experiences, for instance that all emeralds are blue (rather than grue). This explains how prelinguistic Karl can determinately believe such things.

When prelinguistic Karl's dispositions to act become regularly inharmonious with his history of experiences (e. g. he is mentally ill), the best systems theory will assign him beliefs contrary to experiential evidence. The point is that correct interpretations minimize irrationality, not that they impute no irrationality.

The best systems theory is incomplete. It provides a recipe for determining prelinguistic Karl's beliefs and desires *given* a foundation: a rich set of facts involving his *conscious experiences*, his *actions*, and his *reasons*, all of which must be explanatory prior to his beliefs and desires. So best systems theorists must address the following three questions.

First, what determines Karl's *conscious experiences?* Given intentionalism about experience, this is a special case of the hard problem of intentionality. Call it the problem of *source intentionality*.

Second, what distinguishes Karl's *actions* (which are up for rationalization in terms of belief-desire) from his "mere bodily movements" (which are not). Defining actions as movements that are nondeviantly caused by Karl's desires or beliefs would lead to *circularity*, since his desires and beliefs are precisely what the best systems theory is trying to explain.

Third, where do Karl's *reasons* come from? The best systems theory is up to its ears in normativity. It appeals to facts like "given so-and-so experiences, Karl has a *reason* to believe *p*", "so-and-so prior probabilities are rational", "Karl has a basic reason to desire so-and-so intrinsic values", "failing to maximizing expected utility is irrational", and so on. It is very difficult to provide a plausible (not list-like) reductive account of such notions.

These questions constitute the *source problem* for Lewis's best systems theory (Pautz 2013, Williams 2020). Different versions of the best systems theory result when we plug in different answers.

---

[10] This counts against Schwitzgebel's (2002: 269) claim that "what is for a subject to believe something does not require appeal beyond the subject's forward-looking [output-side] dispositions".

[11] If behavioral duplicates of Karl that work by huge input-output "look-up tables" ("blockheads", "marionettes") lack conscious experiences, my *consciousness-based* best systems theory can avoid the mistaken verdict that they have the same beliefs and desires as Karl. I can also appeal to the constraint proposed by Braddon-Mitchell and Jackson 2007: 120-122.

In Stage One, I advanced nonreductive internalist intentionalism about experience. When we plug this into the best systems theory, we obtain:

> *Nonreductive internalist best systems theory*. The best systems theory combined with nonreductive internalist intentionalism about conscious experience (Chalmers 2010, Horgan 2014, Pautz 2010). A crucial source of intentionality is an irreducible relation of conscious acquaintance.

Therefore, in answer to the question of what determines Karl's conscious experiences, I hold with Russell (1912) that they involve an irreducible, internally-determined relation of conscious acquaintance with ostensible states of affairs. So, while Karl's beliefs and desires reduce to facts about his conscious experiences and dispositions to act, these facts cannot in turn be reduced to the austere physical facts about Karl.

In answer to the question of where prelinguistic Karl's reasons come from, I advocate dogmatism (Pryor 2000). Karl's reasons come from what ostensible states of affairs he is conscious of. And I am not especially concerned to reduce facts about reasons and rationality to something more basic.

In answer to the question of how to define Karl's actions, I suggest that they can be picked out prior to his beliefs and desires, thereby avoiding the above-mentioned circularity worry: they are the doings that he experientially represents himself as making happen (Bayne and Levy 2006).

Of course, the best systems theory comes in other forms. Lewis (1974, 1994) himself hoped for a fully reductive form of the best systems theory according to which *all* these facts about Karl ultimately reduce to the austere physical facts about him. But he never provided the details.[12] In this essay, I'm assuming intentionalism about experience. The only well-developed reductive theories of experiential intentionality are externalist. So reductive best systems theorists are led to:

> *Reductive externalist best systems theory*. The best systems theory combined with externalist intentionalism about experience (Dretske 1995, Tye 2019). A crucial source of intentionality is a tracking relation between Karl's brain states and the world.

In his recent book *The Metaphysics of Representation* (2020), Williams defends a best systems theory along these lines. In particular, he favors a teleological tracking theory (Neander 2017). For instance, suppose Karl views a tomato. Karl is in a brain state that has the biological function of tracking a round thing with a red-reflectance. On Williams' view, this "tracking" fact constitutes his experientially representing that a round and red thing is there (2020: 185ff). This is part of his evidence and constitutes his *reason* to believe that a round and red thing is before him (2020: 181ff). Since he has this reason, the best (most-rationalizing) system assigns to him this belief, rather than some twisted belief.

[12] Indeed, Lewis's reductivism faces big problems. For instance, to avoid deviant interpretations of Karl's desires, he says the best systems will tend to assign "reasonable" desires congruent with "the system of intrinsic values" (1974: 336). But then Lewis (1989) reduces values to what Karl and his community would *desire to desire*. This is circular and still faces deviant interpretations.

My argument for nonreductive internalist best systems theory over Williams' reductive externalist best systems theory is simple. Williams' theory is a form of the reductive externalist program. So it faces versions of the problems covered in §1. In particular, it violates the following desiderata:

Internalism about experiential intentionality
Experiential determinacy

It is only nonreductive internalist best systems theory that accommodates these desiderata.

Start with internalism about experiential intentionality. Research in psychophysics and neuroscience suggests that that *BIV-Karl* could have all the same experiences as Karl, including the same tomato-like experience, even though his brain states don't have the function of tracking anything at all (§1.3). Given that BIV-Karl has all the same experiences as Karl, it is obvious that he is *conscious of* ("experientially represents") the (uninstantiated) shape *round*, has a *reason* to believe that a round thing is there, and (mistakenly) *believes* that a round thing is there. But all this is inconsistent with Williams' reductive externalist best systems theory. What is required is a nonreductive internalist best systems theory (Pautz 2013, 2019).

Next, experiential determinacy. Given Williams' teleological tracking theory, it is arguably indeterminate whether, in the middle earth case, Baby Karl's brain state *B* earth represents blue or yellow (Figure 1, *left*). If he also accepts externalist representationalism about phenomenal character (Dretske 1995, Tye 2019), he must say it is consequently indeterminate whether Baby Karl has a bluish or yellowish experience – which is incoherent.[13] By contrast, nonreductive internalist best systems theory avoids radical experiential indeterminacy. What color quality Baby Karl experientially represents is pinned down by his brain state.

Likewise, Williams' reductive externalist best systems theory lacks a plausible account of Karl on black-and-white earth (Figure 1, *right*). In this case, Karl's visual system tracks$_{17}$ the black-reflectance of the outer object and also tracks$_{18}$ the red-reflectance of the inner object. Williams has two options here. First, *indeterminacy:* it's indeterminate whether the experiential representation relation is identical with tracking$_{17}$ or tracking$_{18}$. So it is indeterminate whether he experientially represents black or red, and therefore indeterminate whether he has a reason to believe that a black thing is there or to believe that a red thing is there. The trouble with this option is that, given intentionalism, it implies indeterminacy concerning whether Karl has a blackish or reddish experience – which is incoherent. Second, *arbitrary identities*: it's just a brute fact that the experientially representation relation is identical with (say) the tracking$_{17}$ relation rather than with the intrinsically very similar tracking$_{18}$ relation. Therefore, his tracking$_{17}$ (representing$_{17}$) the outer black-reflectance is part of his evidence, and gives him a reason to believe that a black-reflectance object is there. But his tracking$_{18}$ (representing$_{18}$) the inner black-reflectance is

[13] Williams says that tracking-representational facts constitute Karl's *evidence* (2020, 181, 185). He doesn't explicitly accept the further claim of externalist intentionalism that they constitute the *phenomenal character* of his experiences. But since it is plausible that Karl's evidence and his phenomenal life are inseperable, he is under pressure to accept this further claim (Pautz forthcoming).

*not* part of his evidence, and *doesn't* give him *any* reason to believe that a black-reflectance object is there. That is, tracking$_{17}$ (representing$_{17}$) has epistemic significance but tracking$_{18}$ (representing$_{18}$) has none at all, even though they are nearly identical. Accordingly, the best (most rationalizing) system assigns Karl the belief that a black-reflectance object is there, rather than that a red-reflectance object is there. But this is intolerably arbitrary. It requires the problematic idea that nearly identical relations can differ radically in their reason-grounding significance (Pautz 2017).

Only nonreductive internalist best systems handles this case without indeterminacy or arbitrariness. The austere physical facts are not the only facts. In addition to bearing the tracking$_{17}$ relation to the back-reflectance of the outer object and the tracking$_{18}$ relation to the red-reflectance of the inner object, Karl bears an internally-determined and irreducible relation of conscious acquaintance uniquely to a certain *sensible color* – say, the color red. This constitutes the determinate phenomenal character of his experience. The conscious-of relation is totally different from any tracking relation, and the sensible color *red* is totally different from any reflectance-type. So we accommodate the evident fact that in this situation Karl's relation to a certain color is totally different from his relation to anything else ("stands out"). And, because the conscious-of relation is totally different from both the tracking$_{17}$ relation and the tracking$_{18}$ relation, we can unproblematically hold that it possesses reason-grounding significance that these relations lack. So we have a more plausible account of how Karl uniquely has a reason to believe that a *red* object is there, and (given the best systems theory) determinately has this belief.

In general, like Russell (1912), I think that an irreducible conscious-of relation plays a crucial role in determining Karl's intentional states. Take a superficial functional isomorph of prelinguistic Karl – Robot Karl – that fails to stand in this relation to any states of affairs. For Robot Karl, the austere physical facts are the only facts. Here there are bound to be many equally good, coordinate "global interpretations" of the contents of Robot's Karl's "perceptions", "evidence", "beliefs", "desires" (Pautz 2017). None "stands out". The only reason why there is a (more or less) determinate, "stand-out" interpretation in the case of the actual Karl (as there surely is) is that, unlike Robot Karl, he bears an irreducible, stand-out relation of conscious acquaintance to various ostensible states of affairs. It's *those* states of affairs that his beliefs are determinately about.

The best systems theory of belief and desire can also accommodate the desiderata violated by the inner sentence theory:

The conscious-life constraint
The constitutive experience-belief connection
Prelinguistic limits

Start with the conscious-life constraint. We saw that the inner sentence theory violates it, allowing for "secret scrambling" of Karl's beliefs and desires while his conscious experiences and behavioral dispositions remain the same. The reason is that it is an *inner-state* theory of belief and desire. That is, in the first instance, it assigns contents (or content-plus-attitudes) to individual subpersonal internal states (e. g. inner sentences), which may be temporarily "secretly scrambled" while retaining those contents. By contrast, my favored form of the best systems theory is *subject-based:* in the first instance it assigns a whole system of beliefs and desires to a *subject-at-time*. Moreover, it does so in a way that

it only sensitive to the subject's *conscious experiences* and consequent *dispositions to consciously act* at that time. So, unlike the inner sentence theory, it rules out "secret scrambling" of Karl's beliefs and desires and satisfies the conscious-life constraint.[14]

Next, the constitutive experience-belief connection. Experience is necessarily compelling. For instance, necessarily, if Karl has an experience of a tomato, he is disposed to believe a red and round thing is there. The inner sentence theory violates this (§1.2). By contrast, the best systems theory accommodates it. Further, on this theory, it isn't a brute fact, but something that can be derived from more general truths. First, it is in the essence of experiences to *provide* reasons for belief. Second, it is in the essence of beliefs to be *responsive* to reasons. Therefore, in the absence of contrary behavioral dispositions, he automatically count as believing the contents of his experiences.

Finally, I am especially impressed by how the best systems theory can explain prelinguistic limits. The basic idea: (i) experience-based, prelinguistic *reasons* are necessarily limited ("epistemic limits"); (ii) prelinguistic belief is constitutively connected to such reasons (the best systems theory); therefore, (iii) prelinguistic belief is necessarily limited.

For instance, suppose that prelinguistic Karl has before him a large pile of sea shell beads that have small holes punched into them. He repeatedly places 3 shells on the ground, and then strings them together into a necklace. Since his experience gives him reason to think that there are 3 shells on the ground, the best interpretation is that he believes that there are 3 shells, and he wants to make necklaces with 3 shells. So that is the correct interpretation.

However, there are limits. For example, suppose that Karl has a magical subpersonal mechanism that responds to a pile of exactly 167 shells, and causes him to vigorously wave his arms when and only when there is such a pile (similar to the example in §1.2). One interpretation is that he truly believes that there are 167 shells on the ground, and he wants to wave him arms when there are 167 shells. But there are many others: for instance, he mistakenly believes that there are 168 shells, and he wants to wave his arms when and only when there are 168 shells. Now here the reasons-based gambit for selecting a unique correct interpretation doesn't work. For, while Karl can have an experience-based reason to believe that there 3 rather than 2 or 4 shells there, his experience just doesn't provide a reason to believe that there are 167 rather than 168. So no specific large-number interpretation can ever stand out as "best" or "most rationalizing". Thus, by connecting beliefs to reasons, the best systems theory *explains* the otherwise puzzling fact that, without an outer language, Karl is necessarily unable to have beliefs about specific large numbers. It explains why, beyond small numbers, his numerical beliefs are *necessarily* only approximate.

More generally, prelinguistic Karl has experience-based reasons to believe things *within a certain range*. Let us call this the *perceptual circle* (PC). As noted above, this range extends somewhat beyond the thin contents Karl's sensory-perceptual experiences – but not too far. (So a better name might be "prelinguistic circle".) They include beliefs about small numbers, the sensible properties of things, Rosch's basic level categories (Fodor

[14] While I favor a *subject-based* version of the best systems theory of prelinguistic Karl's beliefs and desires, Williams (2020: 11, 156) favors an *inner-state* version – in particular, one which assigns contents to inner sentences in a Fodorian language-of-thought. I would argue that such an inner-state theory violates the conscious-life constraint, the constitutive experience-belief connection, and prelinguistic limits, for the same reasons given in §1.2. This is why I favor a subject-based best systems theory.

2010: 29), basic spatial and temporal relations, the recent past and near future, generalizations (all emeralds are green), emotional states, basic kinship relations, and types of actions. But prelinguistic Karl can never have experience-based reasons for beliefs far outside this perceptual circle: for instance, beliefs about large exact numbers, very abstract kinds (e. g. democratic socialism), the laws of quantum mechanics, distant objects and people (Socrates), and so on. For his experiences have quite thin contents, concerning only shapes, colors, movements, propensities of movement, sounds, smells, tastes, bodily states – that's it (§3.1). And the "gap" between this meager input and such outside-the-perceptual circle matters is just too great (even with the help of *a priori* connecting principles). So there can never be a unique best (most rationalizing) system that attributes to prelinguistic Karl such a belief. That is why, on the best systems theory, prelinguistic Karl can never determinately count as having such a belief, no matter what he does. I know of no alternative proposal about the nature of belief and desire that explains prelinguistic limits.

This concludes my argument for nonreductive internalist best systems theory. However, given prelinguistic limits, the best systems theory cannot be the whole story. How might Karl eventually form outside-the-perceptual-circle beliefs about the laws of quantum mechanics and the like? Here I will suggest a different story appealing to outer language. Accordingly, Stage Three is an explanation of linguistic meaning (§3.3). Then Stage Four (§3.4) is an account of how Karl can believe outside-the-perceptual-circle contents by accepting outer language sentences expressing those contents.

### 3.3 Stage Three: An Anchored Use Theory of Linguistic Meaning

Imagine, then, that Karl and his tribe invent a language, which eventually comes to resemble modern English.

What fixes the meanings of sentences and expressions of the language? I favor an *anchored use theory*. Briefly, the meanings of some initial, basic expressions were *mentalistically anchored*. This is congruent with a broadly "head-first" approach to meaning (Lewis 1975, Bennett 1976). But mentalistic anchoring only goes so far. For outside-the-perceptual-circle terms, a different story is required. Let us take these points in turn.

As I said, prelinguistic Karl can have a certain limited stock of beliefs involving matters within the perceptual circle. Since such beliefs are explanatory prior to linguistic meaning, they can be used to help explain the meanings of an initial stock of basic expressions. For instance, perhaps "is red" initially came to mean *is red* by virtue of being conventionally associated with the belief that something is red (Lewis 1975, Bennett 1976). So we have:

> *Mentalistic anchoring*. The limited prelinguistic beliefs of Karl and others helped explain the meanings of an initial stock of basic expressions referring to within-the-perceptual-circle matters. Initially, mental content was prior to linguistic meaning.

This initial stock of mentalistically-anchored terms might have included expressions referring to the following:

> small exact numbers

sensible properties
Rosch's basic level categories
basic spatial and temporal relations
emotional states
basic kinship relations
types of actions

However, given prelinguistic limits, mentalistic anchoring cannot help fix the meanings of expressions outside this list, such as "167", "googolplex", "democracy", "neutrino". For instance, we cannot say that "there are 167 shells in the pile" inherits its content from the explanatorily prior belief that there are 167 shells in the pile. For this would arguably imply that Karl could believe this exact-number content without any outer language. And above I argued that this is not the case.[15]

Therefore, we must supplement mentalistic anchoring with:

> *Non-mentalistic use theory.* For any outside-the-perceptual-circle sentence which means that *p*, the correct account of this cannot invoke an explanatorily prior capacity to believe that *p*, because Karl's community lacks such an explanatorily prior capacity. Rather, it typically appeals to ideal regularities of use.

In my view, such outside-the-perceptual-circle expressions include expressions for:

abstract kinds
larger exact numbers
certain natural kinds
theoretical entities
certain normative properties
distant objects

In sum, the anchored use theory is a mixed view consisting of mentalistic anchoring and a non-mentalistic use theory. Thought initially breathed life into language, helping to inject a modicum of determinacy. Then language took on a life of its own. Expressions came to mean things that Karl could not think about without the help of language.

But how does the non-mentalistic use theory work? Like Horwich (2005), I favor metasemantic pluralism. Typically, ideal regularities of use determine meanings, but they differ for different types of expressions. Let me give some examples.[16]

---

[15] For discussion, see Bennett 1976: 96; Blackburn 1984: 137-140; Lewis 1975: 27; and Avramides 1989: 113ff. I think that prelinguistic limits undermines Lewis's (1975, 1992) two-stage mentalistic approach (taken up by Williams 2020: 149ff) on which the contents of *all* uttered sentences are inherited from explanatorily prior beliefs with those contents and then the contents of the all unuttered sentences can be extrapolated. Mentalistic anchoring only applies to a much more meager set of basic, initial expressions. For the rest of language, we need a separate, non-mentalistic use-theory. The result is a messier, more disjunctive story. But I think it is truer to the facts.

[16] A big part of Horwich's own program is his rejection of standard "truth-referentialism" (truth and reference crucial to explaining meaning). But I would prefer to combine his use theory with truth-referentialism and a standard compositional meaning theory. See Horwich 2005: 44ff for discussion.

*Abstract expressions.* Here I mean "semantically stable" expressions for abstract properties and kinds, such as "agent", "philosopher", "knows", "game", and "democracy". For these expressions, I favor a kind of anchored-hierarchical use theory. Karl and others in his community started with an initial stock of within-the-perceptual-circle expressions $O_1$, $O_2$, $O_3$ . . . They are the "original" or "old" expressions. Their meanings were mentalistically-anchored, and they referred to the types and properties listed above. They enabled Karl's linguistic community to grasp scenarios within the perceptual circle. So they could then introduce new expressions $A_1$, $A_2$, $A_3$ . . . governed by certain ideal regularities of the form:

If [$O_1$, $O_2$, . . .], accept sentence [. . . $A_n$ . . .]
If [$O_1$, $O_2$, . . .], reject sentence [. . . $A_n$ . . .]

Once $A_1$, $A_2$, $A_3$ . . . acquired meanings in this way, they able iterate the process, and introduce new words $B_1$, $B_2$, $B_3$ governed by new ideal regularities of use:

If [$A_1$, $A_2$, $A_3$ . . .], accept sentence [. . . $B_n$ . . .]
If [$A_1$, $A_2$, $A_3$ . . .], reject sentence [ . . . $B_n$ . . .]

Finally, they reach very abstract expressions, such as "game", "democracy" and "supervenience". In this way, expressions of Karl's language come to be associated with very abstract properties that would be outside of his cognitive reach without language.

These ideal regularities of usage are determined by the dispositions of Karl and others to use the expressions $A_1$, $A_2$, $A_3$ . . . in response to experienced scenarios and to *correct* each other's usage. Because their dispositions can be in "error" and only cover finitely many cases, there is a gap between them and the ideal regularities. To close the gap, we must appropriate another Lewisian idea: naturalness as a kind of external constraint (Chalmers MSa: 10).

This anchored-hierarchical picture does not require that the "new" expressions are easily definable in expressions of the "old" expressions (think of "game"). Nor does it require that the original, mentalistically-anchored expressions are rich enough to form an analytic scrutability-base for all truths.[17]

*Mathematical expressions*. As I said, the conscious-based best systems theory explains how prelinguistic Karl could have determinate beliefs about small exact numbers. He might also have beliefs about the *next number*. Then his community invented a system of number words. The meanings of the first few number-words could directly mentalistically anchored. For the rest, speakers had the intention that the *next* number-word in the counting sequence refers to the *next number* (Dehaene 1999, Spelke 2003). In this way, some number-words came to refer to larger exact numbers, like 10. This enabled them to have the intentions required for setting up the Arabic numeral system.

[17] Chalmers's (MSb) defends *anchored inferentialism*. Unlike my own view, his view does presuppose a scrutability thesis. There are other differences. Chalmers applies his theory to "mental concepts" and holds that mental content is always prior to linguistic content, while I give priority to outer language for all outside-the-perceptual-circle content. In addition, Chalmers seeks a uniform theory for all non-basic concepts, while I think (following Horwich 2005) that we must settle for a messy, pluralistic approach.

Once they can think of large exact numbers in this way, they were able to introduce the symbol "+". Even though their dispositions are finite and "error-prone", the simplest or most "natural" ideal rule for this expression is: accept instances of "$x+y=z$" iff "$z$" stands for the number which is equal to *x plus y*. This solves the plus-quus problem (Chalmers MSa: 10).

*Logical constants.* Once the contents of sentences are fixed, we can enter into certain inferential practices ("entry rules and exist rules") with them. And this can fix the semantic values of the logical constants (e. g. Williams 2020: 38ff).

*Certain natural kind terms*. When prelinguistic Karl quenched his thirst, it was perhaps indeterminate whether he wanted to drink *water* (the natural kind) or the *watery stuff* (the surface kind). Then his community came up with a term, "water", that specifically refers to the natural kind rather than the surface kind. This enabled him to have beliefs specifically about the natural kind. Maybe meaning of "water" is constituted by the fact that ideal law for its usage is: accept "*x* is water" iff *x* has the underlying nature of the stuff in our seas, rivers, lakes and rain (Horwich 2005: 27).

*Theoretical expressions.* Perhaps the meaning of "neutrino" is fixed by our underived acceptance of "If there is a type of particles that plays the neutrino-role, then there are neutrinos" (Horwich *ibid.*).

This completes my sketch of a *non-mentalist use theory* for outside-the-perceptual-circle expressions of outer language. My discussion has been short on detail. I have only given a "picture". But everyone needs such a theory. Those who favor an inner sentence theory of belief need a non-mentalistic use theory for expressions of the language of thought, where "use" is understood broadly to include asymmetric dependence (Fodor 1990), conceptual role (Williams 2020), and so on. I rejected the inner sentence theory (§1.2). Instead, I think that the explanation starts with expressions of *outer* language. But, no matter where we start, we are all in the same boat: we all need a non-mentalistic use theory of how our representations came to latch onto some outside-the-perceptual-circle contents rather than others.

### 3.4 Stage Four: Language Extends Belief Beyond the Perceptual Circle

In Stage Three, we saw that, since the beliefs of prelinguistic Karl cannot extend beyond the perceptual circle, there must be a theory (a "use" theory) of how sentences of his language came to mean outside-of-the-perceptual-circle contents *which doesn't appeal to an explanatorily prior ability to believe those contents*. Given this, Stage Four suggests that Karl comes to believe outside-the-perceptual-circle contents by understanding and accepting sentences expressing those contents. In general, language gives Karl a new way of believing:

> *The outer sentence theory of belief.* If Karl understands and accepts an outer sentence *s* that means that *p* in his community, then this grounds his believing that *p*.[18]

[18] In some cases *s* can also be associated with a different content *p** (a "primary intension") in Karl's idiolect on the basis of his individual, idiosyncratic (and often not well-defined) use-dispositions (Chalmers MSb).

So Karl has two ways of believing something. One way is given by the best systems theory introduced in Stage Two (§3.2). Call beliefs grounded in this way *language-independent beliefs*. The other is given by the outer sentence theory. Call beliefs grounded in this way *language-mediated beliefs*. The result is:

> *Pluralism* about belief: To believe that *p* is to satisfy either (i) the best systems condition *or* (ii) the outer sentence condition for believing that *p*.[19]

Now we have an explanation of outside-the-perceptual-circle belief as well as inside-the-perceptual circle belief. Karl believes inside-the-perceptual-circle contents by satisfying the best systems condition. In fact, once he has language, he can also believe the same (or similar) contents by accepting sentences that express those contents. For instance, Karl believes that Friedrich is his friend in a language-independent way; given his experience-based reasons and behavioral dispositions, all the best interpretations assign to him this belief. He believes a similar content in a language-mediated way as well: he understands and accepts "Friedrich is my friend". As for outside-the-perceptual-circle contents, Karl has only one way of believing them: by understanding and accepting sentences that express them. The ideal use regularities for the expressions in Karl's community associate them with increasingly abstract properties and kinds lying farther and farther outside the perceptual circle. By accepting sentences employing those expressions, Karl believes contents he couldn't believe without the help of an outer language (e. g. there are 167 shells there, the laws of quantum mechanics). The capacity for outside-the-perceptual-circle belief (Stage Four) evolved simultaneously with outside-the-perceptual-circle linguistic meaning (Stage Three).[20]

Typically, Karl's language-independent and language-mediated beliefs align, as when he believes that Friedrich is his friend. But sometimes they do not. For example, suppose that Karl hallucinates a human face and acts as if he is afraid. But he knows that he is hallucinating and so accepts "there is no face there". Does he believe that there is a face there or does he believe that there is no face? We feel pulled in different directions (Byrne 2018: 146). I think that the right thing to say is that he believes that there is a face there in a language-independent way and he believes that there is no face there in a language-mediated way. Likewise, when a student in a fraternity initiation trick is threatened with a red-hot poker but in fact is touched on his back with a piece of ice, and then says "That's hot!", his language-mediated belief is mistaken but his language-independent belief about his experiential state is correct (Lewis 1999).

---

[19] Bermudez (2003: 66, 150ff), Dennett (1987: 19, 201, 207, 233) and Speaks (2010: 234ff) defend other forms of pluralism about belief.

[20] In §3.2, I defended "epistemic limits": Karl's *experience-based* reasons are limited to *within-the-perceptual-circle* matters (e. g. *there are 3 shells*). I suggested that this, together with the best systems (reasons-responsive) theory of belief, explains why his prelinguistic beliefs are likewise limited. In that case, we need a different story about the source of his reasons for his more sophisticated, language-mediated *outside-the-perceptual-circle* beliefs (e. g. *there are 1,067 shells*). I accept *epistemic pluralism*. There is more than one source of good epistemic standing. Karl's *outside-the-perceptual-circle* beliefs can be "justified" or "reasonable" in the sense that they are reliably formed, or have a high probability given what he knows in a certain way.

Karl's language-independent beliefs, as given by the best systems theory, are rationally constrained. As for his language-mediated beliefs, there is more latitude here. For instance, if Karl acquires Cotard's syndrome, he might believe that he is dead, by understanding and accepting (in some minimal sense) the sentence "I'm dead". In this way, my pluralist theory allows for irrational belief (*pace* Smithies 2019: 150).

The argument for pluralism (or "disjunctivism") about believing is simple. For *prelinguistic* Karl's beliefs, there is a strong case for a subject-based best systems theory. It satisfies the desiderata that the inner sentence theory violates (§3.2). But, as we saw, this theory will only work for Karl's within-the-perceptual-circle beliefs. So, for Karl's outside-the-perceptual-circle beliefs, we need a different theory. Here the outer sentence theory fills the bill. The resulting pluralist theory explains why some thought is possible without an outer language but other forms of thought require an outer language. And, unlike the inner sentence theory of belief, it *remains in line with the conscious-life constraint*, because it only appeals to Karl's conscious experiences and dispositions to consciously act – now including his dispositions to "accept" *outer* sentences.

The pluralist theory says that one way of believing that *p* is by *understanding* and *accepting* a sentence meaning that *p*. How to explain these relations?

Take *understanding* first. I think that no general and simple analysis is possible. The conditions on understanding an expression differ for different types of expressions. For instance, the conditions required for understanding logical expressions differ from the conditions for understanding moral expressions. For some expressions (e. g. "democracy"), counting as understanding them might typically require understanding some other, more basic expressions. The conditions on understanding are never hard and fast. Understanding *admits of degree*. That is why there is no clear answer to the question of whether a 6-year old who says "daddy is a physicist" really thinks that daddy is a *physicist*.[21]

Even though conscious experience has only thin content (§3.1), it anchors all understanding. A robot with no experiences doesn't really understand any words, even if there is a sense in which the robot's words play similar inferential roles to our words. For instance, to understand "167", you need to know what a number is. And that requires having the capacity to have experiences of numbers of things. This is another respect in which consciousness is essential to my account.

Now turn to *accepting*. Since the outer sentence theory proposes that believing an outside-the-perceptual-circle content *p* is to be explained in terms of accepting a sentence that means that *p*, I cannot on pain of circularity say that accepting a sentence that means that *p* is to be explained in terms of believing that *p*. The outer sentence theory requires acceptance is characterizable in belief-independent terms, since it is used to explain how Karl believes outside-the-perceptual-circle contents. (Similarly, since the *inner* sentence theory (Fodor 1990) explains believing that *p* in terms of accepting* - that is, having in one's belief-box - an inner sentence that means that *p*, it requires a belief-independent account of this.) I have no general analysis up my sleeve. As with understanding a sentence, I think that the conditions on accepting a sentence differ for different types of sen-

[21] By contrast, since whether you have a certain cognitive experience with built-in content *p* is presumably a binary, nongraded matter, the cognitive experience theory has the implausible implication that there is a form of understanding or grasping that it is "on-off" and doesn't admit of degree. See Bourget 2015 for an interesting discussion.

tences. Often accepting a sentence involves a disposition to use the sentence in reasoning and planning. Although I have no belief-independent account of accepting a sentence up my sleeve, I am confident that accepting is indeed prior to believing when it comes to outside-the-perceptual-circle matters, so that such an account must be possible in principle.

I have proposed a pluralist theory of the state of believing. What about the activity of thinking? Karl can also count as thinking that *p* in multiple ways. For instance, he can engage in spatial thinking using mental imagery. However, Karl's outside-the-perceptual-circle thinking is generally realized by "inner speech", understood as a quasi-perceptual process representing outer speech (Carruthers 1996: chap. 2; Byrne 2018: 198ff). In these cases, the content of the Karl's thought is just the content (in the context) of the imagined sentence, or a sentence he would take to elucidate it. Since the content of the sentence (e. g. "all the beers are in the fridge", "democracy is in trouble") is bound to be indeterminate and incomplete in various ways, so is the content of Karl's thought.

The resulting pluralist view of belief and thought is superior to the cognitive experience theory of thought that goes with the standard phenomenal intentionality program. We saw in §2 that the cognitive experience theory fails to adequately accommodate the following desiderata:

> Minimize danglers
> Holism
> Prelinguistic limits

By contrast, my pluralist theory of thought nicely accommodates these ideas.

First, the pluralist theory minimizes danglers. To explain how Karl believes sensible contents rather than deviant contents, it needs no special "intentional laws".

For example, return to Stage Two in which Karl lacks an outer language. He might believe that someone is a friend, rather than a friend*. How so? The cognitive experience theory appeals to a special cognitive experience with the built-in content *he is my friend*, together with a special brute intentional law linking this experience to his brain state (§2.2). By contrast, my pluralist theory requires no such special intentional law. For such language-independent beliefs, I accept the *best systems theory*. Karl's history of experiences gives him a stronger reason to believe that the person is a friend (more natural) than to believe that the person is a friend* (less natural). Compare: his experiences give him a stronger reason to believe that emeralds are blue (more natural) than to believe that they are all grue (less natural). The best systems (reasons-based) theory uses this generally-accepted epistemic fact to explain why Karl believes that the person is a friend, rather than a friend*.[22]

Recall that we must explain indeterminacy as well as determinacy (§2.2). For instance, in Stage Four, when Karl thinks that 68 plus 57 equals 15, the content of his thought is perfectly precise and determinate. By contrast, when he thinks democracy is in trouble, the content of his thought is quite indeterminate. We saw that cognitive experience theo-

[22] See e. g. Lewis 1986: 38ff and 1994: 427ff. Lewis is often associated with a toy "use plus naturalness" theory, which uses "naturalness" as a *basic* constraint (and which gives priority to language). In fact, in the case of mental content, he explicitly *derives* his "naturalness constraint" from his more general best systems or "reasons-based" theory of mental content (Pautz 2013: 222; Williams 2020: 62). Some objections to the toy theory do not apply to Lewis's actual view (Pautz 2013: 221-222; Dorr 2019: sect. 4.7).

rists can only say "some brain states produce cognitive experience with determinate contents while other brain states produce cognitive experiences with indeterminate contents". By contrast, my pluralist account provides a real explanation. In the case of such outside-the-perceptual thoughts, I accept the outer sentence theory. Karl doesn't grasp any such contents merely by having certain "cognitive" experiences. Rather, he grasps the contents only by understanding and accepting the sentences - "68 plus 57 equals 125" and "democracy is in trouble" - which mean those contents. This in turn involves use-dispositions and not just his experience at the time (§2.4). Our use of mathematical expressions is highly constrained, and here there is a very "natural" and simple use-rule that fits use, namely the "plus" rule (Chalmers MSa: 10).[23] By contrast, our use of "is a democracy" is less constrained, and here there are many equally natural ideal laws of use that fit our use dispositions, corresponding to different precisifications of "democracy". That explains the difference in content-determinacy without special intentional laws.

Next, holism. Karl's thoughts are holistically bound up with other things – attempts to do things, other thoughts, sensory-perceptual experiences, language-use, and so on. The cognitive experience of thought can accommodate this only by positing a slew of brute and implausible necessary connections between Karl's "cognitive experiences" and these other things (§2.3). By contrast, because my pluralist theory eliminates cognitive experiences, it avoids the need to posit such special, brute necessary connections. In effect, it reduces thoughts with rich contents to complex holistic conditions involving actual and potential sensory-perceptual experiences with thin contents. Holism is a trivial consequence.

Finally, prelinguistic limits. We saw that the cognitive experience theory makes prelinguistic limits totally mysterious. By contrast, my pluralist ("disjunctivist") theory elegantly explains it. The best systems theory uses "epistemic limits" to explain why prelinguistic Karl's thoughts are necessarily limited and why the limits are what they are (§3.2). The only *other* way of having thoughts is given by the outer sentence theory. That is why Karl's thoughts beyond these limits are necessarily language-mediated.

### 3.5 Credo: Thin Experience Reductivism

Let *thin experience reductivism* be the claim that all the mental facts about Karl – including all the intentional facts – reduce to (i) facts about actual and potential sensory-perceptual-emotional experiences with *thin* contents ("thin experiences") and (ii) the functional-behavioral facts about him (including his linguistic dispositions and "wide" functional facts involving his relation to his environment). On this view, the only experiences that must be mentioned in the reductive base for Karl's mental life are his experiences with thin contents. Let the *further fact view* be any view on thin experience reductivism fails. The cognitive experience theory of thought (§2.1) is an example.

---

[23] Horgan and Graham (2010: 328-329) object that an external naturalness constraint would need to be an *extra* "brute fact". (Philip Goff also pressed this objection in discussion.) But all of us *already* believe that *plus* is more natural than *quus*$_g$. So the plus-interpretation "stands out" – and this is the core intuition. Thus, in fact, the naturalness-based solution to underdetermination doesn't require belief anything "extra" beyond what we already accept. Indeed, it is rather Horgan and Graham's *own* solution that requires something extra: a special intentional law (dangler) linking Karl's brain state to his alleged "cognitive experience" that 68 plus (rather than quus$_g$) 57 equals 125 (as we saw in §2.2).

My multistage theory starts with Karl's thin experiences. So it is congruent with thin experience reductivism. Since supplying specific reductive analyses is difficult, any such specific form of thin experience reductivism will be controversial. However, we can offer two general arguments that *some* form of thin experience reductivism is right.

First, the *argument from small steps*. To illustrate, consider mathematical thought. In Stage One, Karl certainly starts off with only thin experiences. So *initially* thin experience reductivism is true. For instance, he repeatedly places 3 shells on the ground, and then strings them together into a necklace. This is *sufficient* for his judgement *that there are three shells there*. A further fact – for instance, a mysterious "cognitive experience" – is not required. Next suppose that he learns a body-based "language" in which he points to different parts of his body (starting with the fingers) to indicate different numbers. He points at a pile of shells and then points at his big toe, thereby communicating the thought *there are 29 shells in the pile* (Dehaene 1997: 93-95). Intuitively, all this might only involve Karl having "thin experiences" of parts of his body. It needn't involve his having, at some specific moment, a totally novel "cognitive experience" with the built-in content *there are 29 shells in the pile*. Finally, suppose that he gradually learns a base-ten number system and different function-names ("plus", "minus", etc.). One day he thinks *68 plus 57 equals 125*. Again, intuitively, at no single step in this process does Karl need to have a wholly novel *kind* of *experience* – a "cognitive" experience - with a built-in rich content *68 plus 57 equals 125*. Intuitively, it's enough that he has thin experiences of new symbols and gradually becomes increasingly competent in using them. The result: merely Karl's thin experiences and increasingly sophisticated linguistic behavior can constitute his thinking *there are 3 shells*, *there are 29 shells*, and *68 plus 57 equals 125*.

Second, a more elaborate *supervenience argument*, which will proceed in two steps. First, thin experience *supervenience* is plausible. Second, thin experience supervenience supports thin experience *reductivism*.

First, thin experience supervenience is supported by reflection on *duplication cases*. Suppose that on different occasions Karl has various mental states. He has an experience of a rock flying towards him and moves out of the way. He believes that a rock is flying towards him. He says "68 plus 57 is 125" and believes that 68 plus 57 is 125. He gets a paycheck and says, while pointing to his financial institution across the street, "I'm bringing this check to the bank", meaning that he is bringing his check to that financial institution. He is happy-go-lucky and says "the future looks bright" and believes that the future looks bright. Now consider *Twin Karl*, a *thin experience duplicate* of Karl. He (i) has all the sensory-perceptual-emotional experiences with *thin* phenomenal contents as Karl and (ii) is like Karl as regards all functional-behavioral facts.

Given the Karl and Twin Karl are *inner-outer* duplicates in all these respects, could their beliefs, thoughts and desires nevertheless radically differ? For instance, could Twin Karl "really" secretly believe the *negations* of everything Karl believes, despite having all the same thin experiences, saying all the same things, and having all the same dispositions as Karl? Or, when he has the same vivid experience of rock flying towards him, could Twin Karl differ from Karl in secretly and irrationally thinking that the rock is moving *away* (§3.2), even though he ducks, says it is headed towards him, and so on? When he says "68 plus 57 is 125", could he differ from Karl in "really" thinking 68 $quus_g$ 57 is 125, even though he does sums the same way as Karl under all possible conditions? When he gets his paycheck and says, while pointing to his financial institution across the street,

"I'm bringing this to the bank", could he differ from Karl in that he "really" means he is bringing it to an *embankment*, even though (like Karl) all his verbal and behavioral dispositions are appropriate to the financial-institution interpretation (Siewert 1998: 279ff)? Could he "really" secretly think the future looks dark, even though all his thin-experiences, speech and dispositions are happy-go-lucky? I do not find such radical variation between Twin Karl and Karl to be clearly conceivable. This supports thin experience *supervenience*.

It might be said that, although Twin Karl's thoughts could not radically *differ from* Karl's, it is conceivable that he should be a "cognitive zombie" who *lacks* all thought, contrary to thin experience supervenience. For instance, Terry Horgan (Horgan 2013: 243-244) says that Twin Karl might be a mere perfect "symbol manipulator" who doesn't really *understand* any English sentences. If Twin Karl might be a complete thin-experiential-cum-functional duplicate of Karl and yet lack understanding, then states of understanding must be elusive "further facts" – for instance, special "cognitive experiences".

But this is not clearly conceivable. To see this, start with rudimentary contents, as in the sequence argument above. Here Horgan's claim is very implausible. For instance, it is quite clear that, just by virtue of having the same sensory-perceptual experience of 3 shells on the ground and the same behavioral dispositions as Karl, Twin Karl will also perfectly well understand the content *there are 3 shells there*. Further, as we go up the conceptual ladder in small steps, at each point, changes in thin experiences and linguistic competence are intuitively enough for changes in thought and understanding. So Twin Karl must think and understand the same things as Karl.

Thin experience supervenience, then, is plausible. The next step of the argument says that thin experience supervenience supports thin experience *reductivism* over the further fact view. After all, if some thoughts and states of understanding were really "further facts" or extra "cognitive experiences" (Kriegel, Siewert, Goff), we would expect they *could* radically differ between Karl and Twin Karl in the above-mentioned ways, while holding everything else fixed, contrary to thin experience supervenience. But we saw this is inconceivable. By contrast, thin experience reductivism offer a simple explanation of thin experience supervenience.

Here is an analogy (Lewis 1994: 413). Take a black-and-white pixel-screen. The gestalt properties of the screen (*containing a square*, *containing a happy-face*) *supervene on* the arrangement of black and white pixels. This suggests that they *reduce* to such arrangements.

Because of the "hard problem of consciousness", Karl's conscious experiences with thin contents are mysterious. But, if we accept thin experience reductivism rather than the further fact view, then we can rest assured that Karl's other mental states with "richer" contents (e. g. the thought that 68 plus 57 equals 125) pose no *additional* profound mystery. To explain them, we don't need to posit dangling "intentional linking laws" (§2.2). For, although it's hard to supply the details, we know that they *somehow* reduce to patterns in Karl's actual and possible thin experiences and relations to the world, just as gestalt features of the screen reduce to patterns of black-and-white.

## 4. CONCLUSION

I argued that both the reductive externalist program and the phenomenal intentionality program miss out on certain desiderata on an adequate theory of intentionality (§§1-2). Then I sketched a multistage theory of intentionality that does satisfy them (§3). Maybe it is along the right lines.[24]

### *References*


Avramides, A. 1989. *Meaning and Mind*. Cambridge, MA: MIT Press.

Bayne, T. and N. Levy. 2006. The Feeling of Doing: Deconstructing the Phenomenology of Agency. In N. Sebanz and W. Prinz, eds. *Disorders of Volition*. MIT Press. 49–68.

Bennett, J. 1976. *Linguistic Behavior*. Cambridge: Cambridge University Press.

Bermudez, J. 2003. *Thinking Without Words.* Oxford: Oxford University Press.

Blackburn, S. 1984. *Spreading the Word*. Oxford: Oxford University Press.

Block, N. 2019. Arguments Pro and Con on Adam Pautz's External Directedness Principle. In A. Pautz and D. Stoljar, eds. *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*. MIT Press. 421-426.

Bourget, D. 2015. The Role of Consciousness in Grasping and Understanding. *Philosophy and Phenomenological Research* 95: 285-318.

Braddon-Mitchell, D. and F. Jackson. 2007. *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Byrne, A. 2018. *Transparency and Self-Knowledge*. Oxford: Oxford Univesity Press.

Byrne, A and S. Siegel. 2017. Rich or Thin? In B. Nanay, ed. *Current Controversies in Philosophy of Perception*. Routledge. 59-80.

Carruthers, P. 1996. *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.

Chalmers, D. 2010. *The Character of Consciousness.* Oxford: Oxford University Press.

Chalmers, D. MSa. Reference Magnetism and the Grounds of Intentionality. Available at <http://consc.net/books/ctw/excursus20.pdf>

Chalmers, D. MSb. Inferentialism and Analyticity. Available at <http://consc.net/books/ctw/excursus19.pdf>

Chudnoff, E. 2015. *Cognitive Phenomenology*. Routledge: London & New York.

Dehaene, S. 1999. *The Number Sense*. Oxford: Oxford University Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dorr, C. 2019. Natural Properties. *The Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/archives/fall2019/entries/natural-properties/>.

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

Fodor, J. 1990. *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, J. 2010. *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Goff, P. 2012. Does Mary Know I Experience Plus rather than Quus? *Philosophical Studies* 160: 223-235.

Goff, P. 2018. Conscious Thought and the Cognitive Fine-Tuning Problem. *Philosophical Quarterly* 68: 98-122.

Graham, G., T. Horgan and J. Tienson. 2007. Consciousness and Intentionality. In M. Velmans and S. Schneider, ed. *The Blackwell Companion to Consciousness*. Blackwell. 468-484

Hawthorne, J. 2006. *Metaphysical Essays*. Oxford: Oxford University Press.



[24] Earlier versions of this essay were presented at Rice University, the University of Cambridge, the University of Leeds, and the Institut Jean Nicod. Thanks to the audiences on those occasions for helpful feedback. I am also grateful to Jacob Beck, David Chalmers, Jim van Cleve, and Robbie Williams for comments on an early draft. Thanks especially to Uriah Kriegel for detailed and very helpful comments on the penultimate draft.

Horgan, T. 2013. Original Intentionality is Phenomenal Intentionality. *The Monist* 96: 232-251.

Horgan, T. 2014. Phenomenal Intentionality and Secondary Qualities: The Quixotic Case of Color. In B. Brogaard, ed.) *Does Perception Have Content?* Oxford University Press. 329-350.

Horgan, T. and G. Graham. 2010. Phenomenal Intentionality and Content Determinacy. In R. Shantz, ed. *Prospects for Meaning*. R. Amsterdam: de Gruyter. 321–344.

Horgan, T. and J. Tienson. 2002. The Intentionality of Phenomenology and the Phenomenology of Intentionality. In D. Chalmers, ed. *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press. 520-533.

Horgan, T. and J. Tienson. Unpublished. Phenomenal Intentionality and Phenomenal Holism.

Horwich, P. *Reflections on Meaning*. Oxford: Oxford University Press.

Hurford, J. 2014. *The Origins of Language*. Oxford: Oxford University Press.

Kriegel, U. 2011. *The Sources of Intentionality*. Oxford: Oxford University Press.

Kriegel, U. 2015. *The Varieties of Consciousness*. Oxford: Oxford University Press.

Kripke, S. 1982. Wittgenstein on Rules and Private Language.

Levine, J. 2019. On the Meta-Problem. *Journal of Consciousness Studies* 26: 148–159.

Lewis, D. 1974. Radical Interpretation. *Synthese* 21: 331–344.

Lewis, D. 1975. Languages and Language. In K. Gunderson, ed. *Minnesota Studies in the Philosophy of Science*, Volume VII. University of Minnesota Press. 3–35.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society* 63: 113–137.

Lewis, D. 1992. Meaning Without Use. *Australasian Journal of Philosophy* 70: 106–110.

Lewis, D. 1994. Reduction of Mind. In S. Guttenplan, ed. *A Companion to Philosophy of Mind*. Blackwell. 412–431.

Lewis, D. 1999. Letter to Timothy Williamson. Available at <http://www.projects.socialsciences.manchester.ac.uk/lewis/letter-of-the-month-november2018/ >

McGinn, C. 1996. Another Look at Color. *Journal of Philosophy* 93: 537-53.

Mendelovici, A. 2018. *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.

Montague, M. 2019. Cognitive Phenomenology, Sensory Phenomenology, and Rationality. In A. Sullivan, ed. *Sensations, Thoughts, Language: Essays in honor of Brain Loar*. Routledge.

Neander, K. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Pautz, A. 2010. Do Theories of Consciousness Rest on a Mistake?
*Philosophical Issues* 20: 333-367.

Pautz, A. 2013. Does Phenomenology Ground Mental Content? In
U. Kriegel, ed. *Phenomenal Intentionality*. Oxford University Press. 194– 234.

Pautz, A. 2017. The Significance Argument for the Irreducibility of Consciousness. *Philosophical Perspectives* 31: 349-407.

Pautz, 2019. How Can Brains in Vats Experience a Spatial World? A Puzzle for Internalists In A. Pautz and D. Stoljar, eds. *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*. MIT Press. 379-420.

Pautz, A. forthcoming. Review of Williams' *The Metaphysics of Representation*. *Mind*.

Pinker, S and R. Jackendoff. 2005. The Faculty of Language: What's Special about It? *Cognition* 95: 201-236.

Pitt, D. 2004. The Phenomenology of Cognition. *Philosophy and Phenomenological Research* 69: 1-36.

Prior, A. N. 1968. Intentionality and Intensionality. *Proceedings of the Aristotelian Society*, Supplementary Volumes, Vol. 42: 73-106.

Pryor, J. 2000. The Skeptic and the Dogmatist. *Noûs* 34: 517-549.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. 1960. *Word and Object.* Cambridge, MA: MIT Press.

Rosen, G. 2010. Metaphysical Dependence: Grounding and Reduction. In R. Hale and A. Hoffman, eds. *Modality: Metaphysics, Logic, and Epistemology*. Oxford University Press. 109–136.

Russell, B. 1912. *The Problems of Philosophy*. London: Williams and Norgate.

Sacks, O. 1989. *Seeing Voices.* Berkeley: University of California Press.

Schwitzgebel, E. 2002. A Phenomenal-Dispositional Account of Belief. *Noûs* 36: 249-275.

Searle, J. 1984. Intentionality and its Place in Nature. *Synthese* 61: 3-16.

Shoemaker, S. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

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

Siewert, C. 2011. Phenomenal Thought. In T. Bayne and M. Montague, eds. *Cognitive Phenomenology*. Oxford University Press. 231-267.

Smart, J. C. 1959. Sensations and Brain Processes. *Philosophical Review* 68: 141-156.

Smithies, D. 2019. *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.

Speaks, J. 2010. Explaining the Disquotational Principle. *Canadian Journal of Philosophy* 40: 211-238.

Spelke, E. 2003. What Makes Us Smart? Core Knowledge and Natural Language. In D. Getner & S. Goldin-Meadow, eds. *Language in Mind: Advances in the Study of Language and Thought*. MIT Press. 277-311.

Strawson, G. 2010. *Mental Reality, 2nd ed*. Cambridge, MA: MIT Press.

Tye, M. 2019. Homunculi Heads and Silicon Chips: The Importance of History
to Phenomenology. In A. Pautz and D. Stoljar, eds. *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*. MIT Press. 545-570.

Varley, R., N. Klessinger, C. Romanowski, & M. Siegal. 2005. Agrammatic but Numerate. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3519–3524.

Williams 2020. *The Metaphysics of Representation*. Oxford: Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.