# Situating Language and Consciousness



Mahrad Almotahari
B.A., Reed College (2005)

Submitted to the Department of Linguistics & Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September, 2011



Signature of Auth·

Department of Linguistics & Philosophy
September, 2011

Certified by:

Robert C. Stalnaker
Laurance S. Rockefeller Professor of Philosophy
Thesis supervisor

Accepted by:

Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Students

SITUATING LANGUAGE AND CONSCIOUSNESS

by

Mahrad Almotahari

Submitted to the Department of Linguistics & Philosophy on September, 2011 in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology

**Abstract**

Language and consciousness enrich our lives. But they are rare commodities; most creatures are languageless and unconscious. This dissertation is about the conditions that distinguish the haves from the have-nots.

The semantic properties of a natural language expression are determined by conventions governing the way speakers use the expression to communicate information. The capacity to speak a language involves highly specialized (perhaps even modular) cognition. Some authors think that one cannot consistently accept both views. In Chapter 1 ('Content and Competence') I explain why one can.

According to the convention-based theory of content determination, propositions are fit to be the contents of both thought and speech. Recently, this view has been challenged. The challenge exploits a series of observations about what it takes to understand semantically incomplete sentences. In Chapter 2 ('Speaker Meaning in Context'), I explain how the challenge can be met.

Physicalists seem to owe an explanatory debt. Why should psychophysical relations appear contingent? In Chapter 3 ('There Couldn't Have Been Zombies, but it's a Lucky Coincidence That There Aren't') I pay the debt on their behalf. My explanation proceeds in three steps. First, I observe that there are necessary coincidences, or accidents. Second, I show that traditional epistemological arguments for dualism merely establish that phenomenal states and corresponding physical states are accidentally, or coincidentally, related. Finally, I suggest that inattention to the distinction between coincidence/accidentality and contingency results in frequent equivocation. Thus the disposition to (correctly) judge that psychophysical relations are coincidences

manifests itself as a disposition to (incorrectly) judge that psychophysical relations are contingent.

In Chapter 4 ('Zombies are Inconceivable') I deny that psychophysical relations appear contingent. The chapter begins with an argument to the effect that zombies cannot be coherently conceived. I then consider and reject various ways of resisting the argument.

Thesis supervisor: Robert C. Stalnaker
Title: Laurance S. Rockefeller Professor of Philosophy

# Acknowledgments

Thanks to my advisors: Agustín Rayo, Bob Stalnaker, and Steve Yablo. In my effort to win their approval, this dissertation improved significantly. Many friends helped me along the way. I'm grateful to James Bondarchuk, Salvatore Florio, Ephraim Glick, Dan Hagen, Adam Hosein, Nick Jones, Heather Logue, Bernhard Nickel, Dilip Ninan, and Damien Rochford. Some friends deserve special recognition. I probably spent more time philosophizing with, and learning from, Ephraim, Adam, and Damien than any of the other members of the MIT Linguistics and Philosophy program. ('Any' ranges over faculty as well as students.) Damien looked at early drafts of every chapter and provided helpful advice and encouragement. James read multiple drafts of the third chapter and suggested that I title my dissertation *I Might Not Have Finished, But It's a Lucky Coincidence That I Did*.

Judy Thomson wasn't part of my committee, but she spent a lot of time teaching me how philosophy ought to be written. I'm still far from mastering the art, but I feel fortunate to have been taught by someone who has. Thanks, Judy!

Thanks also to Dorothy Edgington, Ken Gemes, Jen Hornsby, Øystein Linnebo, and Ian Rumfitt for feedback at a later stage.

I'm *immensely* grateful to my parents: Karim and Tayebeh. Their love and support was constant. This accomplishment is as much theirs as it is mine.

Early drafts of the first chapter were written in the summer of 2009, when Iranians peacefully took to the streets in order to protest fraudulent elections. A mature draft of the third chapter was written in the winter of 2010, when revolutionaries in Tunisia and Egypt brought an end to decades of despotic rule. As I write this, rebels in Libya, rioters in Yemen and Syria, and protestors in Bahrain struggle to achieve the same objective. I dedicate my thesis to the brave Iranians, Tunisians, Egyptians, Libyans, Yemenites, Syrians, and Bahrainis who have suffered for freedom's sake.

# Contents

CHAPTER 1

# Content and Competence

Natural language expressions have semantic properties. 'Mahmoud Ahmadinejad' refers to Mahmoud Ahmadinejad; 'is a fool' expresses, if you like, the property of being a fool; and 'Mahmoud Ahmadinejad is a fool' is true iff Mahmoud Ahmadinejad is a fool. The relationship between an expression of natural language and its semantic properties is not primitive; it is fixed by non-semantic facts about language users (or so it is natural to think). One wants to know, then, what such facts are. Call this *the metasemantic question*, or:

(MQ) What are the more basic facts about us in virtue of which natural language expressions have semantic properties?

This is the first of two questions that we will explore.

Competent language users can interpret indefinitely many novel sentences. Take the longest declarative sentence ever uttered in the English language; add to the front of it 'Mahmoud Ahmadinejad thinks that' and you get a sentence that no one has ever come across before but that any competent English speaker would be able to understand. We possess this virtually limitless capacity despite significant cognitive limitations. One wonders, then, what the capacity consists in. More generally, what does competence with a language involve? This is *the psycholinguistic question*, or:

(PQ) What is it to be a competent language user?

As stated, the question is ambiguous. On the one hand, it can be understood to mean, what must competent language users know, or what are the distinctive abilities they must possess? On the other hand, the question can be understood to mean, what sort of

psychology underwrites this knowledge or set of abilities in actual speakers/interpreters? I have the latter interpretation in mind; that is why I call it the *psycho*linguistic question.[1]

Our questions have been the focus of much attention in the philosophies of language and mind, foundational linguistics, and cognitive science. I want to take a critical look at two dominant answers. I think there is something attractive and something unattractive about both, and the respect in which each is unattractive recommends part of the other view. One half of my project, then, is to motivate a synthesis: content is fixed by conventions governing use, and competence is a complex, special-purpose component of one's psychology that endows the speaker/interpreter with a particular kind of know-how. She knows how to (non-accidentally) exploit content-fixing conventions without participating in them. The other half of my project is to identify and criticize a prevalent assumption blocking the synthesis.

## 1.1 Two questions asked, two sets of answers given

The first set of answers to the questions above draws on work by Noam Chomsky. Famously, Chomsky proposed that competence with a language is a complex psychological state embodied in a special-purpose "language faculty" that represents a finite lexicon and implements a finite set of recursive rules. Together, the rules and lexicon constitute the speaker/interpreter's "I-language". When the language faculty is activated by, say, an English utterance, the speaker/ interpreter's I-language is applied to the verbal input to generate mental representations exhibiting the phonological, syntactic, and/or semantic properties of the sentence uttered (2000, pp. 70-73, 117, 120).[2] Given that one has internalized a finite set of rules, and the meanings of a finite set of morphemes to which the rules can be applied, one will be able to compute the meanings of arbitrarily complex constructions by recursion. Difficulties or mistakes in the

[1] To make vivid the contrast here, consider the difference between asking, on the one hand, what must one know or be able to do in order to be a good free-throw shooter, and, on the other hand, what sort of physiology underwrites such knowledge/abilities in actual free-throw shooters? The answer to the first question might be: one must be able to arc the ball just so under noisy conditions. This, of course, is not an answer to the second question.

[2] It is an open question (about which I will have nothing to say here) how the theorist's representation, or "grammar", of the rules governing one's language relate to the subject's I-language. For an informative, yet opinionated, discussion, see Michael Devitt (2006, pp. 45-84).

process can be chalked up to performance error (forgetfulness, exhaustion, chemical imbalances, etc.). I will refer back to this story as *the Chomskyan account of competence.* It can be filled out in a number of different ways; I will pick one in order to fix ideas. But what's distinctive about each way of filling out the Chomskyan story is that it represents competence with a language as a specialized cognitive capacity that operates independently of one's general intelligence and knowledge (Chomsky 1986, p. 48).

According to one way of filling out the Chomskyan theory of competence, the language faculty is a mental module in the sense defined by Jerry Fodor (1983). It is a domain specific, informationally encapsulated, and highly rapid computer whose purpose is to generate representations of linguistically significant properties.[3] Suppose a speaker utters a sentence, *s*. The interpreter's language module parses the utterance by performing a series of computations at various levels of processing: phonological, syntactic, semantic. The output at each level is a representation that exhibits the corresponding properties of *s*. Suppose (1) below represents *s*'s phrase structure.

(1) $[[\Delta]_{NP}[[\Delta]_{V}\ [[\Delta]_{Det}[\Delta]_{N}]_{NP}]_{VP}]_{S}$

Then at the syntactic level of processing, the interpreter's module generates a representation that exhibits *that* very structure. At the semantic level, those elements of syntactic form which are relevant to semantic interpretation are recovered and another representation is generated, this one exhibiting the semantic properties of *s*. Admittedly, this way of filling out the Chomskyan story relies on assumptions that not all Chomskyans share. My aim was to convey a better understanding of the spirit of the view by looking at how it might be spelled out. Nothing I go on to say depends on the particular way I have filled out the account of competence. This concludes our whirlwind summary of the Chomskyan response to (PQ).

What of (MQ)? The Chomskyan answer is that sentences inherit their phonological, syntactic, and semantic properties from the mental representations that the language module generates as output. (1) correctly represents the phrase structure of *s* in virtue of the fact that an interpreter's language module generates a mental representation at the syntactic level of processing whose structure exemplifies (1). Similarly, *s* means that

[3] A representation generator is more or less informationally encapsulated as it is more or less closed off from one's background knowledge or beliefs. The human visual system provides a good example of informational encapsulation. You may know that two lines are the same length, and yet one line may appear longer than the other because your visual system is encapsulated from the total stock of information you have access to.

$p$ because an interpreter's language module yields, at the semantic level of processing, a mental representation that means that $p$. More generally, for it to be the case that, in one's idiolect, expression $e$ has $v$ as its semantic value just is for one's language faculty to implement an I-language mapping $e$ onto a mental representation with $v$ as *its* semantic value (Stephen Laurence 1996, p. 284 and 1998, p. 212). *Thus, for the Chomskyan, what it is to be linguistically competent,* (PQ), *is theoretically prior to how content is fixed,* (MQ).

The second set of answers to the questions we began with is due largely to David Lewis (1969, 1983), but it also owes much to Paul Grice (1957, 1989). According to Lewis, for $s$ to mean that $p$ in community $C$ is for there to be a convention among members of $C$ to, (*a*), speaker-mean that $p$ by uttering $s$ only when they believe that $p$, and to, (*b*), believe that $p$ when a speaker sincerely and assertively utters $s$. The notion of speaker meaning presupposed here was originally introduced by Grice (*Ibid.*). Roughly, a speaker means that $p$ by uttering $s$ iff, first, she intends to produce the belief that $p$ in her audience by uttering $s$, second, she intends for her audience to recognize her intention to produce such a belief by uttering $s$, and, third, she intends to produce the belief that $p$ in her audience by getting the audience members to recognize her intention to produce such a belief in them. Following Lewis, let's call a convention to (*a*) "Truthfulness" and a convention to (*b*) "Trust".[4] I will refer back to this account as *the convention-based theory of meaning determination.*

A convention is an arbitrary regularity among members of a society that all or most members of the society respect due to a mutually known common interest in complete conformity (members of the society know that there is such an interest, and they know that others know, and so on). Lewis (1983, pp. 164-166) tries to capture this idea in his

[4]In order for an action to be conventional there has to be an element of choice involved in performing it. Conventions are, after all, voluntary; those who participate in them could have coordinated their behavior to satisfy their preferences in other equally good ways. But this observation puts pressure on Lewis's account, because it is widely thought that belief is not a voluntary cognitive reaction. Is it conventional to believe that there is beer in the fridge when one opens the door and sees that there is? Surely the answer is no. Coming to believe that there is beer in the fridge is not, in the relevant sense, an arbitrary response to the visual evidence. So why think that Trust—regularly believing that $p$ in response to a sincere, assertive utterance of $s$—is a convention? Coming to believe that $p$ in response to a sincere speaker's assertively uttering $s$ is not an arbitrary response to the testimonial evidence. I owe this point to Alex Byrne. Fortunately, the Lewisian account can be revised to get around this worry. For $s$ to mean that $p$ in $C$ is for there to be a convention of Truthfulness (speaker mean that $p$ by uttering $s$ only if you believe that $p$) and a *regularity* of Trust (preserved by a common interest in coordination) in $C$.

analysis of conventionality: a regularity, $R$, is a convention in $C$ iff

(2) everyone (or almost everyone) conforms to $R$ in $C$;

(3) everyone (or almost everyone) thinks that others in $C$ conform to $R$;

(4) the belief that others conform to $R$ gives everyone a good and decisive reason to conform to $R$;

(5) members of $C$ prefer general conformity to $R$ rather than slightly-less-than-general conformity;

(6) $R$ is not the only possible regularity meeting conditions (3) and (4), and other possible regularities would satisfy the purpose for which $R$ is useful about as well as $R$;

(7) conditions (2)-(6) are matters of common knowledge.

If we combine the analysis above with the convention-based theory of meaning determination, what we get is *the Lewisian theory of content*: $s$ means that $p$ in $C$ in virtue of the fact that

(2*) members of $C$ ($a$) speaker-mean that $p$ by uttering $s$ only if they believe that $p$ and ($b$) come to believe that $p$ when others sincerely, assertively utter $s$;

(3*) everyone (or almost everyone) thinks that others in $C$ ($a$) and ($b$);

(4*) the belief that others ($a$) and ($b$) gives members of $C$ a good and decisive reason that they ($a$) and ($b$);

(5*) everyone in $C$ prefers that others generally ($a$) and ($b$);

(6*) members of $C$ could just as well have speaker-meant some other proposition, that $q$, by uttering $s$ only if they believed that $q$, and they could just as well come to believe that $q$ when others sincerely, assertively utter $s$;

(7*) it's a matter of common knowledge in $C$ that (2*)-(6*) are satisfied.

I distinguish the convention-based theory from the Lewisian theory in order to emphasize that one can reasonably agree with the former and disagree with the latter on the

grounds that (2)-(6) unsuccessfully capture the idea of conventionality, or that (2)-(6) do not express the proper sense of conventionality for the purpose of addressing (MQ).[5]

The theory can be filled out even further by plugging in Grice's analysis of speaker meaning directly into (2*)-(7*). Doing so makes explicit the complex pattern of higher-order propositional attitudes that the Lewisian theory of content attributes to communities of language users. If the account is correct, then a wide range of language users have beliefs, intentions, and preferences directed at the beliefs, intentions, and preferences of other language users (Laurence 1996, p. 276). Exceptions can be tolerated, no doubt, but such individuals are not, as Lewis would put it, "party to" the convention; they do not participate in, by helping to preserve, the convention. We needn't go through the hassle of actually plugging in the analysis of speaker meaning to illustrate the point here, because even as it is now the Lewisian theory obviously attributes complex attitudes about the communicative beliefs and preferences of other language users. I draw your attention to this aspect of the theory because it will play a role later in our discussion. This concludes the Lewisian response to (MQ).

What of (PQ)? To be competent in the language that *C* speaks, according to Lewis (1969, p. 51), is to participate in, or be "party to", Truthfulness and Trust ('T&T' for short) in *C*. *x* participates in T&T in *C* only if *x* is one of the members of *C* about whom (2*)-(7*) is generally true.[6] Given that (2*)-(7*) attribute higher-order propositional attitudes to *x*, this account of competence predicts that competent language users have a 'theory of mind': that they possess intentional concepts without which they would not be able to think about what others are thinking about, and that they grasp commonsense folk psychological principles which describe the relationship between intentional states and action. Some authors find this prediction highly objectionable; we will look at why in Section 2.3. What I want to emphasize now, before we proceed further, is that *the Lewisian view has reversed the order of explanation: how content is fixed,* (MQ), *is theoretically prior to what it is to be linguistically competent,* (PQ). I am not aware of any discussion that juxtaposes the explanatory structure of the two sets of answers to our questions (Chomskyan, Lewisian), so I want to pause here to say a bit more about the significance of this point.

[5]Authors have identified problems with (2)-(6) as applied to meaning and language. See Tyler Burge (1975) for a nice discussion. The problem that Burge identifies can be fixed with a minor adjustment, as Burge himself acknowledges.

[6]"Generally" true, not always true, because most of us occasionally lie to and distrust others.

## 1.2 Three observations

### 1.2.1 The two sets of answers are importantly similar

Although the two views we looked at differ about where one should begin and end in answering the metasemantic and psycholinguistic questions, this difference should not mask an important similarity regarding their explanatory structure: each answers one question in terms of the way it antecedently addresses the other. The Chomskyan tells a story about competence and then refers back to elements of that story to provide a theory of content determination. The Lewisian gives an account of how content is fixed and then tells a story about competence in terms of that account. This is no accident. It's due to an often implicit assumption shared by advocates of each view. I will state the assumption now, but I will defend the claim that various authors take it for granted a little later.

Let $e$ be an arbitrary expression of natural language; let $P$ be a semantic property, for example, the property of expressing a truth iff Ahmadinejad is a fool; and let $M$ be a set of mental states, like the mental states involved in modular semantic processing, or in participation in T&T. The assumption lurking behind both the Chomskyan and the Lewisian view—the assumption in virtue of which advocates of each view answer one of our two questions in terms of the way they antecedently address the other—states:

(HOMOGENEITY) If $M$ determines that $P(e)$ in $C$, then $M$ is implicated in the cognitive process whereby competent interpreters in $C$ recognize that $P(e)$.

First, a word about nomenclature. I call the assumption the 'HOMOGENEITY' principle because it closes the door on a hybrid or heterogenous account of content and competence—the synthesis I advertised at the beginning of our discussion, if you recall. HOMOGENEITY says that the mental states involved in fixing the semantic properties of an expression are also involved in the process that allows the interpreter to recover the expression's meaning; no other mental states than those implicated in semantic interpretation determine that an expression has its semantic properties. A mental state, $m$, is 'implicated' in semantic interpretation if a complete and accurate psychological explanation of the process refers to $m$.

According to the Chomskyan account of competence, interpreters of English come to recognize, for example, that 'Ahmadinejad is a fool' means that Ahmadinejad is a

fool by generating a mental representation at a particular level of processing which means that Ahmadinejad is a fool. Given HOMOGENEITY, no other mental state determines the meaning of 'Ahmadinejad is a fool'. So, with HOMOGENEITY in the background, the Chomskyan account of competence implies the corresponding Chomskyan theory of content determination. In contrast, the Lewisian theory of content says that 'Ahmadinejad is a fool' means that Ahmadinejad is a fool because there is a pattern of convention-supporting attitudes prevalent among members of our community about the conditions in which 'Ahmadinejad is a fool' can be uttered and the proposition that Ahmadinejad is a fool can be accepted. With HOMOGENEITY in the background, it follows that such attitudes must be implicated in the process whereby competent interpreters come to recognize the semantic properties of 'Ahmadinejad is a fool'. Such attitudes must, therefore, be partly constitutive of the interpreter's competence with her language. So, granting HOMOGENEITY, the Lewisian theory of content implies the corresponding Lewisian account of competence. In Section 3 we will see how the HOMOGENEITY principle manifests itself (sometimes explicitly, sometimes implicitly) in the work of both Chomskyans and Lewisians. But even at this stage it should come as no surprise that, given HOMOGENEITY, the two dominant views in response to (MQ) and (PQ) exhibit the kind of explanatory structure that I have attributed to them. Admittedly, this by itself isn't conclusive evidence for thinking that HOMOGENEITY is actually at work, but it is suggestive.

### 1.2.2 The two sets of answers are importantly different

Whereas the Chomskyan view offers us a theory of idiolectic meaning determination, the Lewisian view provides a theory of how communal linguistic meaning is fixed. This difference reflects a deeper disagreement about the role language is supposed to play in the study of content, competence, and communication.

Which is theoretically fundamental, one might ask, the community's language or the speaker/interpreter's idiolect? Lewis and Grice say very little to directly address this question, so the motivation for the language-first outlook will have to come from elsewhere. According to Michael Dummett, "one cannot so much as explain what an idiolect is without invoking the notion of a language considered as a social phenomenon" (1978, p. 425). An idiolect, says Dummett, is a speaker's more or less incomplete/inaccurate representation of her public language. Idiolects must therefore be answerable to public languages, just as any representation is answerable to what it

represents. But what exactly is it for an idiolect to be 'answerable to' a public language? When a speaker's grammatical or semantic beliefs deviate from the rules or facts that constitute her community's language, and she is made aware of this, the speaker must correct the deviation. Otherwise, Dummett says, communication would be impossible. "An English speaker both holds himself responsible to, and exploits the existence of, means of determining the application of terms which are either generally agreed among the speakers of English ... this has to be so if words are to be used for communication ..." (p. 424). "In employing words of the English language, we have to be held responsible to their socially accepted use, on pain of failing to communicate ..." (p. 429). For Dummett, reference to public language is thus indispensable to a satisfactory account of how communication is possible. If he's right, then insofar as one is interested in communication, public language is the theoretically fundamental notion.

Of course, one might criticize Dummett's argument by flat-footedly insisting that he presents things in the wrong order: idiolects needn't be defined in terms of public languages; public languages have to be defined in terms of idiolects. A public language, one might think, is just a bunch of overlapping idiolects backed by an army and a navy. Dummett is obviously aware of this possible reply and dismisses it (*Ibid.*), but I'm not entirely clear about why. His discussion is extremely obscure. Since I don't want to get bogged down by Dummett exegesis, I will suggest my own way of bridging the gap in his argument.

Suppose, along with our hypothetical critic, that the relationship between a community's public language and the idiolects of its members is like the relationship between the average height of a population and the heights of individuals within it. Then just as a population bottleneck would drastically shift the average height of a community (imagine a catastrophe that killed almost every adult below five feet, seven inches), so would it change the facts that constitute the population's language. Imagine an event that killed every chemical and botanical expert; only ignorant English speakers who incorrectly think that 'gold' and 'elm' are roughly synonymous with 'soft yellow metal' and 'common deciduous tree' survived. Suppose further that the survivors of this mass killing aren't even disposed to correct their semantic beliefs about 'gold' and 'elm' in the light of expert opinion. Then, after the catastrophe, the semantic facts would drastically change; 'gold' and 'elm' would be synonymous with 'soft yellow metal' and 'common deciduous tree'. The idiolect-first definition of public language thus implies that the once incorrect semantic beliefs of the survivors would now be correct. But in-

tuitively the survivors would still be wrong in thinking that 'gold' applies to samples of iron pyrite and 'elm' to beech trees. After all, it would be silly to think that the sudden death of every expert about a given subject-matter could transform ignorance (in this case, about the meanings of 'gold' and 'elm') into knowledge! The upshot of this is that statistical abstractions like average height and idiolect-first public language are unstable. The latter is too unstable to be part of a theory that accommodates intuitive judgments about the correctness of one's semantic beliefs. That is why the notion of a shared language can't be understood in terms of idiolects. Or so one might think. To repeat: I'm offering this argument on someone else's behalf.

I'm not sure how effective our little thought experiment is in covering the hole left open by Dummett's argument. More could be said in defense of the idiolect-first standpoint, but I don't want to get sidetracked. Whether successful or not, the thought experiment points to an important phenomenon that deserves to be accommodated, namely, semantic normativity. I will have more to say about this shortly. Keep it in mind.

The flat-footed objection we considered a moment ago isn't very illuminating. There are more interesting challenges to Dummett's argument, specifically, that it relies on an unrealistic conception of communication. Some German speakers have an easier time communicating with Dutch speakers than they do other German speakers, because some dialects of German resemble Dutch more closely than they do other dialects of German (Chomsky 1980, p. 118). The same is true of regional Scandinavian languages (Alexander George 1990, p. 283). This strongly suggests that conversational partners needn't have a shared language in order to communicate; they need only be able to recognize contextual clues, decipher general themes, and reason properly about the speaker's beliefs, goals, and intentions.[7] This, I take it, was Donald Davidson's point in 'A Nice Derangement of Epitaphs', where he observed that interpreters often adapt to malapropisms without serious damage to the communicative enterprise. If conversation normally takes place under such 'noisy' conditions, and speakers/interpreters are able to carry on just fine, then it mustn't be in virtue of a communally shared conception of what the words they use mean, because there isn't any such thing. Or, if there is a shared conception of what words mean, then it isn't as expansive as it would

[7]Chomsky and George draw a different conclusion from the observation about dialects. They take it to show that the individuation conditions of ordinary languages are unlikely to track anything of theoretical importance for the study of language competence.

have to be to explain the fact that we successfully communicate in such conditions.

My own view regarding the question at stake here is that reference to public language *is* dispensable to accounting for successful communication. But it would take us too far off track to fully defend the position here.[8] I drew your attention to this disagreement not because I want to conclusively resolve it in favor of my preferred view, but because the point of contention—whether one ought to theorize in terms of idiolects or shared languages—will be relevant in a moment when we tally some of the pros and cons of the two sets of answers.

### 1.2.3 You should buy a hybrid

I began our discussion by saying that there is something attractive and something unattractive about both of the views we have been considering, and that the respects in which each is individually unattractive recommends part of the other view. Now that the views have been presented, and some of their commitments have been made explicit, we're in a better position to tally the pros and cons.

One attractive aspect of the Chomskyan view is that it provides a uniform account of how all or many linguistically significant properties are determined (Laurance 1996, p. 285). One wants to know not only how semantic properties are fixed, but how phonological and syntactic properties are as well, and the Chomskyan provides a fully general stroy: *s* means that *p*, exhibits the phrase structure represented by (1), and has such and such a phonological profile in one's idiolect iff one's language module generates as output at each corresponding level of processing a mental representation that means that *p*, exhibits the structure represented by (1), and represents such and such a phonological profile. This accords nicely with the dominant view in generative linguistics, which is that its subject-matter is part of psychology. Some philosophers take a different view about the significance of semantic theories.[9] It may be, they grant, that phonology and syntax characterize aspects of a language user's psychology, but semantics does not. At any rate, sentences have phonological and syntactic properties, and surely something about language users determines the phonological and syntactic properties they have. Additionally, there is no reason to think in advance that content determination should be explained in separate terms, so it's to the Chomskyan's credit that she provides us with a uniform story.

[8]For further discussion see Richard Heck (2006).

[9]See, for example, Scott Soames (1984). Laurence replies to Soames in the appendix of his (2003).

What's nice about the Lewisian theory of content is that it captures the intuitive view of meaning as a conventional and, hence, normative phenomenon. The normative dimension of meaning stems from (3*) and (4*). But Stephen Laurence (1996, pp. 284-285) believes that the Chomskyan view is equally well equipped to accommodate the fact that meaning is conventional. Recall the intuitive description of conventions we gave in Section 1, just before I summarized the Lewisian analysis. Conventions are arbitrary regularities preserved by common interests. Now Laurence correctly observes that under different developmental circumstances, one's semantic processor would map utterances of $s$ onto a mental representation that means that $q$. So the fact that it actually generates a representation that means that $p$ is arbitrary. Laurence observes further that no theory-specific considerations are needed to account for why our semantic processors produce the outputs that they actually do: it's in our common interest to coordinate language use, and communication requires that speakers/interpreters associate roughly the same meanings with the same public signs, which in turn requires that our semantic processors map sentences onto mental representations with roughly the same contents. So given our intuitive understanding of conventionality, the Chomskyan seems to fare just as well as the Lewisian in respecting our conviction that meaning is conventional.

A hallmark of conventions is that they give rise to norms. So to say that meaning is conventional is to say in part that meaning is normative. But what exactly does this phrase—'meaning is normative'—amount to? Rather than define the phrase, let me illustrate with an example the phenomenon I want to draw your attention to. You will probably recognize the case very quickly. It was introduced by Tyler Burge (1979).

Bert is sitting in his doctor's office. His thigh has been bothering him for some time and he wants to get a prescription for fast-acting pain relief medicine. His doctor walks in and says, 'Ok, Bert, tell me about your problem.' Bert then replies, 'Well, doc, I have arthritis in my thigh'. Intuitively, the proposition semantically expressed by the sentence Bert uttered, 'I have arthritis in my thigh', is false. 'Arthritis' applies to an inflammatory condition in one's joints, not a pain in one's muscles. Thus we naturally judge that it's incorrect to use 'arthritis' in the way that Bert did. Bert was wrong partly *because* of what 'arthritis' means. To think that meaning is normative is to think (at least, as I use the phrase 'meaning is normative') that part of the explanation of Bert's error consists in the fact that 'arthritis' applies to an inflammatory condition in one's joints. It's to think that semantic facts play a role in accounting for mistakes of the sort Bert

made. Perhaps some authors mean something in addition to this when they use the phrase 'meaning is normative', but the argument to come relies on nothing more.

So where are we now? Well, we have the intuitive view that meaning is conventional, and we have the observation that a hallmark of conventions is that they give rise to norms. So one is naturally led to think that meaning is normative: that the explanation of mistakes of the sort Bert made partly consists in the semantic facts. Now, for the Chomskyan, the semantic facts about the words in one's lexicon are determined by more basic facts about one's semantic processor, in particular, by the representation that it generates as the output of its processing. So, for the Chomskyan, the explanation of Bert's mistake should consist (partly, anyway) in the fact that his semantic processor generates a representation with a specific meaning. But the representation that Bert's semantic processor generates as output when it's fed 'arthritis' as input is a representation that is synonymous with 'painful condition of the joints or muscles'. So 'arthritis' and 'painful condition of the joints or muscles' *are* synonymous *in Bert's idiolect*. One can't help but wonder, then: what, for the Chomskyan, is the semantic fact that Bert is flouting?

The fundamental problem with the Chomskyan theory of content—why it can't accommodate the normativity and, therefore, the conventionality of meaning—is that it's unable to distinguish between the semantic facts about a speaker's language and her representation of the semantic facts. It is unable to do so because it provides a theory of content determination that ties the meaning of an expression in the speaker's idiolect to its meaning *as represented by her semantic processor*. But there has to be a gap between representation and fact for error to creep in. The Lewisian account allows for there to be such a gap by acknowledging the community's contribution to fixing the semantic facts. The individual speaker's representation of the semantic facts can certainly deviate from her community's (or a privileged subset of her community's), and errors of the sort Bert made consist in such deviation. In light of all this, it is not surprising that Chomsky himself disavows the normativity of meaning. "These concepts of 'misuse of language' ... may be of interest for the study of the sociology of group identification, authority structure, and the like, but they have little bearing on the study of language" (2000, p. 71). "Is there any other concept of 'misuse of language'? I am aware of none. If so, the concept plays no important role in the study of language, meaning, communication, or whatever" (*Ibid.*). "... reference to 'misuse of language', to 'norms', to 'communities', and so on seems to me ... obscure, and it is not clear that

they are of any use for inquiry into language and human behavior" (p. 72). "Rules of language—for example, the principles of [Universal Grammar], or those that guide Mary's [semantic] judgments ...—are not normative ..." (p. 98). Since it's a hallmark of conventions that they give rise to norms, the Chomskyan theory of content is unable, *pace* Laurence, to accommodate the intuitive judgment that meaning is conventional.

To be fair, there is a kind of misuse or error that Chomsky acknowledges as clear and important for inquiry into language use. I quickly referred to it in Section 1 when presenting the Chomskyan account of competence. Performance errors, if you recall, are mistakes due to limitations of memory, computational capacity, and peripheral system interference and malfunction (e.g. auditory/visual difficulty), etc. Perhaps you had a bad burrito for lunch, and it's now affecting your concentration. So you slip up and accidentally use 'disinterested' when you should have used 'uninterested'. But none of this really helps to address the challenge above, because we can simply assume that Bert didn't have a bad burrito, that his memory, computational skills, and peripheral cognitive systems are working just fine.

Although Chomsky himself disavows the intuitive judgment that Bert misused 'arthritis', someone sympathetic to the Chomskyan outlook may be worried by the charge that her theory is unable to accommodate it. She might reply to my argument as follows.[10] The source of our shared judgment that Bert misused 'arthritis' is not the fact that 'arthritis' conventionally expresses a particular meaning, but that Bert is involved in an exchange of information with his doctor and he wants thereby to communicate and recover truths, as any typical speaker/interpreter would. Bert is unlikely to succeed, however, unless the way he uses 'arthritis' coincides with the way his doctor uses it. So given that his use diverges from his doctor's, we naturally judge that Bert has made a mistake: he falls short of his goal of expressing a truth about the condition of his thigh.

But unless the doctor is a completely uncooperative interpreter, which is unlikely, it is simply false that Bert falls short of his goal, because despite incorrectly asserting that he suffers from arthritis in his thigh, Bert correctly asserts that there is something wrong with his thigh, and his doctor comes to believe that there is. Given the context, that seems to be all that matters. Bert's mistake does not in any way thwart his communicative goals, as is illustrated by the fact that the doctor may well choose to ignore Bert's error and simply accommodate, for the purpose of this conversation, that 'arthri-

[10]Ian Rumfitt pushed this response in conversation.

tis' applies to problems of the joints or muscles. So one cannot do justice to the relevant datum by appealing to Bert's objectives qua communicator. His use of 'arthritis' does not frustrate (at any rate, it need not frustrate) his objectives.

Things have been rather one-sided so far, but the Lewisian's view is also flawed. Recall our discussion of competence as she understands it. To be competent with the language of community $C$ one must be a member of $C$ about whom (2*)-(7*) is generally true; it is, as I put it, to participate in T&T in $C$. One of the commitments of this account, as we noted in Section 1, was that a competent language user possesses a theory of mind, that she can represent the mental states of other members of her community in order to coordinate her speech behavior with theirs. So one should naturally expect that as the ability to coordinate one's speech in this way varies, so too does one's competence with the language. But there is compelling empirical evidence that falsifies this commitment. Language skills do not correlate with one's capacity to represent, reason about, and adapt one's behavior to the mental lives of others. There are severely intellectually challenged language users who lack a theory of mind—who are, as it were, almost entirely "mind blind"—but who possess nearly normal language skills. Studies of autism strongly suggest that the deficit involves the absence of a theory of mind. Now typically, subjects with autism have low IQs and poor language skills, but some so-called "high functioning" individuals with autism have normal IQs and close to normal language skills (Laurence 1996, p. 289). Additionally, there are highly intelligent subjects who can communicate by means of empathetic reasoning, but who cannot reasonably be credited with language competence (Chomsky 1980, p. 57). A frequently cited example of such a case is a woman named 'Genie', who grew up in isolation until the age of 13. Researchers describe her as a powerful communicator, but observe that her knowledge of English is equivalent to the average two year old's (Laurence 1996, p. 288). Other cases have been discussed in the literature, but we needn't run through them; the point is clear enough: competence with a language is largely independent of folk psychological knowledge. And that is exactly what one would expect if it were a special-purpose component of one's mental life, as the Chomskyan account of competence implies.

Something interesting emerges from our survey of pros and cons: the respect in which the Chomskyan theory of content determination is flawed recommends the Lewisian theory, and the respect in which the Lewisian account of competence is flawed recommends the Chomskyan account. What bars us, then, from accepting a hybrid of the

two views? Why not, that is, combine a Chomskyan story about competence, which accommodates the independence of one's language skills from general intelligence, with a Lewisian theory of how content is fixed, which respects the normative character of meaning? According to this view, content is fixed by conventions governing use, and competence is a complex, special-purpose component of one's psychology that endows the speaker with knowledge how to (non-accidentally) exploit content-fixing conventions without participating in them. What bars us from doing so is the implicit assumption I drew your attention to in Section 2.1:

(HOMOGENEITY) If $M$ determines that $P(e)$ in $C$, then $M$ is implicated in the cognitive process whereby competent interpreters in $C$ recognize that $P(e)$.

Less abstractly, the HOMOGENEITY principle says that the mental states involved in fixing the semantic properties of an expression are also involved in the process that allows the interpreter to recover the expression's meaning. No other mental states than those implicated in semantic interpretation determine that an expression has its semantic properties. In the discussion that follows, we will see HOMOGENEITY at work. I will then explain why it ought to be rejected and the hybrid theory I proposed accepted.[11]

[11] In his recent book, Devitt (2006, ch. 10) challenges the common view that data such as we have considered above support the Chomskyan account of competence. He then proceeds to criticize the hypothesis that there is a special-purpose language faculty. His argument can be summarized very quickly as follows. If there is a specialized language faculty, then it must be modality-neutral, that is, it must underwrite comprehension and production regardless of whether the speech is vocalized, signed, written, or read. But it appears that the language faculty is largely modality-specific. So, Devitt concludes, there probably is no language faculty. This argument deserves careful consideration, which cannot be carried out in a footnote. Suffice to say, I am skeptical of the first premise. I do not doubt that the language faculty must subserve speech whether vocalized or signed; I doubt whether it must underwrite written and read speech. After all, reading and writing are learned skills, and the language faculty was originally posited as an aspect of the mind in order to explain the acquisition of (what Chomskyans take to be) unlearned competencies. In any event, I am prepared to qualify the major conclusion I want to draw: *even if* the data regarding subjects with autism and Genie support the special-purpose nature of language competence, and there is a language module, that is no challenge to the Lewisian theory of content determination. *My point is that a theory of how content is fixed needn't carry any significant empirical commitments regarding the nature of language processing*. The reason why authors think otherwise is that they presuppose HOMOGENEITY.

## 1.3 HOMOGENEITY is false

The Lewisian's account of competence is probably false. But a couple of authors conclude, on the basis of this, that her theory of content determination is false, too (Dorit Bar-On 1995, and Laurence 1996 and 1998). What justifies this conclusion? Here is Laurence: "...if we do not have reason to believe that the basis of [the Lewisian's theory of content] is satisfied by considering the states required for processing language, what reason do we have for believing that this basis is satisfied?" (1996, p. 296) Switching from the interrogative mood to the indicative, Laurence's point can be paraphrased as follows. 'The only way to verify whether language users actually have the pattern of mental states that the Lewisian's theory of content attributes to them (and, hence, whether "the basis of the account is satisfied") is to look at the mental states involved in language processing. If such states are not implicated there, then it is simply false that they play a content-fixing role.' This paraphrase makes explicit Laurence's commitment to HOMOGENEITY.

Once more, Laurence: "I think there exists substantial evidence for the special purpose nature of the language processor, *and the connection between natural language utterances and their semantic properties is likely to reflect the distinctive nature of this processor*" (Laurence 1998, p. 201, emphasis mine). Now admittedly, it's no altogether clear what it comes to for the language processor to "reflect" the relationship between utterances in natural language and their semantic properties. The statement is highly metaphorical, and Laurence doesn't bother to clarify what precisely he means. But it's very likely that HOMOGENEITY conveys the literal meaning of Laurence's statement. Assuming this is correct, as I think it is, the larger point Laurence is trying to communicate, then, is that the special-purpose nature of the language processor is what motivates HOMOGENEITY. And it is because of his commitment to HOMOGENEITY that Laurence can reasonably conclude that the Lewisian theory of content itself is false on the grounds that the mental states it attributes to language users are not necessary for (implicated in) language processing.

Additionally, in light of considerations that motivate a broadly Chomskyan outlook on things, Dorit Bar-On suggests that we think of the convention-based theory of content as a "rational reconstruction of the condition under which language could emerge". What exactly does this amount to? Bar-On tells us: it is an illuminating but false story about how language might have but actually did not come to be, compa-

rable to, say, a social contract theory of government and political obligation (1995, p. 114). Grice himself understood the view in similar terms; he called it a "myth" that was designed to represent the "conceptual links" between semantic content and the intentional activities of language users. "But how can such a link be explained by a myth? This question is perhaps paralleled ...by the question how the nature and validity of political obligation ...can be explained by a mythical social contract" (Grice 1989, pp. 296-297). Why would Bar-On, a self-professed advocate of the convention-based theory of content determination, concede so quickly that it's literally false on the basis of considerations stemming for the cognitive science of language processing? The answer, I suggest, is that Bar-On implicitly accepts HOMOGENEITY. Only with it in the background is the conclusion she jumps to licensed.

But HOMOGENEITY is false; it needn't therefore bar us from accepting a heterogenous account. My objection begins with a true story.

Howard Engel is a Canadian author who suffered a mild stroke and lost the ability to read. As Engel reports, English letters "looked like Cyrillic one moment and Korean the next." But, oddly enough, Engel was still able to write. "The act of writing seemed quite natural to him, effortless and automatic, like walking or talking. The nurse had no difficulty reading what he had written, but he himself could not .... To his eyes, it was the same indecipherable 'Serbo-Croatian' ..." (Oliver Sacks 2010, pp. 22 & 24).[12]

The ability to read is a complex skill; it consists in several different component skills: the ability to recognize letters and words and thereby associate them with sounds, the ability to recover meanings from phrases and sentences; the ability to keep track of and 'synthesize' recovered meanings in order to form a cohesive interpretation of the whole message. Specialists call the recognitional and associative capacities "reading accuracy"; they label the recovery, tracking and synthesizing capacities "reading comprehension" (Nation *et al.* 2006, pp. 911-912). I will use this terminology.

Accuracy and comprehension in reading are not merely conceptually distinct; they sometimes come apart in actual readers. Here too examples can be found in studies of autism. Children and adults with autism sometimes have above-average recognitional and associative skills, exhibiting reading accuracy well in excess of IQ-expected levels. But even in such cases, understanding is quite low; the reader is typically unable to comprehend the meaning of the text beyond a few isolated words and phrases. Subjects in this condition are called "hyperlexic" (Colasent and Griffith 1998, p. 414).

[12] I want to thank Richard Holton for suggesting that I look at the Sacks article.

As it turns out, reading accuracy involves exercising the object-recognition part of our brains, which, like the language faculty, seems to be a largely special-purpose (maybe even modular) device. Engel's stroke damaged this part of his brain. Along with severely diminished reading accuracy, "he had a large blind spot in the upper-right quadrant of his visual field, and he had difficulties recognizing colors, faces, and everyday objects." But by far the most radical effect was the damage to his reading abilities. Fortunately for Engel, the region of his brain that embodies his recognitional and associative reading abilities is entirely separate from the region of his brain that subserves his writing abilities. That's why Engel's ability to write was unaffected (Sacks 2010, pp. 22-28).

Engel's story isn't unique. The first documented case of reading impairment such as his occurred in 1887, when a Frenchman named 'Mr. C' experienced "pure verbal blindness" (the selective loss of the ability to associate a letter with its characteristic sound) but was still able to write. However, while Engel's verbal blindness was accompanied by more widespread difficulties with object recognition, Mr. C's condition was almost entirely restricted to reading-specific skills. The late 19$^{th}$ century French neurologist, Joseph-Jules Déjerine observed that

> When shown objects, he names them easily. He can name the parts of all the instruments in an industrial design catalogue. At no point during this examination is his memory at fault; drawings immediately prompt the appropriate word and how to use each object .... When handed the newspaper *Le Matin*, which he often reads, the patient says: "It's *Le Matin*, I recognize it by its shape," but he cannot read a single letter in the title (quoted in Stanislas Dehaene 2009, p. 55).

The dissociation in Mr. C between reading accuracy and object recognition more generally "implied the presence in the brain of a cortical 'visual center for letters' specialized in reading" (*Ibid.*). Subsequent research with the assistance of positron emission tomography and fMRI scans revealed that this special-purpose region of the brain is the left occipito-temporal area, which is located at the rear base of the left hemisphere (Dehaene 2009, p. 61-69).

Engel and Mr. C, coupled with recent studies of autism, provide compelling evidence that reading accuracy is neurologically localized and independent of general intelligence; two hallmarks of special-purpose cognition. One might be so impressed

by such data that one may be led to posit a 'reading accuracy processor'. In fact, some researchers have. "We have discovered that the literate brain contains specialized cortical mechanisms that are exquisitely attuned to the recognition of written words. Even more surprisingly, the same mechanisms, in all humans, are systematically housed in identical brain regions, as though there were a cerebral organ for reading" (Dehaene 2009, p. 4). And—just to bring out the analogy with Laurence's views even further—one might then be compelled to think that the connection between letters of the alphabet and the way they should be pronounced when reading is likely to "reflect the distinctive nature of this processor" (Laurence, *Ibid.*).[13] Consequently, one might be attracted to a thesis very like HOMOGENEITY. Namely,

(ANALOGUE) If $M$ determines that, for example, 'a' is to be pronounced $/a/$, then $M$ is implicated in the process whereby readers recognize the association between a token letter 'a' and $/a/$.

I want to emphasize that the motivation for ANALOGUE parallels Laurence's motivation for HOMOGENEITY. In both cases, acceptance of the principle is supported by the special-purpose nature of the relevant capacity. It's difficult to see any principled reason for accepting HOMOGENEITY and rejecting ANALOGUE. The two principles stand or fall together.

Now it seems to me that the association between a letter and the way it should be pronounced is a paradigm of conventionality. If anything is conventional, the relationship between a letter and its characteristic sound is. In fact, this relationship appears to be constituted by the sort of convention that Lewis set out to analyze with (2)-(7). (Call it an 'L-convention'.) For suppose it was not common knowledge, nor was it mutually believed, that community members by and large think that 'a' is pronounced $/a/$, that 'b' is pronounced $/b/$, and so on. It would then be doubtful whether the sequence of letters that a community member uses to inscribe a message would be pronounced in the way that she herself pronounces them. So community members would be unsure as to whether the message they want to convey with their inscriptions would be successfully conveyed. And it's difficult to see, then, how there could be a regular, community-wide, self-perpetuating practice of pronouncing 'a' $/a/$, 'b' $/b/$, etc., because the practice would no longer serve the purpose for which it exists. So it seems that a theory very much in the spirit of the Lewisian analysis of convention accurately

[13]Compare this view with the italicized quotation from Laurence (1998, p. 201) two pages back.

represents the constitutive basis for the association between a letter of the alphabet and its pronunciation. In particular, the theory will attribute higher-order propositional attitudes (common knowledge, or mutual belief) directed toward the beliefs that other community members have about the way letters are pronounced. One then wonders whether such attitudes are implicated in the process whereby readers recognize that a particular letter is to be pronounced in a particular way. The answer, contrary to ANALOGUE, is surely that they aren't, because hyperlexic readers are capable of recognizing the association between 'a' and /*a*/, 'b' and /*b*/, etc. without being capable of thinking about what others are thinking about. The reading accuracy processor doesn't make use of convention-constituting higher-order propositional attitudes. So ANALOGUE is false; *a phenomenon can be conventional even when it turns out that it's sometimes exploited to achieve specific ends by means other than participation in the convention.*

HOMOGENEITY is false for the same reason. The empirical data regarding the special-purpose nature of the language faculty does not jeopardize the view that content determination is (L-)conventional, just as it does not jeopardize the view that letter-sound association is (L-)conventional. Rather, it strongly suggests that the capacity to recognize a natural language utterance as having particular semantic properties is a capacity to exploit T&T without participating in it. This, I propose, is the way we should think about the relationship between how content is determined and what competence consists in.

CHAPTER 2

# Speaker Meaning in Context

I want to do three things in this chapter. First, I want to reconstruct a puzzle about meaning and communication. The puzzle is Ray Buchanan's (2010) but, as he acknowledges, it owes much to Stephen Schiffer's (1995) work on the semantics of definite descriptions. I will focus on Buchanan's presentation, but I will also say something about Schiffer's take on definites.

Second, I want to solve the puzzle. Whereas Buchanan thinks it "calls into question the most basic assumptions of the standard view" of communication by means of language (p. 341), according to which propositions are fit to be the contents of both thought and speech,[1] I think the puzzle shows that advocates of the standard view are committed to the doctrine of *content localism*: that what a speaker means by assertively uttering a sentence is determined only in relation to a set of contextually relevant possibilities, as opposed to a completely unrestricted domain of possibilities. In other words, speech acts never partition all of logical space, only a highly circumscribed sub-region. Because the standard view is so attractive (in fact, even Buchanan admits the strong temptation "to think that some version of the standard view is obviously true" (p. 340)) I think we end up with a rather compelling argument in favor of localism about content. Additional advantages of localism will be discussed as I proceed.

Finally, I want to bring out some similarities between Buchanan's puzzle and a worry one might have about the doctrine of pervasive meaning indeterminacy commonly associated with W. V. Quine and Donald Davidson. The worry I have in mind was expressed clearly and forcefully in an old paper by John Searle (1987): in any con-

[1] This is one of the core assumptions in the Grice-Lewis theory of content determination, which I defended in the previous chapter.

text where one assertively utters 'Lo, a rabbit!' one surely knows about one's own case what was meant by the utterance. This appears to imply, however, that there is a determinate semantic fact about which one can have first-personal knowledge. Once we recognize the similarities between our puzzle and this worry, then we can appeal to localism to resolve the apparent inconsistency between the doctrine of meaning indeterminacy and Searle's observation. This is fortunate because some form of moderate yet far-reaching meaning indeterminacy may well be inescapable.

## 2.1 Preliminaries

The puzzle I want to look at targets "the standard view" of communication. What is the standard view, anyway? Buchanan offers a nice, compact description.

> ... a speaker has a thought—say, a belief that Fichte was a philosopher—which she would like to convey to her audience. This thought has a certain proposition as its content, a proposition which we might specify using the 'that'-clause 'that Fichte was a philosopher'. If the speaker knows that her audience is a competent speaker of a shared language such as English, she can choose some form of words which makes manifest to her audience the proposition she intends to communicate. Perhaps she utters 'Uncle Johann was a philosopher'.... The speaker's audience recognizes her communicative intentions, and communication is successful, only if her audience thereby entertains a thought whose content is the proposition the speaker meant by her utterance (p. 340).

Embedded within this description of the standard view are two assumptions. Following Buchanan, I will call them 'CONTENT' and 'SUCCESS'.

(CONTENT) What a speaker means, or intends to communicate, (at least in cases of indicative speech) must be a proposition.

(SUCCESS) Understanding a speaker's utterance U requires (minimally) entertaining what she meant by U.

The view is neutral as between fine-grained and coarse-grained, structured and unstructured, theories of propositions. To generate the puzzle one merely has to think of

a proposition as the kind of thing that is both the content of one's mental states and the content of sentences one utters to give voice to such states. Typically, however, propositions are also thought of as the kind of thing that determines (either by being or in some other way encoding) a set of truth-conditions. According to Richard Heck (2002), this minimal conception of propositions can be traced as far back as Frege himself.

Advocates of the standard view ("standard theorists") can disagree about the nature of speaker meaning, just as they can disagree about the nature of propositions. But Buchanan (p. 344) offers a pretty good approximation of what it comes to.

(M$^{**}$) A speaker means the proposition P by uttering U only if, for some audience A, and feature $\phi$, she produces U intending (i) that A is to entertain P, (ii) it is mutually obvious between her and A that $\phi$(U), and (iii) A come to recognize her intention (i), at least in part, on the basis of her recognition that $\phi$(U).

Suppose you are channel surfing one afternoon. You come across Fox News. Glenn Beck is at his chalkboard. I point and assertively utter 'He is a real genius'. Although the sentence I uttered conventionally means that Glenn Beck is a real genius, what I mean by uttering the sentence is that Glenn Beck is an idiot. After all, I uttered the sentence intending (i) that you entertain the proposition that Glenn Beck is an idiot, (ii) it is mutually obvious between us that my utterance was sarcastic, and (iii) you come to recognize my intention (i), at least in part, on the basis of your recognition that my utterance was sarcastic. According to CONTENT and the minimal conception of propositions I introduced a moment ago, 'that Glenn Beck is an idiot' refers to an object that determines truth-conditions. According to SUCCESS, you understand my speech act only if you grasp that object. Now what could possibly call this account into question? It feels more like a series of truisms about the conversation rather than a falsifiable theory.

A consequence of SUCCESS and CONTENT is

(LEMMA) If a speaker means a proposition P by her utterance U then the listener must entertain P if she is to understand U.

Buchanan alleges that examples involving implicit quantifier domain restriction and non-sentential assertion falsify LEMMA. Let's look at each sort of case in turn.

We begin with a story. Tim and Chet are hosting a house-warming party. Chet convinces Tim that their more sophisticated guests will want foreign beer.

> To cater to the sophisticates that they hope will show up later that night, they decide to go to a local corner store to pick up several cases of imported bottled beer which they will serve from a giant ice-filled plastic bucket, decorated in a pirate motif, which is to be located in their back yard.
>
> An hour before the party is to begin, Tim asks Chet 'Are we ready to rage?'... Chet responds, 'We are totally ready. The living room totally looks like a pirate ship. The strobe lights are up. Every beer is in the bucket. I just need to find an eye patch to wear with this pirate hat.' Consider (5):
>
> (5) Every beer is in the bucket.
>
> Most people, even philosophers and linguists for that matter, agree on at least two things regarding this case. First, in uttering (5) Chet could have said something true despite the fact that (a) there are numerous bottles of beer nowhere near Chet and Tim's apartment, and (b) there is more than one plastic bucket in the world. Secondly it should be agreed that, it is possible, and in this case probable, that Tim recognized Chet's communicative intentions in uttering (5).
>
> Now the standard theorist might allow that a speaker might mean or say many propositions by uttering what she does. If Lemma is correct, any such proposition the speaker means or says by uttering U will be such that the speaker's audience must entertain it in order to understand the utterance. The problem is that in a case such as (5) there simply is no proposition that has the property that Lemma requires (p. 347, 349).

The problem, according to Buchanan, is that there are many equally good yet non-equivalent candidate propositions for what Chet meant by uttering (5).

(CP1) that every beer *we bought together at the bodega* is in the bucket *in the backyard*

(CP2) that every beer *we will serve at the party* is in the bucket *decorated in pirate motif*

(CP3) that every beer *for our guests* is in the bucket *filled with ice*

(CP4) that every beer *at the apartment* is in the bucket *next to the hot tub*

(CP5) that every beer *we bought together at the bodega* is in the bucket *next to the hot tub*

(CP6) that every beer *at the apartment* is in the bucket *in the backyard*

(There are probably many more candidate propositions, but whatever they happen to be they will differ in (global) truth-conditions.) Buchanan says that none of (CP1)-(CP6) satisfy LEMMA: for no ($CP_i$) is it the case that if Chet meant ($CP_i$), then Tim would have to entertain ($CP_i$) in order to understand Chet's utterance of (5). The reason is that Tim could understand Chet's utterance of (5) by entertaining ($CP_j$), which is distinct from ($CP_i$). Suppose Chet in fact meant (CP2). Well, Tim needn't entertain (CP2) in order to understand Chet's speech act, because entertaining (CP3) would suffice. And since (CP3) differs from (CP2) in truth-conditions, (CP3) is not the same proposition as (CP2). So LEMMA is false. Or so the argument goes, anyway.

The argument above challenges LEMMA. If it succeeds, then we know that at least one of either CONTENT or SUCCESS is false. But which is the culprit? Buchanan offers a more direct argument for the falsity of CONTENT. It begins with the observation that if a (competent) speaker means P by utterance U, then she has good reason to expect that her audience is likely to entertain P on the basis of U. But no candidate proposition is such that Chet has good reason to expect that Tim will entertain it on the basis of his utterance. "Even if the speaker uttering (5) in some sense 'has in mind' one of these propositions...she cannot mean it. She has no reason to think that her audience will recognize her as having meant just this proposition" (p. 350). So Chet didn't mean any of the candidate propositions. And if Chet meant a proposition, then he meant one of the candidate propositions. So there mustn't be a proposition that Chet means. As Buchanan puts it, "...any account that respects the generality and indifference characteristic of speaker's communicative intentions must give up Content..." (p. 357).

The argument from non-sentential assertion exemplifies a very similar structure.

> Suppose after graduating Chet and Tim get jobs working as short order chefs at a restaurant in their college town. While at work, Tim spots an oddly dressed man in the dining room curiously sniffing a plate of chicken fried steak that Chet had just prepared. Tim rushes over to Chet, taps him on his shoulder, nods in the direction of the man, and utters (8):

(8) A health inspector.

> In uttering (8) Tim presumably meant something, and Chet quite plausibly understood him, but what proposition could he have meant (p. 351)?

There are many equally good yet non-equivalent candidate propositions.

(CP7) that *he is* a health inspector

(CP8) that *the man we are looking at is* a health inspector

(CP9) that *the customer sniffing his plate is* a health inspector

(CP10) that *the guy who is frowning at his chicken fried steak is* a health inspector

(CP11) that *the guy with the strange mustache is* a health inspector

And just as in the first example, the listener (in this case, Chet) could have understood the speaker's utterance by entertaining any one of the candidate propositions, whether or not it was actually meant by the speaker.

So LEMMA appears to be false. And, as before, there is a direct argument that lays the blame on CONTENT: a speaker means P only if she has good reason to expect that her listener is likely to recover P on the basis of her speech act. But no candidate proposition is such that Tim has good reason to expect that Chet will entertain it on the basis of his utterance, because the conventional meaning of (8) provides no more evidence for any one of the candidate propositions to be identified as the thing that Tim meant than it does for any other proposition. Since Tim (a competent speaker of the language) knows this, he mustn't intend to communicate any of the candidate propositions. But if Tim meant a proposition at all, then surely he meant one of the candidates. So he mustn't have meant a proposition.

What the two cases have in common, as Buchanan notes, is that "the character of the lexical material uttered ... falls far short of constraining the options down to some one proposition, or set of propositions" (p. 351). It is likely that such constructions are the norm in everyday conversation, which often involves sentence fragments and incomplete phrases. So the argument does not rest on some anomaly that advocates of the standard view can just idealize away.

Perhaps the conventional meanings associated with the lexical items occurring in (5) and (8) do not pin down a unique proposition, but there is a lingering feeling that,

*given the contexts in which the speech acts take place*, all of the propositions that can reasonably be considered candidate speaker meanings amount to the same thing. The reason why Tim can understand Chet's utterance in the first case, regardless of which candidate proposition he in fact entertains, is that the differences between (CP1)-(CP6) are nullified by contextually salient information. Similarly, the reason why Chet can understand Tim's utterance in the second case, regardless of which candidate proposition he in fact entertains, is that the differences between (CP7)-(CP11) are eliminated by contextually salient information. If one had a context-relative view about speaker meaning that identified (CP1)-(CP6) and (CP7)-(CP11) in light of the common stock of information at each context, then the problem would be solved. Such a view would say that (CP1)-(CP6) and (CP7)-(CP11) are, in the relevant contexts, one and the same proposition meant, though it may well be that relative to some other context some or all of the candidate propositions are distinct. According to content localism, whether P and Q are the same proposition meant by the speaker is not a context-independent matter; one must fix on a particular context in order to answer questions about sameness/difference of proposition meant. In the next section I want to spell this view out more clearly in order to do justice to the lingering feeling I expressed a moment ago.

## 2.2 Speaker meaning localized

I want to help myself to a picture of content and conversation defended in Robert Stalnaker (1978, 1984). With this picture in view, we will be able to solve Buchanan's puzzle.[2]

The purpose of an assertion is to partition a space of possibilities. What possibilities? Well, the relevant ones, of course. But which are those? Proposal: they are the possibilities consistent with (left open by) what the speaker and her audience mutually presuppose for the purpose of the conversation. Call this space of possibilities the *context set*. Given that the purpose of an assertion is to eliminate possibilities from the context set, how might we represent the content of an assertion in such a way as to model successful and unsuccessful assertions? The simplest solution is to identify the

[2]The framework that I'm helping myself to is based on a number of assumptions that have been criticized by different authors. The most challenging criticisms, to my mind, are due to Hartry Field (1986), Stephen Schiffer (1986), Scott Soames (2006), and John Hawthorne and Ofra Magidor (2009). See Stalnaker (1986, 2006, 2009) for replies.

content of an assertion with a subset of the context set, specifically, the subset of possible worlds, S, in the context set at which the uttered sentence, given its conventional meaning at each world $w \in$ S, is assigned the value True in $w$. Stalnaker calls S the *diagonal proposition*.[3] So according to the possible worlds theory of conversation, the content of an assertion just is the diagonal proposition determined by the uttered sentence and the context set. This theory implies the following principle about asserted content.

(SAMENESS) If updating context C with either P or Q results in the same updated context C′, then, relative to C, P and Q are the same asserted content.

The solution I am about to suggest relies on this commitment of the Stalnaker framework.

Buchanan notes that there is a constitutive connection between asserted content and speaker meaning (p. 345). He captures this connection with the following principle: if a speaker asserts P by uttering U, then at least one of the propositions that she means by U must be P. Given this principle, if P and Q amount to the same asserted content in C, then 'they' are 'both' meant by the speaker in C. Consequently, if one can show that the various candidate propositions from the two puzzle cases are really the same asserted content (and, hence, the same proposition meant by the speaker), then it will follow that by entertaining one candidate speaker meaning the listener entertains all. So on the assumption that the speaker meant ($CP_i$), we will be able to say that the listener entertains ($CP_i$). Thus LEMMA will be satisfied and neither CONTENT nor SUCCESS will be in danger of falsification.

Let's focus on the first case first. Why think that (CP1)-(CP6) are, in the relevant context C, the same asserted content? More precisely, why think that updating C with any one of (CP1)-(CP6) will result in the same updated context C′? Well, because it is mutually presupposed by Chet and Tim that the following identities hold.

(BEER) the beer that we bought together at the bodega = the beer that we will serve at the party = the beer for our guests = the beer at the apartment

[3]It's easy to come away from Stalnaker's (1978) with the impression that the diagonal proposition is asserted only in abnormal cases of apparent conversational infelicity, but that would be a mistake. When no conversational norm appears to be violated, the diagonal proposition just is the conventional meaning, or "horizontal proposition", expressed by the sentence. In cases where apparent violation occurs, the two come apart and the audience identifies the asserted content of the speaker's utterance with the diagonal proposition. Either way, what gets asserted is the diagonal. That is why I omit any discussion of the horizontal above.

(BUCKET) the bucket in the backyard = the bucket decorated in pirate motif = the bucket filled with ice = the bucket next to the hot tub

So BEER and BUCKET are true at every world in the context set. So the quantifiers in (CP1)-(CP6) range over the very same set of objects at each world in the context set. Consequently, updating C with any one of the candidate propositions will result in C'. Given SAMENESS, it follows that (CP1)-(CP6) are really the same asserted content in C. And given the constitutive connection between asserted content and speaker meaning, it follows that (CP1)-(CP6) are the same candidate speaker-meanings in C.[4] Indeed, *any way of restricting the quantifiers in (5) so as to yield a proposition that would suffice for Tim to understand Chet's utterance, were he to entertain it, will be such as to generate the very same updated context as any other such proposition, if Tim were to accept it.* To illustrate this point, let's revise the story a little bit.

Tim was on beer patrol. He purchased a case of Pabst, a case of Schlitz, and a case of Hamm's, which the guests will like very much, but which Chet, given his refined taste, refuses to drink. So he tells Tim that he is going to the bodega to purchase a different brand of beer for himself. It is mutually presupposed, let's say, that Chet is selfish and stingy; he is not going to share his beer with the guests. A little later, Chet returns with a case of Belgian beer in hand. Party time draws closer and Tim starts to worry. He asks, 'Are we ready?', wondering whether the beer that they will serve the guests is ready. Chet knows that Tim is asking about whether the beer for the guests is ready and replies by uttering (5). Now suppose Tim entertains (CP6)—that every beer at the apartment is in the bucket in the backyard—on the basis of Chet's utterance. (Perhaps Tim sees Chet looking at the receipt for his Belgian beer and, due to its salience, takes 'every' to range over all the beers in the apartment.) Intuitively, Tim did not understand Chet's utterance. He might reveal his misunderstanding by following up: 'Really?! You are going to share your case of Belgian?' To which Chet

[4]In conversation, Dorothy Edgington put the following question to me: (CP1)-(CP6) are intuitively different. How can your view account for the difference? I think there's a straightforward explanation of why one might feel that (CP1)-(CP6) are different: one knows that there are contexts relative to which they would partition the live possibilities differently. The source of Edgington's intuition, I claim, is this knowledge. Having said this, let me say a little more. It seems to me that there's a perfectly good sense of 'proposition' such that, intuitively, (CP1)-(CP6) are the same proposition. It's just the plain old sense we began with, according to which a proposition is simply the unit of information communicated by a speech act. Since the unit of information one conveys or retrieves is always partly determined by the preexisting stock of information that one possesses, and since the stock of information shared by Chet and Ted comprises BEER and BUCKET, I think intuition is on my side.

says, 'No, fool, you know me better than that. I don't waste good beer.' Because any way of restricting the quantifiers in (5) so as to yield a proposition that would suffice for Tim to understand Chet's utterance will be such as to generate the very same updated context, all of the candidate propositions will count as the same asserted content and, hence, the same proposition meant by Chet.

Recall the more direct argument that Buchanan gave for rejecting CONTENT. That argument relied on the premise that the speaker "has no reason to think that her audience will recognize her as having meant just this proposition" (p. 350). But given the highly restricted space of worlds relative to which the asserted content of Chet's utterance is determined, it is perfectly reasonable for Chet to expect that Tim will recognize that he meant just this particular subset of the context set. After all, no other subset of the context set is carved out by any of the candidate propositions. So CONTENT is unscathed, if combined with a localist picture of content and conversation.

My response to the second case parallels my response to the first very closely. As Buchanan sets things up in the second case, Tim taps Chet on the shoulder and nods in the direction of the oddly dressed man. So, presumably, Chet sees the man and Tim knows that Chet sees the man. It is mutually obvious, then, that the following identities hold.

(INSPECTOR) he = the man we are looking at = that customer sniffing at his plate = the guy who is frowning at his chicken fried steak = the guy with the strange mustache

So INSPECTOR is true at every world in the context set. Consequently, at each of those worlds, one and the same guy satisfies all of the candidate ways of picking out the oddly dressed man toward whom Tim draws Chet's attention. So, at that context, (CP7)-(CP11) eliminate the same worlds. SAMENESS implies that the candidate propositions are really the same asserted content. Given the constitutive connection between asserted content and speaker meaning, it follows that (CP7)-(CP11) are the same proposition meant.

Suppose that it was not mutually obvious that the guy who is frowning at his chicken fried steak is the guy that Tim and Chet are looking at. And suppose Chet entertained (CP8)—that the man we are looking at is a health inspector—on the basis of Tim's utterance. Chet might then reply, 'Why are you so worried? He seems to be enjoying his hamburger.' To which Tim might say, 'No, not that guy; *that* guy. The

one frowning at his chicken fried steak.' Intuitively, Chet has not understood Tim's utterance. What this minor revision to the original story illustrates is that any way of identifying the oddly dressed man so as to yield a proposition that would suffice for Chet to understand Tim's utterance will be such as to generate the very same updated context if it were accepted by Chet.

Buchanan rejects the standard view. He opts for a revisionary picture of linguistic communication, positing a level of content that mediates the relationship between sentence uttered and proposition expressed. Speaker meaning is identified with an object at this intermediate level of representation. According to Buchanan's theory, by uttering (5) Chet means a "restricted proposition-type", which is a complex consisting of, first, a propositional template, or structure—

(TEMP) [the $y$: bucket($y$) $\wedge$ – ($y$)] ([every $x$: beer($x$) $\wedge$ – ($x$)]($x$ is in $y$)

—that represents the Kaplanian character of the sentence and, second, a "vague range of restrictions on how that structure is to be completed. So long as Tim constructs some one or more propositions of the form given by (TEMP) within a (vague, contextually-restricted) range of propositions, he will have understood Chet's utterance" (p. 358). Buchanan then combines this view of speaker meaning with a semantics for 'said that' and 'means that' constructions—e.g.,

(13) Chet said/meant that George W. Bush lives in Washington.

—according to which the 'that'-clause denotes a restricted proposition-type, the proposition-type conveyed by the subject of the report.

Setting aside the considerable intuitive appeal of the standard view, which Buchanan acknowledges but which his proposal lacks, there is a good reason why the localist solution ought to be preferred to the theory of restricted proposition-types. Localism is more explanatory. No special provisions were added to, nor was anything left out of, the Stalnakerian framework in order to generate the desired result: an account that specifies the object understood by the listener consistent with the impoverished lexical meanings associated with (5) and (8). I merely drew attention to an often overlooked part of the framework, namely, SAMENESS. This principle says, basically, that the asserted content of a sentence can never outrun the common stock of information that conversational partners use to interpret speech. It thereby incorporates the proper sort of context sensitivity without abandoning CONTENT or SUCCESS. So the Stalnakerian

framework enjoys the added benefits of predictive success. Not so for Buchanan's theory; restricted proposition-types were posited for the purpose of getting the desired result. If they did not succeed at the modest task for which Buchanan introduced them, then some other object would have been posited. Buchanan's theory merely accommodates the relevant data. The predictive success of a theory typically provides us with further evidence in its favor. Therefore, the localist response is better confirmed than the theory of proposition-types.

One might be tempted to reject Buchanan's account for yet another reason. According to Buchanan, although the 'that'-clause in (13) denotes a proposition-type, 'that'-clauses occurring in belief reports denote propositions. So one might worry that Buchanan is unable to accommodate the truth of 'mixed' reports, such as 'Chet both believed and said that George W. Bush lives in Washington'. After all, there is only one 'that'-clause in this report. What does it denote: a restricted proposition-type, or a full-blooded proposition? Relatedly, Buchanan acknowledges in the last footnote of his (2010) that the following inference appears valid even though his account predicts that it is invalid:

(i) Chet meant that Fichte was a philosopher.

(ii) Tim believes what Chet meant.

(iii) Therefore, Tim believes that Fichte was a philosopher.

The complement phrase in (ii), 'what Chet meant', is anaphorically linked to 'that Fichte was a philosopher' in (i). So the former denotes whatever the latter denotes, and, according to Buchanan, the latter denotes a restricted proposition-type. But given that the complement phrase in (iii) denotes a proposition, the inference is simply a non-sequitur.

Such arguments are often used in semantic theorizing. I like to call them *piggy-backing* arguments, since they rely on elliptical or anaphoric piggy-backing to illustrate that assigning a particular denotation to an expression misrepresents that expression's semantic contribution. But sometimes piggy-backing of the sort that such arguments are supposed to rule out (call it *shifty* piggy-backing) is semantically kosher. Consider Chomsky's famous example: 'France is hexagonal and a republic' or 'France is hexagonal and it is a republic'. This example is the sort of thing that one might reasonably say (according to Chomsky, it is licensed by English grammar), but it also exemplifies the allegedly problematic sort of piggy-backing, like 'Chet believed and said that GWB

lives in Washington'. A region of space is hexagonal, but not the sort of thing that could be a republic. A particular kind of social institution is a republic, but not the sort of thing that could be hexagonal. So 'France' must somehow be doing double duty, just as the 'that'-clause in the mixed report must also be doing double duty. Consider, further, the argument below.

(iv) France is hexagonal.

(v) It is a republic.

(vi) Therefore, France is hexagonal and a republic.

If English grammar does indeed license 'France is hexagonal and it is a republic', then (iv)-(vi) should be a valid inference. But notice that it involves shifty anaphoric piggy-backing. The upshot, I think, is that piggy-backing is more complicated than is often supposed. Until we understand how it works better than we now do, I think we should be reluctant to put too much weight on the apparent felicity of 'Chet believed and said ...' and the apparent validity of (i)-(iii).

Buchanan drew our attention to a theoretically important aspect of linguistic communication: often, lexical meaning is too impoverished to pin down a unique proposition meant by the speaker. But rarely does this ever get in the way of successful exchanges. Why? Because so much is mutually presupposed in any given context that a speaker can reasonably expect her meaning to be recovered, and a listener needn't discriminate between such closely related bits of information as the candidate propositions above. A localist conception of speaker meaning, which identifies contents that effect the same partition in the context set, is able to model this ubiquitous phenomenon consistent with the standard view of communicative interaction. Buchanan is right, then, to suggest that the proper way to explain the phenomenon is to acknowledge a kind of context relativity in the notion of speaker meaning, but wrong to think that the appropriate sort of context relativity forces us to give up on the standard view of communication.

I want to emphasize the ubiquity/frequency of the phenomenon. Why? Well, originally, the localist picture of conversation was presented as a repair strategy designed to accommodate abnormal cases, specifically, necessary *a posteriori* and negative existence statements (Stalnaker 1978). Viewed as a response to Buchanan's examples, however, it appears to be motivated by normal cases of communication. Ordinary speech very often involves implicit restrictions on quantifier domains, non-sentential assertion, and

incomplete definite descriptions (more on this in a moment), so the abnormal cases seem to be those where conventional sentence meaning fully determines asserted content. And as the discussion in the previous two sections illustrates, whereas standard non-localist conceptions of speaker meaning run into problems when we look at such cases, the standard localist view fares quite well. It is nice when a theory can handle abnormal cases that its rivals have a difficult time with, but when it can handle normal cases that its rivals seem unable to—well, what more favorable evidence can one reasonably hope for in this area of philosophy?

One naturally wonders whether the kind of solution recommended here can be mimicked within a framework that employs structured propositions. Perhaps, but it is difficult to see how without committing oneself to localism. On the 'structuralist' view, as I will call it, a lot of information is encoded in a proposition's internal make-up. This is typically viewed as one of the chief virtues of the theory, allowing it to respect intuitive distinctions between contents. Somehow this information has to be 'filtered out', as it were, so that the candidate propositions above aren't inadvertently represented as being distinct asserted contents. For ease of exposition, let's just focus on the first case. The structuralist will need some way to discard the bits of information that differentiate ($\text{CP}_i$) from ($\text{CP}_j$), thus getting at the 'common denominator': a unit of information shared by (CP1)-(CP6). Furthermore, the filtering device has to be available to all conversational partners, otherwise it would be unreasonable for the speaker to expect that every member of her audience can recover the common denominator. Given such constraints, what could play the role of this filtering device? The obvious answer is the context. So, regardless of how the structuralist chooses to model the context, it has to play the kind of role that it plays in Stalnaker's framework, that is, a device that strips away aspects of conventional meaning so as to identify what from the perspective of compositional semantics are distinct contents. Thus the structuralist incorporates localism into her theory of communication.

## 2.3 Incomplete descriptions

Very early on in our discussion I said that Buchanan's puzzle closely resembles Schiffer's work on the semantics of definite descriptions. As far as I can tell, there are only two differences. First, Buchanan's argument employs a wider class of constructions. Second, he, unlike Schiffer, takes the reasoning to show that the standard view itself is

defective, not any particular analysis of the sample constructions.

Schiffer's argument was designed as an objection targeting the Russellian analysis of definites. It begins with a story. Rather than over-populate our cast of characters, I will simply put our old friends, Chet and Tim, back to work.

> Suppose Chet and Tim are in a lecture hall waiting for a famous philosopher to enter and give a talk. As the philosopher enters, he stumbles. Chet then says: 'The guy is drunk.' Even before [Chet's] utterance, it was mutually evident that [Chet and Tim] had knowledge of the professor under numerous shared definite descriptions—the author of *Smells and Tickles*, the only man within sight wearing a yellow jacket and red golf pants, the man we are waiting to hear, the man now staggering up to the podium, and the list, in any realistic situation, will go on and on ... Imagining myself as your audience, I do not see how I could have identified any one individual concept, however complex, as the one which figured into the proposition that you asserted. And yet it would seem that I understood your utterance perfectly well (Schiffer 1995, p. 376).

The point is that "in any realistic situation" there is no one proposition that the listener can reasonably be expected to recover on the basis of the speaker's utterance, 'the F is G', because there are multiple candidate propositions consistent with the conventional meaning of the uttered sentence. How, exactly, does Schiffer's observation challenge Russellianism about definites? Well, presumably the Russellian wants to say that Chet asserted the proposition that [the $x$: $x$ is male $\wedge$ F($x$)]($x$ is drunk), where 'F' is a schematic predicate-letter standing in for the contextually relevant descriptive term that specifies the asserted content of Chet's utterance. But—and here is Schiffer's point—there is no predicate with which 'F' can be replaced to yield a proposition that Tim can reasonably be expected to recover. The reason is that there are multiple equally well qualified contextually salient candidate predicates, each doing its part to express a distinct proposition when plugged in for 'F'. Since Chet is a competent speaker, he himself must be aware of the fact that Tim can't reasonably be expected to recover any of the candidate propositions. So the Russellian is unable to say what she wants, because a speaker asserts P only if she believes it reasonable to expect of her conversational partners that they be able to recover the proposition P.

Russellians have replied to Schiffer's argument by tweaking the naive account in

various ways.[5] But, as Buchanan and Ostertag (2005) observe, there are problems with the proposed amendments. The problems suggest, according to them, that the standard view is to blame, not Russellianism. Additionally, one should expect relevantly similar phenomena to be subsumed under the same explanation. So given the obvious similarity between Schiffer's argument and the problem of implicit quantifier domain restriction and non-sentential assertion, one should expect that a satisfactory response to Schiffer's objection would solve all three problem cases (implicit domain restriction, non-sentential assertion, definite descriptions) simultaneously. It is misguided, then, to look for the solution in the analysis of any one construction. What the examples bring out, as I noted a little earlier, is that asserted content/speaker-meaning is infected with context relativity, and localism provides a way of modeling that relativity consistent with the standard view of communication.

The localist response to Schiffer's argument should be clear at this point. The problematic assumption is that the candidate propositions are different asserted contents in the context at which the speech act takes place. As Schiffer sets things up in the passage I quoted above, Chet and Tim mutually presuppose that the author of *Smells and Tickles* is the only man within sight wearing a yellow jacket and red golf pants, and so on. So all of the relevant identifying descriptions pick out the same guy at each world in the context set. So each candidate proposition determines the same partition of live possibilities. So according to content localism, they count as one and the same asserted content/speaker meaning in that context.

## 2.4 Indeterminacy and the first person

Not every promise has been kept. I advertised that I would have something to say about meaning indeterminacy and its relation to Buchanan's puzzle. In this section, I intend to deliver.

The doctrine of pervasive semantic indeterminacy is often attributed to Quine (1960, ch. 2). As it appears in Quine's work, it is a thesis about translation: there are multiple, inconsistent mappings from the object language to the metalanguage, each respecting patterns of stimulus and response among native speakers of the object language. I prefer to focus on the case for indeterminacy as it appears in Davidson's (1984), since it gets rid of unnecessary behaviorist trappings. We can simplify things even further

[5]See, for instance, Stephen Neale (2004).

if we focus on a special case of the indeterminacy doctrine: the thesis of reference inscrutability, which says that there are multiple, inconsistent ways of assigning referents to singular terms in the object language, each of which is fully confirmed by all of the evidence, known and unknown. Before I get to Searle's objection, let me quickly motivate this counterintuitive doctrine.

Suppose you are charged with the task of constructing a compositional semantic theory for a hitherto unknown language, L. Davidson called this project "radical interpretation". How are you, the radical interpreter, to carry out your task? To start, you need to determine what a core set of sentences in L mean. But how are you to do this? Well, if you had some idea of what speakers of L believed/wanted in contexts where they produce speech behavior, then, Davidson suggested, you could use that knowledge to triangulate what they meant. Once you observe a stable pattern of meaning P by uttering a particular sentence, *s*, you can reasonably hypothesize that *s* means P. One might worry at this point that you have no more insight into what speakers of L believe/want than you do into what the sentences of L mean, since your primary way of knowing what someone believes/wants is by understanding the sentences she utters. The worry isn't altogether baseless, but it exaggerates the problem. True, the ascription of beliefs/desires to speakers of L is largely guess work, but it isn't entirely unprincipled. You presume that speakers of L are not massively confused or mistaken: that they agree with you about a great many things, and that they want what you would want if you were in the position they happen to be in. Ascriptions of divergent attitudes are justified only insofar as one can supply a good explanation for the divergence, e.g., that the speaker is blind and, hence, can't see that the cat is on the mat. Once you have assigned meanings to a large enough set of sentences, you can parse the sentences into smaller elements (names, predicates, quantifiers, pronouns, etc.) and assign the elements denotations. This part of the project is governed by one important constraint: that the assignment allow you to systematically compute the meaning of a whole sentence as a function of the meanings you assign to its parts and how they have been arranged. An assignment is acceptable only if it satisfies this condition.[6]

Now suppose you have your semantic theory of L in hand. What would count as evidence for it? What would it take to empirically verify the theory? Davidson,

[6]One might place further demands on the theory: that it assign only objects with which speakers of L have had the appropriate sort of causal contact, etc. It's likely that even then some degree of indeterminacy will persist.

following Quine, proposed that "all of the evidence for or against a theory of truth (interpretation, translation) comes in the form of facts about what events or situations in the world cause, or would cause, speakers to assent to, or dissent from, each sentence in the speaker's repertoire" (1984, p. 230). If your semantic theory, when coupled with your psychological profile of the speaker, predicts that she would assent to '*Gavagai*!' in such and such conditions, and the speaker does indeed assent to it under such and such conditions, then your semantic theory is verified. Where this sort of evidence gives out—where it falls short of deciding between two inconsistent theories—so too do the semantic facts about L. Consequently, once we recognize that there are multiple empirically equivalent, fully satisfactory yet inconsistent reference assignments, we must acknowledge that no one reference assignment is the right one and, hence, that referring terms in L don't determinately refer to any of the candidate referents left open.

For heuristic purposes we, following Davidson, assumed that L was a foreign language, and that speakers of L were members of an alien community, but this assumption is dispensable. The point is: when you interpret me, you go through a similar process, relying on your representation of my psychological profile to determine what I mean. You have no better insight into what I mean by my speech acts than you have into what speakers of L mean by theirs. Don't say that, in our case, you have greater insight because we speak the same language; how do you know we speak the same language? Granted, the pattern of sounds I emit closely resembles the pattern of sounds you emit, but for all you know I speak a different language with a remarkably similar phonology. As Davidson says, "The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same?" (p. 125) The hypothesis that you and I speak the same language is verified in precisely the same way that hypotheses about L are: if it allows us (by and large) to successfully coordinate, then it's true. If it doesn't, then it isn't.

Davidson (1984, p. 229-231) provides a nice example to help illustrate and (to some extent) motivate the idea that there are multiple empirically equivalent yet inconsistent semantic theories. Suppose everything has a shadow. Then we can semantically decompose the sentence 'Wilt is tall' so that 'Wilt' refers to Chamberlain and 'is tall' denotes the property of being tall. Alternatively, we can take 'Wilt' to refer to Chamberlain's shadow and 'is tall' to denote the property of being the shadow of something that is tall. The first theory implies that 'Wilt is tall' is true iff Wilt Chamberlain is tall;

the second implies that 'Wilt is tall' is true iff Chamberlain's shadow is the shadow of something that is tall. With enough compensatory adjustments elsewhere in one's semantic theory, the second theory can be made to accord with all of the constraints on interpretation. And, since Chamberlain is tall iff Chamberlain's shadow is the shadow of something that is tall, the two semantic theories will be confirmed by exactly the same facts about assent/dissent. Given the argument from two paragraphs back, it is indeterminate whether 'Wilt' refers to Chamberlain or his shadow or, for that matter, any other distinct object appropriately related to Chamberlain (e.g., a particular time-slice of Chamberlain: Chamberlain@NOW) and his height. So there are multiple non-equivalent candidate propositions that an interpreter can assign to a speaker's literal, assertive utterance of 'Wilt is tall'.

(CP12) that Chamberlain is tall

(CP13) that Chamberlain's shadow is the shadow of something tall

(CP14) that Chamberlain@NOW is tall

.

.

.

But suppose I were to utter 'Wilt is tall' in a literal, assertive tone of voice. Wouldn't *I* know what I meant? Searle (1987) thinks so, and he says we can respect this first-personal judgment about speaker meaning only if we regard the doctrine of indeterminacy as a *reductio* of the principles from which it was derived.

> ...on Davidson's view the indeterminacy follows only if we assume from the start that different semantic facts must necessarily produce different "publicly observable" consequences. ...But, I submit, we know quite independently that this conclusion is false, and, therefore, the premises from which it is derived cannot all be true. How do we know the conclusion is false? We know it because in our own case we know that *we mean*, e.g., Wilt as opposed Wilt's shadow, rabbit as opposed to rabbit stage (1987, p. 141, emphasis added).

Observe that Searle's point is about speaker meaning, not sentence meaning: we know the indeterminacy doctrine has to be false because "we know that *we mean*, e.g., Wilt as

opposed to Wilt's shadow...". There is a passage in Davidson that addresses the sort of worry that Searle has in mind.

> Perhaps someone (not Quine) will be tempted to say, "But at least the speaker knows what he is referring to." One should stand firm against this thought. The semantic features of language are public features. What no one can in the nature of the case figure out from the totality of the relevant evidence cannot be a part of meaning. *And since every speaker must, in some dim sense at least, know this, he cannot even intend to use his words with a unique [meaning] for he knows that there is no way for his words to convey the [meaning] to another* (1984, p. 235, emphasis added).[7]

Davidson's response is more than a little puzzling. If one doubts his picture of interpretation on the grounds that it conflicts with a deeply felt first-personal conviction—that one knows what one meant by uttering 'Wilt is tall'—it helps not at all to be told that "in some dim sense" one knows that one really does not know. This is all the more puzzling, given that elsewhere Davidson wrote, "Kurt utters the sentence 'Es regnet' and under the right conditions we know that he has said that it is raining" (p. 125). Is Davidson suggesting that third-personal knowledge of speaker meaning is more secure that first-personal knowledge? Note further the italicized part of the passage; it strongly suggests the kind of argument that Buchanan and Schiffer used to undermine CONTENT and the Russellian analysis of definites, respectively. A speaker means P by her utterance U only if she has good reason to expect that her audience is likely to entertain P on the basis of U. But there is no candidate interpretation such that I have good reason to expect that you will assign it to my words on the basis of my utterance. So I "cannot even use [my] words with a unique [meaning] for [I know] that there is no way for [my] words to convey the [meaning] to another". And if I determinately meant anything, then surely I meant one of the candidate interpretations. Therefore, I didn't determinately mean anything. And since I am a competent speaker, I must "in some dim sense at least" know all this.

Davidson and Searle are wrong in thinking that the indeterminacy doctrine and the first-personal conviction are inconsistent. Suppose I utter 'Wilt is tall'. Then according to Davidson it is indeterminate whether the sentence I uttered means (CP12), (CP13),

[7] In the original text, 'reference' occurs in place of the bracketed additions. Since Searle's point is put in terms of meaning, I have taken the liberty of amending Davidson's preemptive response.

(CP14), or ..., but one should keep in mind that indeterminate sentence meaning can, in context, result in determinate speaker meaning. We saw how in Sections 3 and 4, and I think the very same move can help explain where the mistake above lies. Suppose the asserted content of my utterance, and, hence, what I mean by performing my speech act, is defined in terms of the live possibilities left open by what we mutually presuppose. So, if (CP12)-(CP14) partition the live possibilities in exactly the same way, they are the same asserted content in the context where I uttered 'Wilt is tall'. Now in any normal context (CP12)-(CP14), along with any other candidate interpretation of 'Wilt is tall', would partition the context set in precisely the same way. (Think about what the conversational partners would have to presuppose in order for (CP12)-(CP14) to effect different partitions. Any such context would be highly bizarre!) Because the interpretation of speaker meaning occurs within such a narrowly restricted sub-region of logical space, to which you and I both have cognitive access, I have good reason to expect that you will recover my meaning. Davidsonian indeterminacy is, therefore, perfectly consistent with the first-personal conviction that each of us, in our own case, know what we meant. Localism is the key to reconciliation.

## 2.5 In closing

Arguments come and go. Sometimes they re-appear in different trappings to support different conclusions. The so-called Frege-Gödel Slingshot is an example of one such argument. It can be found in Church, Quine, Davidson, and Neale. Each author puts it to use in his own way. The argument we have looked at here is another example. It shows up in Buchanan and Ostertag (2005), Buchanan (2010), Schiffer (1995), and in an exchange between Searle (1987) and Davidson (1984). Buchanan (and Ostertag) use it to generate a puzzle for the standard view of communication. Schiffer relies on it to criticize the Russellian theory of definite descriptions. And Davidson puts it to use in defense of pervasive meaning indeterminacy. I think the argument is flawed in all its incarnations. The flaw in each case is the implicit assumption that content localism is false. By acknowledging and correcting the flaw, the standard view, Russellianism about definites, and the doctrine of meaning indeterminacy escape the problems that allegedly confront them. Since the standard view is so attractive (indeed, even its critics think so) we end up with a rather compelling argument in favor of localism about content.

CHAPTER 3

# There Couldn't Have Been Zombies, but it's a Lucky Coincidence That There Aren't

I want to ask an old question and provide a new answer, or perhaps dress an old answer up in new, more attractive, clothes. The question is why should the correlation between the phenomenal and the neurophysiological appear contingent even to advocates of physicalism? Why should (some) physicalists be at all disposed to judge it contingent that pain accompanies c-fiber firing, given that they think it's metaphysically necessary that it do so?[1] Admittedly, not all physicalists are so disposed. Some are hard-nosed; they say that zombies—physical duplicates of actual humans in whom consciousness is absent—are not just metaphysically impossible, they can't even be coherently imagined. Such physicalists are convinced that there is no explanatory gap between the phenomenal and the physical, or that the gap can be closed with relative ease once we determine how the brain processes, stores, and transfers information. Hard-nosed physicalists face a different (perhaps more difficult) challenge: why has the incoherence of the zombie scenario escaped the attention of so many sharp philosophers of mind for as long as it has?

In this chapter I will assume a kinder, gentler, more concessive sort of physicalism. Call it 'soft-nosed physicalism'.[2] The soft-nosed physicalist thinks that there couldn't

[1] I use 'c-fiber firing' as a schematic expression to stand in for whatever correctly describes the neurophysiological basis of pain in us.

[2] In the next chapter I will defend hard-nosed physicalism.

have been zombies, but feels considerable pressure to say that they're possible. (This appears to be the dominant view among philosophers.[3]) The question I'm concerned with is why does she feel such pressure? What is its source? The urgency of the question is largely polemical; without a persuasive answer that coheres with her commitment to physicalism, the appearance of contingency tells in favor of dualism. *With* such an answer—one that explains both why the appearance of contingency should be expected to arise and why it should be as persistent (hard to shake) as it is—one major consideration in favor of dualism will be neutralized.

My answer can be divided into three parts. First, I observe that there's a distinction between accidentality or coincidence on the one hand and metaphysical contingency on the other. It can be an accident—a sheer coincidence—that $p$ even when it's metaphysically necessary that $p$. Second, I show that the epistemological considerations that traditionally challenge physicalism merely establish that phenomenal states or properties and corresponding neurophysiological states or properties are accidentally, or coincidentally, related. These considerations dispose even the (soft-nosed) physicalist to (correctly) judge that it's an accident, or a coincidence, that there's something it's like for one's c-fibers, say, to fire. Finally, I suggest that our inattention to the distinction between something's being an accident, or a coincidence, and its being metaphysically contingent results in frequent equivocation. As a result, the disposition to (correctly) judge that a phenomenal state or property and its neurophysiological basis are accidentally/coincidentally related manifests itself as a disposition to (incorrectly) judge that they are contingently related. I will show that the error of equivocation runs quite deep here. It's not something one can avoid by simply updating one's semantic beliefs. In this respect, the disposition to incorrectly judge that psychophysical relations are contingent resembles an optical illusion, which persists even when the subject acknowledges that the way things visually appear diverges from the way things really are.

I suspect that readers will be highly skeptical, even at this early stage. It may seem to them that my account is just contradictory. Consequently, some may want to stop reading here. But I ask that readers be patient. As odd as it may seem, I will rely on the fact that my answer is contradictory-sounding when accounting for the persistence of the contingency illusion.

Here is a quick preview of things to come. I begin in Section 1 by clearing some

[3]The basis for this judgment is the recent PhilPapers Survey.

ground. I explain why the model of modal error that we've inherited from Saul Kripke's celebrated work is defective. My purpose in this section is to demonstrate that modal rationalism is false, and, therefore, that an error theory of the sort I want to provide needn't satisfy the constraints that a rationalist would impose. I turn to my central question in Sections 2-4, unpacking the account I quickly sketched a moment ago. My goal in this part of the discussion is to identify and clarify a notion that's both sufficiently like and importantly unlike metaphysical contingency so as to accommodate what is perhaps the most forceful consideration in favor of dualism. There is, according to the soft-nosed physicalist, a kernel of truth in the dualist's challenge to her metaphysical doctrine; it would be a shame if partisan bickering barred us from acknowledging it. I close in Section 5 by addressing a series of possible objections to my account and summarizing the take-home message.

## 3.1 Why modal rationalism is false

According to one interpretation of recent philosophical history—an interpretation that has its home in Australia—one of the major contributions of Kripke's lectures, *Naming and Necessity*, was an outline of a general theory of modal error—a theory that promised to explain why merely apparent possibilities are actually impossible in such a way as to render it unsurprising that there are modal illusions. As if that wasn't enough, Kripke's explanation was supposed to be so compelling that it would (i) convince even one's interlocutor of the inaccuracy of her modal judgment, and (ii) vindicate modal rationalism—the doctrine that conceivability (of a sort) implies possibility (of a sort). Kripke's outline has recently been filled in by David Chalmers (1996, 2002, 2010), and the resulting theory plays a key role in his 'two-dimensional' argument for dualism. To concisely state the Kripke-inspired theory of modal error, we will have to help ourselves to some philosophical jargon.

According to Chalmers, content is two-dimensional; sentences of natural language semantically express both a *primary* and a *secondary* intension. The secondary intension of an arbitrary declarative sentence is the set of possible worlds (considered as counterfactual) *of* which the sentence is true. To determine whether a particular world, $w$, is a member of the sentence's secondary intension, you hold fixed the actual extensions of the terms occurring in the sentence and ask whether it's true of $w$. So the secondary intensions of 'water quenches thirst' and 'heat is present in Room 203' are, respectively,

the set of worlds at which $H_2O$ quenches thirst and the set of worlds at which high mean kinetic energy is present in Room 203. In contrast, the primary intension of a sentence is the set of worlds (considered as actual) *in* which the sentence is true. To determine whether $w$ is a member of the sentence's primary intension, you allow the extensions of the terms occurring in the sentence to vary with the facts in $w$ and ask whether the sentence is true in $w$. So the primary intensions of our sample sentences are, respectively, the set comprising every world, $w$, at which the most abundant clear, odorless, tasteless, drinkable liquid at $w$ quenches thirst at $w$, and the set consisting of every $w$ such that the phenomenon in $w$ that produces sensations of heat is present in my actual office in $w$. To illustrate the distinction between the two dimensions of content still further, take any old sentence you like. Replace any names or natural kind terms occurring in the sentence with descriptive phrases that capture the way a competent speaker might mentally represent the corresponding extensions. The resulting sentence will express the primary intension of the original. Applying this technique to the examples above we get 'the most abundant clear, odorless, tasteless, drinkable liquid around here quenches thirst' and 'the phenomenon that produces sensations of heat in us is felt in my actual office'.

According to the theory of modal error that Kripke inspired and Chalmers defended, when it's necessarily true that $p$ and yet one seems able to conceive of a situation in which it's false that $p$, different dimensions of $p$'s content are involved. What's necessarily true is $p$'s secondary intension; what one conceives to be false is $p$'s primary intension. Since $p$'s primary intension corresponds with the subject's way of mentally representing its secondary intension, it's not surprising that her ability to conceive of the primary intension's truth would dispose her to think that $p$'s secondary intension is contingent. But—and here is where modal rationalism is vindicated—if one can't rule out the coherence of $p$'s primary intension on the basis of *a priori* reflection under optimal circumstances, then $p$'s primary intension *is* possible.

Following Chalmers (2002, 2010), let *ideal primary conceivability* be the inability to rule out the primary intension of a sentence by *a priori* means under optimal cognitive conditions, and let *1-possibility* be the sort of possibility that attaches to a sentence's primary intension. Then we can express the sort of modal rationalism that falls out of the above story as follows.

(MR) If it's ideally primarily conceivable that $p$, then it's 1-possible that $p$.

The motivation for MR is that it appears to accommodate remarkably well all of the paradigm examples of necessary *a posteriori* truth. With MR in hand, Chalmers is able to mount his two-dimensional argument for dualism.

Alas, MR is false. To illustrate my point, I want to slightly tweak a standard example of empirical necessity and ask how the theory handles the slightly revised case. It will become clear, I think, that the Kripke model is rather unconvincing when applied to the new case, and that we thus have a genuine counterexample to MR.

Consider Kripke's explanation of why it merely appears to us as though heat could have been something other than high mean kinetic energy (HME). "This very phenomenon [the actual microphysical nature of heat] could have existed, but due to differences in our neural structures and so on, have failed to be felt as heat" (Kripke 1980, p. 133). Given a different neural structure, *low* mean kinetic energy (LME) could have produced sensations of heat. According to Kripke, we imagine a situation where such is the case when we mistakenly judge that something other than HME could have been heat. But doesn't it also seem to us that the microphysical nature of heat could have been different from what it actually is without a change in our neural structure? It certainly seems possible to me, by which I mean that no amount of purely *a priori* reflection is going to rule out the supposition that, holding our internal constitution fixed, LME is the microphysical basis of heat. In fact, when I judge that heat could have been something other than HME, I'm imagining a situation that departs from actuality in a relatively minimal way. It's built into the content of my imagining that my neural structure is just as it is. How, then, would Kripke explain away the appearance of this possibility? Bracket the question aside for a moment. We'll come back to it shortly. For now, contrast Kripke's remarks about heat with his comments about the necessity of material constitution.

My dining room table is actually made of wood. Could it, then, have been made of some other kind of stuff? There's some temptation to say yes, it could have. What's the source of this temptation? According to Kripke (*Ibid.*, p. 114), the temptation stems from the fact that we imagine a situation in which I have a distinct dining room table—one that's made of ice, say, and that superficially resembles my actual dining room table to such a degree that my sensory faculties *as they actually are* can't tell the difference. It's not surprising that I would confuse a possibility involving something *I* couldn't tell apart from my dining room table with a possibility involving my dining room table. Why would I confuse $x$ with $y$ when, given my sensory architecture, they

appear different? Explanations of modal error are convincing only when a genuine epistemic counterpart of the relevant phenomenon is involved in the explanation.

In order for Kripke's explanatory strategy to work in the case of heat, there has to be something (Stephen Yablo (2008) calls it "fool's heat") that stands to HME as the epistemic counterpart of my dining room table (the qualitatively identical table made of ice) stands to my wooden dining room table. Fool's heat (like fool's gold) is something that perturbs our actual sense faculties in such a way as to cause sensations of heat (appearances of gold) in the absence of HME (Au). One wonders whether there is such a thing. And I feel considerable pressure to say no, there isn't. "It may be possible to slip a cleverly disguised icy table in for this wooden one while preserving visual appearances. *But it is not possible to slip cleverly disguised* LME *in for* HME *and have it feel just the same*" (Yablo 2008, p. 189, emphasis in original). Kripke too must have thought there isn't any such thing as fool's heat, otherwise his explanation would likely have relied on it rather than on observers with different sensory faculties. Without an epistemic counterpart of heat, the Kripke-inspired explanation of the apparent possibility is unconvincing.

But that's not all. Let 'NA' abbreviate a complicated sentence that characterizes, in precise technical terms, our actual neural architecture. Then it's ideally conceivable that

(1) NA and both heat and LME are present in Room 203.

The secondary intension of (1) is, I presume, impossible. Heat, after all, just is HME. So the Kripke-Chalmers theory predicts that what one actually conceives is the primary intension of (1), which is captured by

(2) NA and both the phenomenon that produces sensations of heat and LME are present in my actual office.

But even *that* is impossible, as Kripke himself tacitly acknowledges by assuming that our neural structure has to differ at worlds where the sensation of heat is produced by something other than HME. So substituting (1) for $p$ in MR results in a falsehood, because it's ideally primarily conceivable yet not 1-possible. Since the two-dimensional argument for dualism indispensably relies on MR, it cannot succeed.

But it would be premature for physicalists to rest content. One still wants to know why even *they* should be unable to shake the temptation to say that psychophysical

relations are contingent. In the next three sections, I take up this explanatory challenge. But let's be clear about what the challenge is asking for.

I began this essay by suggesting that there's one explanatory project: why do those of us with physicalist sympathies feel as though psychophysical correlations aren't necessary? I want to cancel the suggestion. There are at least two different projects one might have in mind when asking this question. One can try to answer it in the way that I will, by showing that the feeling of contingency should be expected even if the corresponding necessity statement were true. Or, alternatively, one might try to answer the question by telling a story that convinces the unconvinced that psychophysical relations are indeed necessary. The former project is defensive in nature (it takes physicalism for granted, the goal being to force a stalemate) while the latter is offensive.

Kripke's strategy in *Naming and Necessity*, when discussing non-phenomenal empirical necessities, was offensive; he sketched the beginnings of a theory that was supposed to undermine our grounds for thinking that the correlation between water/$H_2O$, heat/high mean kinetic energy, lightning/electrical discharge, etc. is contingent. We were then supposed to just recognize—due to the absence of any (bad) reason to the contrary—that the corresponding theoretical identifications are necessary truths, provided that they're actually true. Now I'm typically of the mind that the best defense is a good offense, but sometimes one has to make do. I doubt that there's a winning offensive strategy in this game. My skepticism isn't restricted to illusions of psychophysical contingency; I doubt whether there's a successful offensive strategy for even some of the standard cases. It would be a significant blow to physicalism if all illusions of contingency save the zombie intuition could be explained away on the Kripke model. What's more likely: that this highly attractive, very well motivated model actually breaks down when applied to psychophysical illusions, or that the zombie intuition can't be explained away because it accurately tracks modal truth? Fortunately, the model breaks down when we consider non-phenomenal empirical necessities.[4] But even if I'm wrong about this last point, I think everyone involved in this debate—even Chalmers, the consummate modal rationalist—should agree that there's no generally successful offensive strategy for explaining away modal errors. Let me explain why.

[4]Chalmers (2010, p. 182) and Eli Hirsch (2010, pp. 118-119) seem to demand of a convincing physicalist theory that it be offensive in nature. The motivation for this demand is, presumably, that paradigm empirical necessities can be accommodated with an offensive explanatory strategy. The objection to MR undermines this motivation.

The literature on modal epistemology is full of putative counterexamples to MR. One of the more compelling challenges is itself a modal statement. Consider the sentences below.

(3) There's a necessary creator of the universe.

(4) There isn't a necessary creator of the universe.

The primary intension of (3) coincides with its secondary intension. The same is true of (4). Additionally, both sentences are ideally conceivable (Yablo 2000). But, assuming $S_5$, (3) and (4) can't both express possibilities, otherwise a contradiction would follow. Since MR implies that they both *do* express possibilities, it must be false. Relatedly, the primary and secondary intensions of MR coincide, and $\neg$MR is (allegedly, anyway (Howell 2008)) ideally conceivable. So if MR is true, then $\neg$MR is possible. But presumably, if MR is true, then it's necessarily true. So, assuming S5, contradiction follows. In response to both objections, Chalmers (2010, pp. 178-180) denies the first step; (3) and $\neg$MR aren't, according to him, ideally conceivable.[5] Perhaps Chalmers is right (I'm skeptical), but even if he's right, a further question is left open: why should (3) and $\neg$MR appear possible to his critics? Since their primary and secondary intensions coincide, one can't simply rely on the Kripke-inspired error theory to resolve the puzzle. And Chalmers provides no alternative explanation. It must just be that his critics have failed to think things through properly.[6] But that's not the sort of answer that's likely to convince *them* that, in this case, the appearance of possibility is mere.

So, to sum up, we all have to play defense. Some defensive strategies are, however, better than others. It would certainly make me feel uncomfortable if the best we physicalists could do was allege that dualists have failed to think things through properly—end of story. That's not much of an error theory. One wants something more informative and less dogmatic. And I think I can offer just that.

[5]My (3) and (4) differ slightly but importantly from Yablo's sample sentences. Yablo's objection was put in terms of God, an omniscient, omnipotent, omnibenevolent, necessary being. It may well be that the notion of God *is* incoherent—the logical puzzles arising from the supposition that there is such a thing are common knowledge—but there's much less content to the notion of a necessary creator. Similar puzzles don't arise on the supposition that there is, say, a necessary First Cause. So I think the response that Chalmers favors to Yablo's version of the objection is rather unattractive when extended to my version above.

[6]Indeed, comments in Chalmers (2010, p. 179) suggest as much.

## 3.2 Coincidence without contingency

Within the practice of mathematics, experts sometimes classify particular results, observations, even definitions as accidents/coincidences. (They use the terms 'accident' and 'coincidence' interchangeably.) In the earliest paper on the subject that I'm aware of, Philip J. Davis—a professor of applied mathematics at Brown University—writes that "for the working mathematician, coincidence exists. He feels it, he identifies it, he uses it as an inductive and constructive element" (1981, p. 320). One might quibble over particular cases, but there are some highly compelling examples of mathematical coincidence/accidentality that I'm going to take at face value. The best way to begin making sense of the phenomenon is by looking at such examples. We can then ask what they have in common, how they differ from nonaccidental truths. Once we have clarified the notion of mathematical coincidence, the temptation to resist thinking that there are such things is, I wager, likely to fade away, because it will then be clear that one's resistance was based on a misunderstanding.

### 3.2.1 Pinning down the phenomenon

FIRST EXAMPLE –

"...consider the decimal expansion of $e$, which begins 2.71828 1828...It is quite striking that a pattern of four digits should repeat itself so soon—if you choose a random sequence of digits then the chances of such a pattern appearing would be one in several thousand—and yet this phenomenon is universally regarded as an amusing coincidence ..." (William Timothy Gower 2007, p. 37).

SECOND EXAMPLE –

A *factorion* is a natural number that equals the sum of the factorials of the values expressed by the individual digits in its numerical expression. For example:

$$145 = 1! + 4! + 5!$$

There are just four such numbers: 1, 2, 145, and 40 585. Why shouldn't there be any other factorions? Don't waste your time trying to come up with an answer; it's merely a coincidence.

THIRD EXAMPLE –

Take the decimal representations of $\pi$ and $e$. On average, the same term seems to occur once every ten digits. This hasn't been proved, of course, but the pattern is robust. If true, one wonders why $\pi$ and $e$—outgrowths of two separate branches of mathematics—should be related in this way. It's probably just an accident (Davis 1981, p. 313).

FOURTH EXAMPLE –

Consider the following approximation:

$$1782^{12} + 1841^{12} \approx 1922^{12}$$

(The terms flanking the approximate identity symbol express values that differ by about $3 \times 10^{-8}$ percent.) This is one of several cases that come damn close to disproving Fermat's Last Theorem, which says that for any integer value, $n > 2$, $x^n + y^n \neq z^n$. But one shouldn't take such 'near misses' seriously. We know that Fermat's Last Theorem is true. So it's nothing more than a noteworthy accident that there *are* near misses (Noam Elkies, unpublished).

FIFTH EXAMPLE –

In 1976 Kenneth Appel and Wolfgang Haken proved The Four Color Theorem: given any division of a plane into contiguous regions (e.g., a map of the United States) the regions can be colored with at most four colors so that no two adjacent regions have the same color. Part of the proof involved a computer program that was used to run through 1476 different maps on a case by case basis. Why was no counterexample found? There's no informative answer. It's just a remarkable coincidence that there isn't one (Alan Baker 2009).

The first two examples involve patterns that unexpectedly give out, while the others involve more robust regularities that suggest something deeper at work. When we discover that the suggestion is false, we judge the regularity to be a coincidence. In light of this, one might naturally think that mathematical accidents/coincidences are just truths in mathematics that evoke, or have the capacity to evoke, false expecta-

tions. Call this *the false expectation theory* of mathematical coincidence. It's an attractive proposal, but ultimately incorrect. There are mathematical truths that generate false expectations—or, at any rate, have the capacity to—but which aren't accidents. Let's look at one such case.

Consider the equation below.

(5) $x = \sqrt{1141y^2 + 1}$

Now let T be the truth that there's no value for $y$ between, say, 1 and 10 000 inclusive which renders $x$ an integer. T would probably engender the expectation that there simply is no value of $y$ that would make $x$ an integer. But, as it turns out, this expectation is false. The first value of $y$ that renders $x$ an integer is 30 693 385 322 765 657 197 397 208. Still, T is no coincidence. Why? Because the theory of Pell's equation tells us that if the decimal expansion of $\sqrt{n}$ has a long period,[7] then the first solution to the Diophantine equation below is bound to be really, *really* large.[8]

(6) $x^2 - ny^2 = 1$

Letting $n = 1141$ and solving (6) for $x$ gives you (5). So, given the length of $\sqrt{1141}$'s period, we have an explanation of why $x$ isn't an integer when $y = 1, 2, 3, \ldots 10\,000$. Each value of $y$ is less than a really, *really* large number (Davis 1981, p. 313).[9]

Our counterexample to the false expectation theory provides us with a clue about how to properly understand mathematical coincidence/accidentality. There's an important relationship between mathematical coincidence/accidentality and mathematical explanation. It's because we have an explanation of T that it's not a coincidence. Furthermore, the coincidences in each of the five examples that I presented lack a (unified) explanation. Let me quickly run through them to illustrate my point.

Reconsider the first example: it's a coincidence that a pattern of four digits should repeat itself so soon in the decimal expansion of $e$. One can compute the value of $e$ in different ways. However one chooses to do so, the computation would prove that the third through sixth digits are '1828' and that the seventh through tenth digits are

[7]The period of a decimal expansion is the length of its smallest sequence of repeating digits. For example, the period of 5/7 (= .**714285**714285) is 6.

[8]A Diophantine equation is a polynomial whose variables can take only integer values.

[9]Drawing on Nick Lord's work, Marc Lange (2010, p. 313) presents a different counterexample to the false expectation theory of mathematical coincidence.

'1828', but it wouldn't explain why the seventh through tenth digits repeat the third through sixth. "There is ... no reason why this pattern of digits repeats. It just does" (Marc Lange 2010, p. 324). Similarly, we can determine that each of 1, 2, 145, and 40 585 is a factorion by verifying the corresponding equations, but that wouldn't explain why there shouldn't be any other such numbers. Nor is there some other explanation of the fact that there aren't. Consider now the third example: that the decimal representation of $\pi$ and $e$ coincide (on average) every tenth digit. To verify this, we first have to compute the value of $\pi$ and then the value of $e$; we don't possess a single procedure, much less an explanatory procedure, to verify the convergence. The next example resembles the last three. We have to compute the near misses individually; there's no general explanation of why there should be equations that come so close to disproving Fermat. The final example is no different. The Appel-Haken proof of The Four Color Theorem is widely regarded as unexplanatory; it shows that the theorem is true, but it's far too disjunctive to explain why it is. Perhaps each one of the 1476 individual cases in the proof explains why a particular map isn't a counterexample, but there's no unified explanation of why the theorem is true.

### 3.2.2 Explanation in mathematics

Talk of explanation in mathematics may seem puzzling. Typically we think something admits of explanation only when it could have been otherwise. After all, the explanation is supposed to provide information about why things actually aren't otherwise. But mathematical truths couldn't have been otherwise. So isn't the very notion of a mathematical explanation confused? What exactly is a mathematical explanation?

Puzzling though it may be, there's no denying that explanation plays an important role in mathematics. To demonstrate the point, consider the passages below. The first appears in a review of a recent text on the quadratic reciprocity theorem: $x^2$ is congruent to $p$ modulo $q$ and $x^2$ is congruent to $q$ modulo $p$.[10] The second passage is an autobiographical description by a well known contemporary mathematician.

> We typically learn (and teach!) the law of quadratic reciprocity in courses on Elementary Number Theory. In that context, it seems like something of a miracle. Why should the question of whether $p$ is a square modulo $q$ have any relation to the question of whether $q$ is a square modulo $p$? After all, the

[10] $a$ is congruent to $b$ modulo $c$ if $(a - b)/c$ is an integer.

> modulo $p$ world and the modulo $q$ world seem completely independent of each other... The proofs in the elementary textbooks don't help much. They prove the theorem all right, but they do not really tell us why the theorem is true. So it all seems rather mysterious ...and we are left with a feeling that we are missing something.[11]

> I remember one theorem that I proved, and yet I really couldn't see why it was true. It worried me for years and years ...I kept worrying about it, and five or six years later I understood why it had to be true. Then I got an entirely different proof ...Using quite different techniques, it was quite clear why it had to be true (Atiyah 1988, p. 305).

There are hundreds of different proofs of quadratic reciprocity, eight of which were discovered by Gauss. But the feeling that we were missing something led Hilbert to set the "proof of the most general law of reciprocity in any number field" as ninth on his list of central problems (Tappenden 2008a, p. 261). What Hilbert wanted, of course, was an *illuminating* proof, one that removed the feeling that we were missing something. Admittedly, this observation doesn't address the puzzlement I gave voice to in the last paragraph. It's merely intended to discourage one from thinking that, in light of the puzzlement, explanation really has no home in mathematics.[12]

A mathematical explanation is typically a proof,[13] but not just any proof qualifies. Highly disjunctive proofs, such as the one constructed by Appel and Haken, don't count. An explanatory proof subsumes apparently dissimilar mathematical truths under common principles or definitions. It *unifies*. Unification involves generality; the

[11]Quoted in Jamie Tappenden (2008a, pp. 260-261).

[12]Frege explicitly acknowledged the importance of explanation in mathematics. In his *Grundlagen* Frege wrote: "The aim of proof is ...not merely to place the truth of a proposition beyond all doubt, but also to afford us insight into the dependence of truths upon one another." "After we have convinced ourselves that a boulder is immovable, by trying unsuccessfully to move it, there remains the further question, what is it that supports it so securely?" (§2) According to Tyler Burge's Frege, there's a natural ordering of mathematical truths from more to less fundamental, and explanatory proofs should reflect this objective order (Burge 1998). Thanks to Ian Rumfitt for the references here.

[13]Not all explanations in mathematics take the form of a proof. For example, there's an explanation of why Von Neumann ordinals (not Zermelo ordinals) correctly represent the natural numbers: "The principle generating the finite Von Neumann ordinals generalizes naturally to infinite numbers while the one generating Zermelo's ordinals doesn't" (Tappenden 2008b, p. 285). But this explanation isn't a proof; it's just an argument of a more general kind.

weakest assumptions necessary for proving one case should suffice for proving other cases. Prime examples of mathematical unification are proofs that demonstrate a special case, and that can then be used to demonstrate the corresponding generalization by simply discharging various quantifier restrictions. Marc Lange (2010, pp. 310 & 320) discusses one such example:

(7) $1 + 2 + 3 + \ldots + (n - 1) + n = (n + 1)n/2$

whether $n$ is even or odd. He contrasts a proof that unifies the two separate cases—one in which $n$ is assumed to be even, the other in which it's assumed to be odd—with a proof that doesn't. The difference between them is that, in the former, no new assumptions have to be made in order to prove the theorem for odd (or, as the case may be, even) $n$; one simply has to drop the quantifier restriction to even (odd) $n$. So embedded within an explanation of the one case we have (the makings of) an explanation of the other, and thus of the theorem itself. Not so in the latter proof; it's merely a conjunction of two separate arguments, and it involves more than is needed to demonstrate each case on its own. When you dispense with what isn't needed to prove (7) for even (odd) $n$, you're unable to then go on and prove (7) for odd (even) $n$. This illustrates something that I take to be constitutive of unified explanation. Let $\Phi$ be a general or derivative fact that supervenes on some set of particular instances or comparatively non-derivative/basic facts, $\phi_1$, $\phi_2$, etc. Then:

(PRINCIPLE ONE) If E is a unified explanation of $\Phi$, then no part of E is dispensable in explaining $\phi_1$, $\phi_2$, etc. individually.

I will rely on this principle in the next section.

"Unnatural" proofs—those that rely on "incorrect"/"artifical" definitions or "remote" assumptions—are also likely to be unexplanatory. The reason why Gauss and others were dissatisfied with the available proofs of quadratic reciprocity was that they involved artificial definitions, and the proofs were disappointingly indirect. Gauss wrote, "Although these proofs leave nothing to be desired as regards rigor, they are derived from sources much too remote ... I do not hesitate to say that until now a *natural* proof has not been produced."[14]

What, exactly, is an "incorrect" or "artificial" definition? Surely we're free to stipulate any meaning we like for a (meaningless) term. Why, then, is it appropriate to speak

[14]Quoted in Tappenden (2008a, p. 261, fn. 10).

here in terms of correctness and incorrectness? Furthermore, since *we* are the ones who stipulate definitions, aren't they always artificial?

Often, when a theorist defines a term, she does so intending for the definition to allow her to achieve specific theoretical goals: to formulate various hyptheses, make various projections, ask various questions, etc. Perhaps the definition succeeds in this respect. But as inquiry advances, the theorist might come to have new goals, and she might come to think the notion she latched onto with her old definition, and that satisfied her earlier purposes, should also satisfy her newer purposes. This belief might be accompanied by the unfortunate recognition that her original definition didn't latch onto a notion that *can* accomplish all she would now like it to. The theorist might then propose a new definition for her term. Alternatively, she might come to regard what was once considered a substantive principle—formulated with the help of her term—as constitutive of the term's meaning. Whichever route she takes, the theorist is likely to view her original definition as an abortive effort to come into cognitive contact with the notion that's only now available to her. If the theorist is altogether unable to define a notion that satisfies both purposes (old and new), then she will have to make do with distinct notions, both of which are artificial in light of her conviction that a single concept should be doing double duty. Let's call the kind of theory change that takes place when a theorist *is* able to (re)define a notion that satisfies all of her theoretical goals, *semantic revision*.

Hilary Putnam (1962) famously suggested that semantic revision occurred when physics abandoned the Newtonian definition of 'kinetic energy' to embrace the relativistic definition. Furthermore, the revision was responsive to empirical considerations. Putnam concluded that the analytic/synthetic distinction, as it was traditionally understood, is highly problematic. More recently, Jamie Tappenden (2008a, pp. 267-268) suggests that something very similar occurred in mathematics with the definition of 'prime number'. The standard definition for this predicate is: $n \neq 1$ is a prime number if it is evenly divided by only 1 and $n$. An alternative characterization, equivalent over the natural numbers, is: $n \neq 1$ is a prime number if, whenever $n$ divides a product, $ab$, then $n$ divides $a$ or it divides $b$. This alternative characterization was originally a theorem—a substantive result—of elementary number theory. When applied to complex numbers, however, the two ways of characterizing 'prime number' are inequivalent. And, as it turns out, the alternative characterization, not the standard definition, is correct: it better satisfies the theoretical purposes that drive number theory. In fact,

Tappenden (2008b, p. 278) observes that, from the perspective we occupy when we take the alternative characterization to be definitional, the standard definition is "accidental". More precisely, it's an accident that the standard definition coincides with the alternative characterization when both are restricted to the natural numbers.

What does Tappenden's example have in common with our earlier examples of mathematical accidents? Well, it also involves a kind of explanatory disunity: the standard definition isn't sufficiently general; when we drop the restriction over the natural numbers, it doesn't generalize in a theoretically satisfying way, while the alternative characterization does. The generality of the alternative characterization explains why (restricting $n$ to the natural numbers) $n \neq 1$ is a prime number if, whenever $n$ divides a product, $ab$, then $n$ divides $a$ or it divides $b$. In contrast, the standard definition lacks an explanation, or principled basis.

In light of all this, we can reasonably hypothesize that it's a mathematical coincidence/ accident that $p$ iff there's no unified explanation of the fact that $p$.[15] The first four examples we looked at—that a pattern of four digits should repeat itself so soon in the decimal expansion of $e$; that there are just four factorions; that (on average) the same term occurs once every ten digits in the decimal representations of $\pi$ and $e$; and that there are values of $n > 2$ such that Fermat comes damn close to falsification—lack explanations. *A fortiori*, they lack unified explanations. Our hypothesis correctly predicts that they're coincidences. The Four Color Theorem lacks a unified explanation. So it, too, is a coincidence. (3), on the other hand, isn't a coincidence because it has a unified explanation.

Before we proceed, I want to discuss one more example. In conversation, Stephen

[15]Alan Baker (2009) and Marc Lange (2010) recommend similar accounts. This project began when I came across Baker's article. Halfway through I encountered Lange's. My account expands on their work by incorporating the idea of naturalness in the notion of mathematical explanation, thus accommodating judgments about the accidental correctness of some definitions. My account also regiments talk of coincidence in a different way. Whereas Baker and Lange take coincidences to be conjunctive facts, I don't. Consider again the passage I quoted under my first example. The fact we identified there as a coincidence isn't conjunctive. What the author takes to be a coincidence is the fact "that a pattern of four digits should repeat so soon" in the decimal expansion of $e$. One can gerrymander the case to make it fit Baker and Lange's way of talking. One can say it's a coincidence that the third through sixth digits of $e$ are '1828' and the seventh through tenth digits are '1828', but it's far more natural to express the thought in the way the author actually expressed it.

I would like to thank Marc Lange for an informative e-mail exchange about his work and an earlier draft of this paper.

Yablo asked whether Euler's identity—

$$e^{i\pi} = -1$$

—is an accident. It can seem quite surprising that the reference-fixing stipulations governing '$e$', '$\pi$', and '$i$' are such as to determine values that, when related by exponentiation and multiplication, are equivalent to -1. After all, the stipulations were outgrowths of distinct branches of mathematics (analysis, geometry, algebra), and motivated by (responsive to) different kinds of considerations. One might naturally wonder, then, how the formula expressing Euler's identity can manage to express a truth. Perhaps someone might express her puzzlement by semantically descending: how could transcendental and imaginary numbers, related as $e$, $\pi$, and $i$ are, be identical to (nothing other than) -1? The properties singled out by the reference-fixing stipulations governing '$e$', '$\pi$', and '$i$' are, of course, necessary properties of $e$, $\pi$, and $i$. So the puzzlement here may just as well be expressed as follows: *how could one and the same quantity be -1 and a complex mathematical relationship between quantities with such and such necessary properties?* Benjamin Peirce (a leading nineteenth century mathematician/philosopher at Harvard, and father of C.S. Peirce) is reported to have been puzzled by just such a question. Interestingly, he was well aware of the proof that demonstrated the truth of Euler's identity (Eli Maor 1998, p. 160). But the consensus among experts today is that proofs of Euler's identity are fully explanatory. In fact, it was Gauss's opinion that if Euler's identity didn't immediately strike an advanced student of mathematics as obviously true upon first encountering it, he would never become a first-rate mathematician (John Derbyshire 2004, p. 202).

There are no lingering questions about how Euler's identity could be true, so I don't want to say that it's an accident. But I do want to say that if (*per impossibile*) the questions above had no answers—if Peirce, not Gauss, were right—then Euler's identity would be a coincidence, because it wouldn't have (though it would surely call for) a specific sort of explanation.

### 3.2.3 Empirical coincidences

Hallmarks of explanation in mathematics are also hallmarks of explanation in other branches of inquiry. Disjunctive empirical theories are typically not that explanatory, nor are theories that rely on gruesomely gerrymandered predicates and classification schemes, nor are theories that rely on remote causes. One naturally wonders, then,

whether our judgments about empirical accidents/coincidences are also sensitive to the presence or absence of unified explanations. I think they are. Indeed, the absence of a unified explanation for an empirical fact or occurrence typically suffices for its being a coincidence. I'll rely on this point a little later, so it deserves emphasis.

(PRINCIPLE TWO) If there's no unified explanation of the fact that $p$, then it's just a coincidence (merely an accident) that $p$.

Coincidentally, I share a birthday with Russell Crowe, the Academy Award-winning star of *Gladiator*. An explanation of the fact that Crowe was born on April 7th combined with an explanation of the fact that I was born on April 7th doesn't provide a *unified* explanation of the fact that there's a day on which both he and I were born. Granted, if you traced the sequence of events that led to our births far enough back in time, they may converge on a temporally remote common cause. But citing such an event typically doesn't qualify as the relevant kind of cause. In most ordinary contexts, it would be quite odd to say that Crowe and I share a birthday because of something that happened so far in the past. It may well be true, but in ordinary contexts it's not explanatory. That's why it's often odd to say such things.

PRINCIPLE TWO coheres nicely with an old observation in H.L.A. Hart and Tony Honoré's classic book about *Causation in the Law*. "We speak of coincidence whenever the conjunction of two or more events in certain spatial or temporal relations ... occur without human contrivance and are independent of each other" (1959, p. 74). In this context, 'independent' means *causally independent*. And one should expect the co-occurrence of causally independent events to lack a unified explanation.

I want to cancel any suggestion that the notion of (mathematical) explanation is conceptually prior (whatever that means) to the notion of (mathematical) coincidence/accidentality. I haven't tried to analyze the concept of mathematical coincidence. What I've tried to do in this section is motivate the first part of the error theory I sketched in the introduction: some necessities—namely, mathematical truths—are accidents/coincidences. I also provided an account of what makes a particular mathematical truth an accident, or coincidence. I then suggested that the sort of condition that renders a mathematical truth accidental suffices to make empirical phenomena accidental.

Some philosophers, both in print and in conversation, are resistant to the claim that there are mathematical coincidences/accidents on the grounds that coincidence/accidentality involves contingency (Michael Potter 1993). Perhaps the reader sympathizes.

There's definitely something to this worry, and I will have more to say about it, but only after my view is fully spelled out.

## 3.3 Psychophysical coincidence

In this section I'm going to motivate the second part of my error theory: the epistemological considerations that traditionally support dualism merely show that pain and c-fiber firing are accidentally, or coincidentally, related, not that it's metaphysically possible for the one to occur without the other. Some readers might anticipate how this is going to go. I will show that explanations of psychophysical occurrences are disunified, and that efforts to unify them generate a notorious "explanatory gap". More generally, the strategy here is to analogize mathematical coincidences and psychophysical phenomena in the light of our much discussed and suboptimal epistemic predicament *vis-à-vis* the latter.

Suppose I prick my finger with a pin. C-fibers begin to fire and pain is immediately felt. We have a pretty good account of why each singular event occurred. C-fibers (as opposed to, say, d-fibers) began to fire in response to my punctured skin because a cascade of specific biochemical signals was triggered, resulting in neurophysiological effects that culminated in the firing of c-fibers. The pin prick was painful because it punctured my skin, and pain is supposed to alert us to a damaged body; that's its function. This naive teleological account can be unpacked in terms of evolutionary biology. Felt pain generates and reinforces an aversion to conditions which produce the sensation. Animals with an aversion to damaged skin are more likely to survive and reproduce than animals that aren't similarly averse. So animals that feel pain in response to damaged skin are more fit in the evolutionary sense. The capacity to feel pain has a genetic basis, so it can be passed from generation to generation. Therefore, one should expect animals to evolve the capacity to feel pain when skin is damaged (as opposed to some other or no sensation).[16]

But although we have a pretty good account of why each singular event occurred, we don't thereby have a *unified* account of why there's something it's like for my c-fibers to fire—that is, of why the realization of a phenomenal state or property should

[16] I don't expect there to be an adaptationist story for very many sensations, but physicalists should think that some kind of evolutionary story (in terms of genetic drift or exaptation or whatever) applies to them all.

be correlated with the firing of my c-fibers. Perhaps the two separate theories, when combined, go some way toward answering this further question, but the resulting conjunctive theory doesn't qualify as a unified explanation. Recall

(PRINCIPLE ONE) If E is a unified explanation of $\Phi$, then no part of E is dispensable in explaining $\phi_1$, $\phi_2$, etc. individually.

Let $\Phi$ be the derivative fact that there's something it's like for one's c-fibers to fire; let $\phi_1$ and $\phi_2$ be the comparatively basic facts that one's c-fibers fire when such and such conditions are met and that one feels pain when such and such conditions are met. (Both occurrences of "such and such..." are place-holders for a description of the very same conditions.) The conjunctive theory doesn't satisfy the consequent of PRINCIPLE ONE, because the teleological/evolutionary component is dispensable in explaining $\phi_1$ and the neurophysiological component is dispensable in explaining $\phi_2$. Omitting the superfluous component in the explanation of $\phi_i$ renders the theory incapable of explaining the other comparatively basic fact. Thus PRINCIPLE ONE implies that the conjunctive theory isn't a unified explanation of $\Phi$. The situation resembles a case we looked at earlier: the disunified proof of (7). So if there weren't any better explanation of the fact that there's something it's like for one's c-fibers to fire, it would follow by PRINCIPLE TWO that it's a coincidence (merely an accident) that there's something it's like for one's c-fibers to fire.

But perhaps there *is* a better explanation in the offing. Ned Block and Robert Stalnaker rightly observe that "Identities allow a transfer of explanatory and causal force not allowed by mere correlations. Assuming that heat = [mean kinetic energy], that pressure = molecular momentum transfer, etc. allows us to explain facts that we could not otherwise explain" (1998, p. 24). More specifically, positing such identities allows our physical theories—formulated with the help of a vocabulary within which folk terms like 'heat', 'pressure', and 'boiling' have no place—to explain not merely why things *correlated* with heat and pressure jointly cause something *correlated* with boiling, but why heat and pressure *together* cause boiling. The identities close the door on a question left open by our physical theories: what causes boiling? Note further that they close the door on other questions: why should facts about heat, for example, be correlated with facts about mean kinetic energy? Perhaps, then, if we posit a psychophysical or functional identity in the case that interests us here, we will be able to close the door on the question left open by the separate neurophysiological and teleological theories

we looked at earlier—why should a phenomenal state or property be correlated with my c-fibers firing—and thus have a perfectly unified explanation of the relevant fact.

Let pain be a specific functional state: the (second-order) state of being a (first-order) state that tends to cause such and such other (first-order) states or properties (wincing, wanting relief, etc.), and that tends to be caused by such and such *other* (first-order) states or properties (punctured skin, etc.). And let c-fiber firing be the neurophysiological state or property that realizes this complex functional state in animals like us. Or, alternatively, we could simply take pain to be identical to the firing of c-fibers. Whichever reduction we prefer, the question, why should a phenomenal state or property be correlated with the firing of my c-fibers (why is there something it's like for my c-fibers to fire), appears to be answered. A neurophysiological account of why c-fibers fire in response to pin pricks just is an account of why pain is felt in such conditions. Similarly, a teleological/evolutionary account of why I felt pain just is an account of why my c-fibers fired. Positing the identity seems to provide us with a perfectly unified explanation. So what's the problem?

The problem is that the identity we've posited gives rise to an "explanatory gap". Authors disagree about how to precisely characterize the gap, but they agree in thinking that psychophysical/functional identities are, in an important respect, unlike 'heat = mean kinetic energy', 'pressure = molecular momentum transfer', etc. For my purpose, it doesn't really matter which of the competing characterizations is correct; however one chooses to characterize our impoverished epistemic situation with respect to psychophysical/functional reductions, it will follow that in some way or other they are left unexplained. So they are coincidentally, or accidentally, true. Let me elaborate.

Some authors take the explanatory gap to be a lingering why-question: assuming that psychophyscial/functional identities do indeed hold, it makes sense to ask *why* they do (Joseph Levine 1983, 1993; Colin McGinn 1991; Thomas Nagel 1974). Why is c-fiber firing the neural basis of, or identical with, *this* rather than *that* phenomenal state or property? Similar questions can't, according to these authors, be asked about, say, heat and its physical basis. Assuming that heat = mean kinetic energy, it makes no sense to ask why the former should be identical to the latter. Let's call reductions of the first sort, which give rise to the kind of why-questions identified a moment ago, *gappy*. (The term is Levine's.) What accounts for the difference between gappy and non-gappy reductions? One answer (due to Levine, Chalmers and Jackson) begins with the observation that non-gappy reductions are justified in part by

conceptual analyses of the higher-level phenomenon: heat, pressure, boiling, water, or whatever it may be. Our grounds for accepting them can be factored into a purely conceptual or *a priori* component, and a purely empirical one. Conceptual analysis reveals that heat is (for lack of a better way of putting the point) the heatish stuff. 'Heatish' abbreviates some lengthy causal/structural description that *exhaustively* unpacks the folk notion of heat. Microphysics then informs us that the heatish stuff is mean kinetic energy. So we can reasonably conclude that heat is mean kinetic energy. Now it may well be that if we analyze the concept of pain, we'll be able to specify a complex causal/structural/functional description that aptly characterizes the role pain plays in our lives. But, unlike the analysis of heat, this description wouldn't (according to the authors I've mentioned) exhaustively unpack our notion of pain. Some part of that notion would be left unaccounted for: *viz.*, qualitative character, or the *what-it's-like* component of pain, which is constitutive of the way we represent pain in thought. In general, causal/structural/functional descriptions of qualitative states/properties don't *a priori* imply what it's like to be in that state or exemplify that property. (Frank Jackson's Mary illustrates this.) So, to borrow an age-old metaphor, the concept of what it's like to be in pain mustn't be part of the concept expressed by the appropriate sort of causal/structural description of pain. Consequently, the causal/structural/functional analysis of pain is incomplete. It's because our concepts of heat, pressure, boiling, water, etc. don't incorporate a qualitative, what-it's-like element that we can identify their referents with physical phenomena without any further questions calling out for answers. If one accepts this way of characterizing our sub-optimal epistemic situation with respect to psychophysical/functional identities, then it follows by PRINCIPLE TWO that such identities are accidents, or coincidences, because they have no explanation.

But almost every major claim in the story I just summarized about why psychophysical/functional identities are gappy, whereas other identities aren't, has been challenged.[17] I want to voice an additional worry that seems not to have been registered in the back and forth on this matter. Assuming for the sake of argument that reductions generally proceed by way of conceptual analysis, why think that purely causal/structural descriptions of non-phenomenal entities (heat and light, say) *exhaustively* unpack the corresponding folk notion (HEAT and LIGHT)? Isn't information about the phenomenology of heat or light (how it feels or looks) part of HEAT/LIGHT? Kripke, in his typically cautious way, expressed some sympathy for thinking so.

[17]See Block and Stalnaker (1998) and Block (2002) for objections. See Jackson and Chalmers (2001) for replies.

> ...we identify heat and are able to sense it by the fact that it produces in us a sensation of heat. It might be so important to the concept that its reference is fixed in this way, that if someone else [an alien, say] detects heat by some sort of instrument, but is unable to feel it, we might want to say, if we like, that the concept of heat is not the same even though the referent is the same (1980, p. 131).
>
> ...a blind man who uses the term 'light', even though he uses it as a rigid designator for the very same phenomenon as we, seems to us to have lost a great deal, perhaps enough for us to declare that he has a different concept (1980, p. 139).

And if phenomenal information—information about how something feels or looks—*is* part of the corresponding folk notion, then the account of gappy identities over-generates: it predicts that heat = mean kinetic energy should also be gappy. Let me be clear: my point isn't that non-phenomenal identities can also be gappy; it's that if there *are* gappy identities, then the difference between them and non-gappy identities can't consist in our capacity to reduce by means of conceptual analysis the higher-level phenomenon to some microphysical phenomenon. Even putatively non-gappy reductions don't work by way of conceptual analysis.

Critics of the view that reduction proceeds by way of conceptual analysis don't think there is a lingering why-question; they think, rather, that psychophysical/functional identities generate a lingering how-question. Let Q be the phenomenal quality associated with a particular visual experience, and let's assume (following Francis Crick and Christof Koch (1990)) that cortico-thalamic oscillation is Q's neural correlate. Now even if we suppose that Q = cortico-thalamic oscillation, "We have no idea *how* it could be that one property could be identical both to Q and cortico-thalamic oscillation. How could one property be both subjective and objective? ...This is what we need in the case of subjective/objective identities" (Block 2002, p. 395, emphasis added). "...it remains puzzling, in the pain/brain-state case, *how* one and the same thing can have the two different kinds of essential properties [subjectivity/objectivity]" (Edgington 2004, p. 18, emphasis mine). Even according to this alternative way of characterizing our impoverished epistemic situation *vis-à-vis* psychophysical identities, something about them is left unexplained. The situation here is quite similar to Peirce's bewilderment concerning Euler's identity. Recall: how could one and the same quantity be -1 and a complex mathematical relationship between quantities with such and such necessary

properties? The only relevant difference is that there really is a question left unanswered in the psychophysical case. And without an explanatory answer, PRINCIPLE TWO implies that psychophysical identities are just accidents/coincidences.

This concludes the second part of my error theory. In the next section I will elaborate on the third and final part of the account: that we equivocate between metaphysical contingency on the one hand and accidentality/coincidence on the other.[18] It seems as though there could have been zombies because it's a coincidence that there aren't and we typically don't distinguish coincidences from contingencies. Admittedly, I'm not the first to suggest that the appearance of contingency somehow arises from our epistemological predicament in relation to psychophysical/functional identities. To my knowledge, Joseph Levine (1983, 1993) was. What the discussion above demonstrates is that one can accept Levine's point about the source of the contingency intuition without committing oneself to his questionable assumptions about the nature of reductive explanation (by means of conceptual analysis) and the distinction between gappy and non-gappy identities. It's enough that psychophysical/functional identities leave us wondering *how* they could be true for us to correctly judge that they are coincidences.

## 3.4 Semantic encapsulation

Consider the title of this essay. It certainly sounds contradictory to the ordinary reader. That's partly why I chose it; I thought it might provoke interest. But it also serves to illustrate an important point. Why *should* it sound contradictory? Answer: often, we're not sensitive to the distinction between coincidence and genuine metaphysical contingency. But this can't be the whole story, because the title also sounds contradictory to those of us who acknowledge the distinction. You and I, for example, know that there are necessary coincidences, and that coincidentality consists in the absence of a unified explanation, not in the fact that some (but not all) ways the world might have been are such that .... And yet my title has a contradictory flavor to it. Why? The answer must be that we conflate the two notions at a level of cognition that can't be quickly penetrated by conscious rational judgment. I want to further motivate this explanation by looking at a series of analogies that come closer and closer to the case at hand.

The Müller-Lyer illusion involves two lines of equal length, but which appear unequal due to inverted arrows at the ends of each line. The illusion persists even for the

[18]The *Oxford English Dictionary* lists 'accidental' and 'incidental' as synonyms of 'contingent'.

enlightened, that is, even for individuals who know the lines are the same length. The reason is that visual perception is modular and, therefore, encapsulated from judgment centers of the mind (Jerry Fodor 1989). This means the 'wetware' involved in vision does not have access to the information we possess about the actual lengths of the lines. The optical illusion is thus screened off from (and, hence, can't be corrected by) rational judgment.

The kind of encapsulation, or informational screening off, involved in the Müller-Lyer phenomenon isn't restricted to optical illusions; something very like it is present in other cognitive domains. For example, consider the way people react to the 'Monty Hall problem' after being told why the correct answer is *yes*,[19] or consider how people typically feel about the Berkeley sex bias case even after they verify it by running through the statistics.[20] In both cases the incorrect intuitive answer/feeling continues to appear correct. Each example illustrates that cognitive appearances are sometimes screened off from our rational judgments. So, just as with visual illusions, cognitive illusions persist even for the enlightened among us.

Not too long ago I used to think (and I'm embarrassed to admit this, but it serves my point too well to keep it hidden) that 'disinterested' and 'uninterested' are synonyms, and that 'livid' meant red with anger. (It actually means bluish grey.) Even after my semantic beliefs were consciously corrected, I was disposed to use 'disinterested' in contexts where 'uninterested' was called for. Uses of the word 'livid' continued to be automatically interpreted as meaning red with anger. It took a concentrated effort to avoid such mistakes. My revised semantic beliefs didn't immediately register at a deep dispositional level. (It was a little bit like learning a second language, but on a much, *much* smaller scale.) I'm happy to report that my efforts succeeded. But one

[19] "Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice?" (Craig Whitaker 1990, p. 16)

[20] The University of California, Berkeley was sued for male bias in graduate admissions. The admissions figures for 1973 showed that male applicants were more likely than female applicants to be accepted. The gap between male and female acceptances was so large that many assumed it couldn't have been due to chance. When statistics from individual departments were gathered it was discovered that no department had a non-random bias in favor of males. In fact, most departments had a statistically significant bias in favor of females (David Freedman *et al.* 1998, p. 19). The example is an instance of "Simpson's paradox": a statistical trend present in each individual case which is reversed when the data from each case is aggregated.

wonders why traces of the semantic error lingered after my beliefs had been revised. Hypothesis: semantic processing is to some degree encapsulated from one's faculty of conscious judgment. There appears to be some independent evidence for thinking so. Meaning retrieval is a highly rapid, mandatory or automatic occurrence that speakers/interpreters take advantage of to perform a relatively narrow range of tasks. According to Fodor (1983), these are hallmarks of encapsulated processing. It's useful, then, to contrast two different kinds of equivocation: *shallow* and *deep*. The former is exemplified by my incorrect semantic belief that 'disinterested' and 'uninterested' are synonyms; the latter is exemplified by my disposition to incorrectly use (and interpret uses of) 'disinterested'/'livid' even after I revised my semantic beliefs.

I want to suggest that the semantic 'capsule' that embodies one's competence with 'coincidence' and 'contingency' is screened off from one's enlightened judgment that the terms are semantically inequivalent, much like how my semantic processor's representation of the meaning of 'disinterested' and 'livid' were largely unaffected by conscious belief revision. In other words, by acknowledging that some necessities are (or can be) coincidences we thereby correct the error of shallow equivocation, but deep equivocation persists. This, I propose, is the reason why my title may sound contradictory to even those of us who overtly acknowledge the distinction between coincidence and contingency, and it further explains why the zombie intuition is as hard to shake as it is. At a deep dispositional level, the judgment that, coincidentally, there aren't zombies is registered as, contingently, there aren't zombies, which contradicts the statement that there couldn't have been zombies. With enough time, and a concentrated effort, the encapsulation can be overcome. In this respect, cognitive illusions brought about by semantic encapsulation differ from optical illusions. In my own case, the zombie intuition is significantly less compelling than it once was. I'm confident that over time it will continue to lose its force. For soft-nosed physicalists, coincidence is contingency enough.

## 3.5 Objections, replies, and closing remarks

I anticipate a number of objections. In this section I want to summarize them and explain why I think they're unconvincing.

One might worry that my account is self-undermining. The worry might be put as follows. "You say it appears contingent to the physicalist that pain accompanies c-fiber

firing (partly) because she correctly judges that it's a coincidence that pain accompanies c-fiber firing. So your account relies on the principle below.

(PRINCIPLE THREE) If one judges it to be a coincidence/accident that $p$, then it will appear contingent that $p$.

But this principle seems to predict that mathematical coincidences should appear contingent. Yet they don't. Mathematical propositions are transparently necessary. So the very thing you rely on to motivate part of your story actually falsifies it! How embarrassing for you!"

I think this apparently critical observation really points to something that further supports the story I've been telling. Earlier, if you recall, I observed that some philosophers, both in print and in conversation, want to resist the idea that there are mathematical coincidences. Michael Potter, for example, writes, "The image of mathematical sentences being true by accident is an arresting one. It is plainly repugnant ..." (1993, p. 308). The source of this repugnance is the transparent necessity of mathematical propositions. As Potter explains, "God simply does not ... have the freedom to decide their truth or falsity" (*ibid.*). It's precisely because candidate mathematical truths don't appear contingent that Potter and others refuse to judge that they're coincidences. This provides indirect support for PRINCIPLE THREE by directly supporting its contraposition. Philosophers take for granted that mathematical propositions are necessary and lose sight of the fact that they can be accidental. Similarly, in the psychophysical case, many of us take for granted that the correlation between phenomenal and neurophysiological occurrences is a coincidence and, consequently, we feel pressure to say that c-fibers could have fired without pain being felt. In both cases, however, the mistake is accompanied by—and, I claim, the result of—shallow equivocation. The reason why mathematical coincidences don't appear contingent to you, me and other 'enlightened' folk is that, unlike Potter, we don't (shallowly) conflate coincidence and contingency.

One might put pressure on the semantic part of my account as follows: "If the 'unenlightened' typically equivocate, then why isn't the discourse fragment below contradictory-sounding?

(8) It's contingent but not accidental that Arsenal won the match. After all, on paper, Arsenal was the better team.

By your lights, shouldn't equivocators interpret (8) as follows:

(9) It's contingent but not contingent that Arsenal won the match. After all, on paper, Arsenal was the better team.

Again it seems your proposal gets the facts wrong."

(8) certainly isn't as bad as (9). In fact, I think it sounds ok. But my informants tell me that (8) would sound really bad if the follow-up sentence—'After all, ..."—were omitted. By appealing to the standard analysis of modals in linguistics, due in large part to Angelika Kratzer (1981), I think we can explain the data here consistent with my suggestion that ordinary interpreters shallowly equivocate.

According to Kratzer, modals are doubly context-dependent quantifiers over possible worlds. The context supplies information that (a) restricts the space of worlds in the domain of quantification (called the "modal base") and (b) determines an ordering of worlds in the restricted domain. Distinct modals within a single sentence can be assigned different modal bases. (E.g., "Goldbach's conjecture might turn out to be false, but that would be surprising.") *Given that ordinary interpreters typically treat 'accidental' as semantically equivalent to 'contingent', one would expect that the interpretation of 'accidental' in (8) is sensitive to the same parameters that the interpretation of 'contingent' is sensitive to.* But the parameters must be resolved by different bits of contextually salient information. The modal base of 'accidental' is (while the modal base of 'contingent' isn't) partly determined by the follow-up sentence. This explains why, when the follow-up sentence is absent, interpreters judge (8) to be semantically marked, because the coherence of (8) critically depends on a shifted quantifier domain. The modal base of 'It's contingent that' is a relatively expansive set of possibilities, in some of which the comparative talent of Arsenal and its rival differs from the way it actually is: the rival is about as good as, or better than, Arsenal. The modal base of 'accidental', in contrast, is restricted to a proper subset of that domain where, in each world, Arsenal is better than its rival. If we let W be the modal base associated with 'It's contingent that', then, assuming as I have that ordinary interpreters equivocate, (8) should be interpreted as (8*).

(8*) In some but not all of the worlds $w \in W$, and in every $w$ where Arsenal is better on paper than its rival, Arsenal won the match.

And this is perfectly coherent. But now one wants to know why (9) is ineligible for this sort of interpretation. Why can't it be assigned a coherent reading? The answer is that the use of 'accidental' in (8) is the speaker's way of lexically signaling that a different

modal base is now operative. The infelicity of (9) is due to the fact that such a signal is required but missing.

Even if my account successfully identifies the source of our temptation to say psychophysical relations are contingent, one might still worry that it offers us less than we might have hoped for. The question many philosophers of mind have been concerned with since Kripke is: if zombies aren't metaphysically possible, then what is the real content of the mental act which so strongly suggests to us that they are? The urgency of this question stems partly from the pressure we physicalists feel to diagnose the source of the zombie intuition and to ultimately massage it away. Getting rid of this pressure is the *dialectical challenge*. The urgency of the question also partly stems from the idea that there has got to be some interesting, though perhaps convoluted, epistemic connection between the kind of mental act we perform when we seem to imagine a zombie world on the one hand and genuine metaphysical possibility on the other. How else can we come to know that, for example, Sarah Palin might have been vice-president? Without a story about what the real content of our mental act is—a story that flows naturally from the way we explain away other instances of illusory contingency (the correlation between water and $H_2O$, heat and molecular motion, etc.)—we will not be able to clearly spell out the connection between imagination and possibility. Telling such a story is the *epistemological challenge*.

Typically, philosophers try to resolve both challenges at once. They begin by offering a (sketch of a) theory about how imagination (or conception) and metaphysical possibility are epistemically connected, then they apply the theory to the dialectical challenge.[21] My account exemplifies a different strategy: rather than starting with a general theory about the relationship between imagination and possibility which is then applied to exorcise the zombie intuition, I try to diagnose the source of the zombie intuition while keeping quiet about the epistemological challenge. But habit sometimes breeds the expectation that things ought to go as they typically do. So some philosophers might think that since the account I have offered is silent about how imagination and possibility are epistemically connected, it isn't as informative as it ought to be. But from a dialectical point of view, this is one of its strengths. There is little agreement about the epistemic connection between imagination and metaphysical possibility.[22] It would be nice if one could diagnose the source of the modal illusion and massage it

[21]For examples of this strategy, I refer the reader to Kripke (1980), Thomas Nagel (1984), and Robert Stalnaker (2002).

[22]See David Chalmers (2002), Stalnaker (2002), and Stephen Yablo (2000, 2008).

away without being forced to take a stand on the epistemological controversy, which rapidly escalates into larger disagreements about whether semantic content is multi-dimensional, and how robust the *a priori*/empirical distinction is. Doing so would bypass worries ultimately stemming from a prior commitment to an alternative semantic and epistemological framework.

What are we to make of all this? Well, it would certainly be nice if we physicalists had a modal epistemology that could eliminate the zombie intuition in the way that Kripke's theory promised (but failed) to dissolve modal illusions generally—an epistemology that was independently motivated and that could convince the unconvinced. But sadly the prospects for such a theory are grim, and the problems aren't restricted to the mind-brain case. The best we can hope for (here and elsewhere) is a stable defense. And that, I think, we have. Like any good error theory, mine identifies the kernel of truth in the interlocutor's objection. There is something very like contingency that is true of the relationship between pain and c-fiber firing. So much so that one could hardly be blamed for confusing this property with its metaphysical cousin. But even if the difference is easy to overlook, it can't be denied. And it's precisely because of this difference that the physicalist needn't be scared of zombies.

CHAPTER 4

# Zombies are Inconceivable

## 4.1 Opening remarks

Zombies present a challenge to physicalism, just as super Spartans present a challenge to behaviorism. The challenges are quite similar.[1] One seems able to conceive that there are creatures of a particular sort: in the former case, physical duplicates of actual human beings who inhabit a world with the same physical laws but who lack phenomenal consciousness; in the latter case, human beings who never exhibit pain behavior, even when they're experiencing intense pain. More specifically, each challenge rests on the premise that no amount of *a priori* reflection—not even *a priori* reflection under optimal cognitive conditions—will uncover a latent inconsistency in the supposition that there are such creatures. Furthermore, both challenges rely on the premise that there's a reliable epistemological connection between what one can conceive and what's genuinely possible. In light of Saul Kripke's (1980) highly compelling examples of empirical necessities and *a priori* contingencies, this connection is rather difficult to articulate. But if conceivability and possibility aren't reliably connected, then one wonders how we could have the rich body of modal knowledge we ordinarily take ourselves to have.[2] We thus appear to have reliable evidence that zombies and super Spartans could have existed, contrary to physicalism and behaviorism.

I would wager that, nowadays, almost all (if not all) physicalists think that the super Spartan argument is rather convincing. It's surprising, then, that the most common

[1] The similarity, and its significance, was recently observed by Daniel Stoljar (2005, 2006).

[2] For illuminating discussions that try to articulate the connection between conceivability and possibility, see Stephen Yablo (1993, 2002) and David Chalmers (2002).

objection to the zombie argument targets the epistemological relationship between conceivability and possibility. Typically, physicalists criticize the zombie argument on the grounds that, in this particular case, conceivability leads us astray. But why should the reliability of our conceivings suddenly breakdown when we stop thinking about super Spartans and start conceiving of zombies?[3] This question has received very little attention.

A satisfactory response to the zombie argument should identify an error that doesn't inadvertently sabotage the super Spartan argument. In this chapter, I would like to provide such a response. My strategy is very simple. I will present an argument to the effect that zombies are inconceivable. The argument relies on something distinctive about zombies, so it doesn't support the analogue claim about super Spartans. I will then address what I take to be the most challenging objections to my central argument. Along the way, I will discuss related epistemological issues about introspective awareness and phenomenal judgment. But before I plunge ahead, I want to consider two questions with more care: what is the zombie argument supposed to achieve, and what do zombies have to be in order for the argument to achieve it? A clearer understanding of the answers than my opening remarks provide is necessary for my central argument and its defense.

## 4.2 Stage setting

The zombie argument is supposed to justify three separate conclusions about the nature of mind, all of them negative. The first is rather straightforward and calls for almost no commentary.

(ZAC1) Consciousness does not metaphysically supervene on the physical.[4]

If, as we physicalists think, consciousness supervenes metaphysically on the physical, then a physical duplicate of a conscious being should also be conscious. Since you and I are conscious, but our zombie doppelgängers are supposed to be unconscious, (ZAC1) follows. Therefore, either consciousness itself is a fundamental property—metaphysically on a par with such properties as mass, spin, and charge—or it

[3]Stoljar criticizes the so-called "phenomenal concept strategy" and Robert Stalnaker's (2002) defense of type-B physicalism by asking a very similar question.

[4]Dualists happily grant that consciousness nomologically supervenes on the physical.

supervenes on fundamental "protophenomenal" properties. In any case, the upshot is that there are more fundamental kinds of property than are dreamt of in the physicalist's philosophy.

Now, for physicalists, the supposition that $x$ and $y$ are physical duplicates is going to carry a further commitment. Specifically, the supposition implies that $x$ and $y$ have the same *functional organization*. Let me elaborate.

Specifying the functional organization of a system is a three-step process. First, one identifies the explanatorily relevant components of the system. This step can be carried out in different ways, depending on what one's theoretical objectives are. Second, one describes the possible states that each component can realize. This, too, can be carried out in different ways, depending on how finely one individuates component states. The third and final step involves characterizing the system's *causal profile*: the pattern of possible causal relations between (i) the system's internal states, (ii) its environmental inputs, and (iii) its behavioral outputs. ('Behavioral output' should be understood rather broadly. Any sort of reaction to an external stimulus—whether it be a bodily movement, or a change in one's internal state—qualifies as behavioral output in my sense.) The outcome of the final step in the process is a *functional description* of the system. Two things have the same functional organization if they satisfy the same functional description, and they satisfy the same functional description if they have the same causal profile.

Now, according to physicalists, a thing's causal profile is preserved by physical duplication. If $x$ and $y$ are physical duplicates (inhabiting worlds with the same physical laws), then the pattern of possible causal relations between $x$'s internal states, its inputs, and its outputs should be the same as the pattern of possible causal relations between $y$'s internal states, inputs, and outputs. So given our definition of sameness of functional organization, it follows that physical duplication preserves functional organization. Since zombies are supposed to be physical duplicates of conscious beings, it follows that they have the same functional organization as their conscious counterparts. But since they apparently lack consciousness, (ZAC2) follows.

(ZAC2) Consciousness does not metaphysically supervene on functional organization.

This, I gather, is what David Chalmers was getting at in the passage below.[5]

> What is going on in my zombie twin? He is physically identical to me, and we may as well suppose that he is embedded in an identical environment. He will certainly be identical to me *functionally*: he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be *psychologically* identical to me ... All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions (1996, p. 95).

My argument for the inconceivability of zombies will rely heavily on the notion of sameness of functional organization. And, admittedly, difficult questions about this notion haven't, as far as I'm aware, been answered. How, for example, are we to individuate environmental inputs, or stimulus conditions? How are we to individuate behavioral outputs? Such questions aren't entirely unrelated to my central argument. Sadly, I don't have a theory to offer in response. So readers might have two preliminary worries about my project: first, that it might presuppose a questionable way of individuating reactions and stimuli so as to merely generate the illusion that zombie worlds are inconceivable; second, that we don't have a firm enough grip on the idea of functional organization to know what we're talking about on this particular occasion.

Let me respond to the second worry first. In general, our inability to provide an independent criterion of individuation for behaviors and stimulus conditions does little to challenge our conviction that we know what we're talking about when we use the term 'functional organization'. After all, functionalism in the philosophy of mind is a significant theory. I see no reason why this inability should pose a unique challenge in this context. Our intuitive understanding of sameness and difference of behavior/stimulus is robust enough to confer content on, and license, the judgments about functional organization that I hope to elicit momentarily. And that shouldn't be surprising, since our intuitive understanding of what constitutes the same behavior/stimulus subserves a wide range of uniform judgments about particular cases, even when the cases happen to be rather outlandish. For example, most of us think that China could have had the same functional organization as Ned Block.[6] This pre-

[5]Chalmers accepts (ZAC2), but thinks that consciousness nomologically supervenes on functional organization. He calls this doctrine the *principle of organizational invariance*. See his (1996, ch. 7).

[6]See Block (1978).

supposes that there's some way of individuating behaviors and stimulus conditions so as to count the inputs and outputs of China and Block as the same. No judgment about functional organization that I will rely on presupposes a more objectionable way of individuating things.

In response to the first worry I can only say this: be patient. It will become clear that I'm relying on little more than the intuitive understanding that anchors our judgments about the functional organization of different systems.

One last preemptive remark before we move on. It may be that some cognitive states constitutively involve phenomenal properties. A reasonable candidate for such a state might be the state of believing that *this* is what it's like to see red, where the referent of '*this*' is fixed by inwardly demonstrating the phenomenal character of one's visual experience. Cognitive states of this sort are often called *phenomenal beliefs*. But since zombies don't have any phenomenal properties, they might not be able to occupy such states. This, of course, puts pressure on the thought that zombie-me and I have the same functional organization, since it suggests that there are some states which I can occupy by means of introspective reflection that he can't. In order to bypass such worries, Chalmers proposes that 'judgment' be used as a neutral term designating "what is left of a belief after any associated phenomenal quality is subtracted", namely, the underlying information-carrying state that my zombie and I have in common (*Ibid.*, p. 174). It is at this level of abstraction that, according to him, my zombie and I have the same functional organization.

One might be skeptical about whether there's anything that answers to the reference-fixing description governing 'judgment', but Chalmers assures us that such skepticism is unwarranted. The reason is that

> For a start, the disposition to make verbal reports of a certain form is a psychological state; at the very least, we can use the label "judgment" for this disposition. Moreover, whenever I form a belief about my conscious experience, there are all sorts of accompanying functional processes, just as there are with any belief. These processes underlie the disposition to make verbal reports, and all sorts of other dispositions. If one believes that LSD produces bizarre color sensations, the accompanying processes may underlie a tendency to indulge in or to avoid LSD in future, and so on. We can use the term "judgment" as a coverall for the states or processes that

> play the causal role in question. At a first approximation, a system judges that a proposition is true if it tends to respond affirmatively when queried about the proposition, to behave in an appropriate manner given its other beliefs and desires, and so on (*Ibid.*).

The definition of 'judgment' that Chalmers provides in this passage is a technical stipulation. If one is resistant to the idea that zombies make judgments, then simply substitute for 'judgment' 'judgment*'. Nothing hangs on terminology. The point is simply that zombies and their conscious counterparts realize the same information-carrying states, and that *that* is what matters for sameness of functional organization.

The third and final conclusion that the zombie argument is supposed to justify is relatively straightforward.

(ZAC3) Psychophysical laws do not metaphysically supervene on physical laws.

A corollary of (ZAC1) was that either phenomenal properties or protophenomenal properties are fundamental; the distribution of such properties can't be explained in terms of more basic phenomena. And, as Chalmers writes, "Where we have new fundamental properties, we also have new fundamental laws. Here the fundamental laws will be *psychophysical* laws, specifying how phenomenal (or protophenomenal) properties depend on physical properties" (*Ibid.*, p. 127). So, for Chalmers, (ZAC3) is a consequence of (ZAC1).

The point might be put in a slightly different way. We have good reason to believe that the relationship between physical events and conscious experience is governed by law. But remember: zombies have the same physical makeup as we have *and* they inhabit a world with the same physical laws as the world we inhabit. So, if the laws governing the relationship between physical events and conscious experience supervened on physical laws, then physical events at zombie worlds should be related to conscious experience in exactly the same way that physical events at the actual world are related to conscious experience. But at zombie worlds physical events aren't related to conscious experience in the way that they are at the actual world. Some physical events here cause pain, others cause pleasure. But pain and pleasure don't exist at zombie worlds. So the relevant laws mustn't supervene on the physical laws. They must be fundamental.

Now that we have a clearer understanding of what the zombie argument is supposed to achieve, and of what zombies are supposed to be in order for the argument to achieve it, we're in a position to assess whether zombies are conceivable. I hope to convince the unconvinced that they aren't.

## 4.3 The central argument stated and defended

There's a well-known heuristic device that illustrates the difference between physicalism and dualism quite nicely. (The use of this device can be traced back to Kripke (1980).) According to the physicalist, once God determined all of the physical facts, he determined all of the facts about our world, even the facts about our mental lives (whether we're experiencers, what we're experiencing, what it's like to be experiencing it, etc.). So there was nothing left for God to do. According to the dualist, however, when God determined all of the physical facts, the facts about conscious experience were left undetermined. God had more work to do. Fortunately, he kept working in our world. But the inhabitants of other worlds—even worlds physically identical to our own—weren't so fortunate. God simply chose not to wave his consciousness-conferring wand over them.

But suppose that God comes to regret his choice and resolves to change things midstream. God decides to bestow consciousness on our zombie twins. Now imagine that he does so in a way that involves as little effort on his part as possible: he just waves his consciousness-conferring wand indiscriminately over the entire region of modal space occupied by worlds physically identical to our own. So he waves his wand over the actual world, too. Thus the causal process that God's wand-waving initiates impinges on me as well as my zombie twin. In my twin's case, it generates the intended effect. Zombie-me is *de-zombified*. In my case, the process doesn't generate that effect. I'm already conscious, after all. But despite the difference between the phenomenal effects that God's wand-waving brings about, it generates the very same stimulus in both my case and my twin's. To clarify this last remark, it may help to consider an analogy.

Imagine two machines connected by separate pathways to the same switchboard. Imagine further that two different buttons on the switchboard regulate whether the machines are on or off. You press the green button over here to turn the machines on; you press the red button over there to turn the machines off. One of the machines happens to be on already; it was connected to the switchboard while in operation. The

other machine happens to be off. But let's say you want to turn it on. So you press the green button, which sends the very same signal through both pathways. While the signal produces the intended effect in the machine that was off, it doesn't produce that effect in the machine that was on. But the same stimulus impinges on both machines. God's wand-waving, and the causal process that it initiates, is relevantly like your button-pushing, and the signal that it transmits. Despite different phenomenal or operational effects on the targets, the environmental input is the same.

The question that now deserves consideration is whether my de-zombified twin and I will make the same judgments in response to our encounters with the same stimulus. In other words, will the *cognitive* effects of the imagined stimulus be the same?[7] The background assumptions of the zombie argument—that my twin and I are physical duplicates and, therefore, that we have the same functional organization—predict that the answer is *yes*. Intuitively, however, the answer is *no*. My de-zombified twin's cognitive reaction will differ from mine. But things are hardly ever this simple. As Gareth Evans observed in an unrelated context, "with pliant enough intuitions you can swallow anything in philosophy" (1973, p. 192). So intuition-mongering is unlikely to change any hearts and minds. Fortunately, there's a compelling argument in support of the negative answer. I set it out in the next paragraph.

God's wand-waving didn't bring about a radical change in my mental life. Things are about as colorful, noisy, and vibrant for me now as they were just moments before. My cognitive state, then, will be virtually the same. To be sure, if I *had* experienced a radical change in my mental life, it would have registered at a cognitive level. For example, if I were to experience "qualia inversion"—if, that is, the phenomenal properties associated with my experiences were systematically switched around—that would surely result in a judgment to the effect that things were never like *this* before (inwardly demonstrating). After all, my faculty of introspection works well enough to detect such massive changes. So, presumably, does yours.[8] But since my mental life wasn't radically altered in the case at issue, I won't make any such judgment. My de-zombified twin's mental life, however, *has* radically changed. The change is even more radical

[7]The two machines in my analogy occupy different physical states, so the contrasting operational effects that your button-pushing has on them is perfectly consistent with assuming that they have the same functional organization. My de-zombified twin and I, however, are supposed to be physical duplicates. So any difference at the level of judgment that God's wand-waving may bring about is inconsistent with supposing that he and I have the same functional organization.

[8]For an interesting defense of the reliability of our introspective capacities, see Jakob Hohwy (2011).

than qualia inversion would be in my case. After all, there wasn't anything it was like to be him; now there is. Since he and I are supposed to have the same functional organization, and my faculty of introspection can reliably detect significant changes in my mental life, his faculty of introspection should also be able to reliably detect such changes in his mental life. Cognitive faculties are input-output mechanisms. It follows that their abilities should be preserved by sameness of functional organization. So he *will* judge that things were never like this before. The upshot is that my twin and I must have had different functional organizations all along. The same stimulus elicits different cognitive reactions from us. But then we mustn't have been physical duplicates, since physical identity implies identity of functional organization. So, contrary to dualism, the absence of consciousness must involve a physical difference.[9]

The reasoning above was a paradigm of *a priori* reflection. We considered a hypothetical situation that elicited an intuitive judgment from which a substantive conclusion was deduced. Further *a priori* considerations were then marshaled in support of the intuitive judgment. It illustrates, then, that we can rule out the possibility of zombies *a priori*. So I conclude that zombies are, in a widely recognized sense, inconceivable.[10]

[9]One wants to know what exactly the physical difference consists in. Given how impoverished our present state of empirical knowledge is regarding the physical underpinnings of conscious experience, one can't confidently answer the question. But one can speculate. Assuming that Francis Crick and Christof Koch's (1990) hypothesis about the correlation between visual consciousness and reverberatory activity in pyramidal cells of the lower layers of the visual cortex is correct, the difference may partly consist in the absence of such activity.

[10]The central argument of this chapter rests on the premise that, when my zombie twin and I encounter the same stimulus, his cognitive reaction to it will differ from mine. This premise was supported by a series of considerations indicating that, while my de-zombified twin will judge that things were never like this before, I won't. Undoubtedly, some readers will want to resist this part of the argument. I can think of two ways to do so. One will be discussed in the main body of this paper momentarily. The other can be spelled out as follows. In order for my twin to judge that things were never like this before, he would have to compare what things are like for him now with what things were like for him prior to his de-zombification. (Presumably, it's by a comparison of this kind that I would recognize that my qualia had been inverted.) But things weren't like anything for my twin before he was de-zombified. So he can't undertake such a comparison. The upshot is that when my twin is granted the gift of consciousness, he *won't* judge that things were never like this before, and his cognitive state will be the same as mine.

I think we can suppress the worry above by looking at a case very similar to the one at issue, but where the relevant sort of comparison between phenomenal states is clearly unnecessary for the phenomenal judgment that things have never been like this before. Imagine someone completely blind—Alice, say—but whose visual faculties can retrieve and process information about her environment. Alice's condition is called *blindsight*, and is the subject of many recent studies in cognitive science. (More specifically,

The central argument has relied thus far on a case of de-zombification. We can strengthen it further by now considering a case of zombification. We begin, as before, by supposing that God regrets conferring consciousness on some but not all creatures with our physical makeup. But this time we suppose that God decides to equalize things by stripping *us* (and all of our conscious physical duplicates at nearby worlds) of phenomenal consciousness. Again, he chooses to do so with as little effort as possible. So he waves his zombification wand over the same region of modal space as before. The causal process that this initiates impinges just as much on my zombie twin as it does on me. The phenomenal effects, however, are different. My zombie's mental

---

the kind of blindsight I'm imagining is called "Type 1" blindsight. It's the most extreme form of the condition. Type 2 blindsight involves limited conscious awareness (Larry Weiskrantz 1998).) It has even received some attention in philosophy. (See John Campbell (2002).) Blindseers can visually detect what's going on in front of them. They can also use this information to navigate through their environment (Helen Briggs, 2008). Remarkably, some blindseers can even discriminate between colors, if the samples are presented in rapid succession (Iona Alexander & Alan Cowey, 2010). What blindseers lack is conscious visual awareness. The gathering and processing of visual information takes place at an unconscious level. The explanation is that information gathered by the retina is transmitted via the optical nerve to two separate regions of the brain. The first is the primary visual cortex, which subserves normal vision. The second is located at the base of the brain, where more primitive information processing takes place. When the primary visual cortex is damaged but the retina, the optical nerve, and the vision processing center at the base of the brain are left intact, blindsight results (Jake Young, 2009).

Now imagine that Alice is extremely ignorant about her visual condition. She's blindsighted, but thinks she can see. Alice's ignorance is partly due to her remarkable ability to get by (for the most part, anyway) on her own, and partly to some other cognitive deficit that shields her condition from introspection. If you ask her what it was like to see the last lunar eclipse, for example, or what it was like to see the roof of the Sistine Chapel, Alice will find it immensely difficult to answer, but you won't be able to convince her that she's actually blind. She's quite stubborn. Admittedly, it's not easy to imagine Alice's situation from her point of view, but that by itself shouldn't engender skepticism about whether the case is conceivable in the relevant sense. After all, one can't imagine my zombie's situation from his point of view. He doesn't have one. Additionally, there are actual cases of blind subjects who sincerely believe that they can see. The condition is called "Anton's Syndrome". Ned Block (2001) draws our attention to this remarkable sort of self-ignorance in order to illustrate his distinction between access consciousness and phenomenal consciousness. What I'm asking you to suppose is that Alice suffers from Anton's Syndrome.

Now suppose that one day the damage to Alice's primary visual cortex is repaired, thus enabling her to see. Suppose further that her other cognitive deficit—the condition that shielded her blindness from introspection—is also treated. How is Alice likely to react when she suddenly comes out of the coma induced by the restorative operation? Suppose you ask her, 'Have things ever been visually like they are for you now?' She will reply with a definite *no*. Of course, she can't compare what things are visually like for her now with what things were visually like for her before the operation, because there wasn't anything that things were visually like for Alice back then. She was completely blind.

life is unchanged. Mine, however, is radically altered. The flame of consciousness is extinguished. But what sort of cognitive effects will the stimulus have? In my twin's case, none at all. He will go about his business as he would have if God had done nothing. He will continue to falsely judge that he's conscious. If asked, *are you conscious*, he would say *yes*. In my case, the cognitive effects will be significant. I will judge that I'm no longer conscious. If the question, *are you conscious*, were put to me, I would say *no*. My faculty of introspection is working well, after all, and we can coherently suppose that zombification has no affect at all on its ability to register massive phenomenal change. So just as I would detect sudden qualia inversion, I would detect zombification. It may be that some psychophysical laws will be altered by the sudden absence of consciousness from our world, but there's no reason to believe that the alteration will be so drastic as to shield the change to my mental life from introspective detection. Given that in this revised case my cognitive reaction differs from my twin's, we mustn't have had the same functional organization all along.

But suppose we encounter a dualist who thinks that, conceivably, some radical changes to one's mental life will go undetected by means of introspection. The sort of dualist I have in mind agrees that qualia inversion will (as a matter of contingent fact) be detected, but thinks it *conceivable* that I continue to judge that I'm conscious after I'm zombified. So, according to her, it's *possible* that I continue to judge that I'm conscious after I'm zombified. And a zombified-me who continues to judge that he's conscious will have had the same functional organization as his zombie doppelgänger. Since one *was* conscious while the other wasn't, physicalism is false.[11] How might we reply to this dualist?

I grant that, conceivably, I will continue to judge that I'm conscious after being zombified. We can coherently suppose that God zombifies me and tampers with my faculty of introspection in such a way as to preserve my judgment. But that isn't the case that I was imagining. In the relevant case, my faculty of introspection is left unchanged by zombification, and it's by means of this faculty—*both before and after I'm zombified*—that I judge that I'm conscious. After all, my zombie doppelgänger judges that he's conscious by exercising the very same faculty as I do. So the question at stake here is whether it's conceivable that, after being zombified, I continue to judge that I'm conscious by way of the same (highly reliable) faculty that underwrites my present

[11] In an e-mail exchange about an earlier draft of this chapter, Chalmers criticized my argument on similar grounds.

judgment that I'm conscious. If this proposition *is* conceivable, then we should be able to coherently suppose it. But the argument below shows that we can't.

(1) My phenomenal judgment that I'm conscious is justified.

(2) If one's phenomenal judgment is formed via faculty F, then: the judgment is justified only if: were the evidence for it not available, and the reliability of F intact, one's judgment wouldn't be formed via F.

(3) It's via my faculty of introspection that I formed the phenomenal judgment that I'm conscious.

(4) So, the phenomenal judgment that I'm conscious is justified only if: were the evidence for it not available, and my faculty of introspection continued to operate reliably, I wouldn't judge that I'm conscious via introspection.

(5) But suppose that I'm actually zombified a minute from now (thus the evidence for the judgment that I'm conscious will no longer be available, but my faculty of introspection will be operating reliably) and that I continue to judge via introspection that I'm conscious.

(6) Then my phenomenal judgment that I'm conscious isn't justified.

I take it that (1) is an obvious truth. It's a datum that I'm justified in judging about myself that I'm conscious, just as you're justified in judging about yourself that you're conscious. But when we unpack what (1) requires, and suppose for the sake of argument that, after being zombified, I continue to judge that I'm conscious by way of the same faculty that underwrites my present judgment that I'm conscious, we end up with a contradiction. (6) follows from (4) and (5), and (4) follows from (2) and (3). I can't imagine that anyone would want to deny (3)—introspection just is the faculty by means of which we ordinarily form judgments about our mental lives—so the argument above is a *reductio* of either (2) or (5).

One might be skeptical about (2) on the grounds that counterfactual conditions on knowledge and justification are typically susceptible to counterexample. The counterexamples exploit the fact that one might actually have formed one's belief in a 'good way' even when it could easily have been formed in a 'bad way'. One wouldn't want to say about some such cases that the belief falls short of being justified. It's useful to have one clear example of this sort in mind, in order to explain why it doesn't undermine (2). Suppose that a mad scientist really wants me to believe that I'm sitting

in my armchair, typing on my laptop. Now here I am in my armchair, typing away. My sensory and cognitive faculties are working properly, so as it happens I do believe that I'm in my armchair, typing on my laptop. Because the mad scientist's wish is satisfied, he doesn't interfere. But if I weren't in my armchair, the mad scientist would zap me with his belief-inducing ray gun, thus compelling me to think that I was in my armchair, typing on my laptop. The mere fact that the mad scientist is off in the corner ready to zap me doesn't mean that, as things actually are, my belief that I'm in my armchair isn't justified. It's true that if I didn't have just this visual evidence available to me now, I would still believe that I'm in my armchair typing. But that doesn't undermine my justification. The thing to keep in mind about such examples is that in the hypothetical bad case—the case where I'm zapped by the mad scientist—either my belief-forming mechanism is rendered unreliable, or I form my belief in some way other than the faculty by means of which I actually formed it. But given the highly qualified counterfactual condition that (2) imposes on justification—were the evidence for the phenomenal judgment not available, *and the reliability of F intact*, one's judgment wouldn't be formed *via F*—such cases don't pose a challenge.

Now it's one thing for a principle to withstand the force of popular counterexamples that falsify naive variations on it; it's quite another for the principle to be counterexample free. After so many years of 'puncture and patch' epistemology, wouldn't it be overly optimistic to think that no one with enough time and ingenuity will be able to falsify (2)? Perhaps it would be, but I'm not at all confident that there's a counterexample to (2) which falsifies the spirit rather than merely the letter of the premise. By that I mean I'm not at all confident that there's a counterexample to (2) that can't be avoided by suitably reformulating it and thus deriving (6). What we're after is a response to (1)-(6) that not only eliminates the inconsistency in its present trappings, but also the *threat* of inconsistency in revised form. As long we have no independent reason to think that (2) is false in spirit, the threat of inconsistency lingers.

(1)-(6) is a *reductio* of (5), not (2). And because (5) generates an inconsistency, it mustn't be conceivable. So the dualist's objection to my central argument fails.

I promised to reply to the zombie argument in a way that wouldn't accidentally undermine the super Spartan argument. I believe the reasoning in this section fulfills my promise. To conceive of a super Spartan is to conceive of someone in pain whose outward bodily movements duplicate the bodily movements of someone who isn't in pain. A super Spartan isn't supposed to have the same functional organization as her

counterpart. Indeed, if a super Spartan is a human being, then the dependency relations between her internal states will differ quite a lot from her counterpart's. But it's precisely this supposition about zombies—that they have the same functional organization as conscious humans—which allows us to generate the problem that I have identified here. So my objection to the zombie argument avoids the risk of inadvertent sabotage. To that extent, it should be preferred over other avenues of resistance.[12]

## 4.4 The central argument defended further

After this chapter was drafted, I was informed that (1)-(6) resemble an old argument due to Sydney Shoemaker. There's no concise statement of the argument in Shoemaker's essay, but by assembling various disconnected fragments, it becomes clear that he and I are getting at the same thing.

> Supposing such cases of 'absent qualia' are possible, how might we detect such a case if it occurred? And with what right does each of us reject the suggestion that perhaps his own case is such a case ...? It is, of course, a familiar idea that behavior provides inconclusive evidence as to what qualitative character, if any, a man's mental states have. But what usually underlies this is the idea that the man himself has a more 'direct' access to this qualitative character than behavior can possibly provide, namely introspection. And introspection, whatever else it is, is the link between a man's mental states and his beliefs about ... those states. ... And if ... we take qualitative character to be something that can be known in the ways we take human feelings to be knowable (at a minimum, if it can be known introspectively), then it is not possible, not even logically possible, for a state that lacks qualitative character to be functionally identical to a state that has it (Shoemaker 1975, pp. 189-191).

In more recent work, Chalmers (1996, pp. 192-198; 2003, pp. 293-294) reformulates and criticizes Shoemaker's argument. His reformulation is provided below (with very minor adjustments for the sake of clarity).[13]

[12] In particular, the phenomenal concept strategy and type-B physicalism. See footnotes 1 and 3.

[13] Ned Block (1980) provides a rather different summary of Shoemaker's argument: "... if absent qualia are possible, then the presence or absence of the qualitative character of pain would make no difference to

(7) If zombies are possible, then they have the same phenomenal beliefs as their conscious twins, formed by the same mechanism.

(8) If zombies are possible, then their phenomenal beliefs are untrue and unjustified.

(9) If it's possible that there are beings with the same phenomenal beliefs as a conscious being, formed by the same mechanism, where those phenomenal beliefs are false and unjustified, then the conscious being's phenomenal beliefs are unjustified.

(10) Therefore, if zombies are possible, every conscious being's phenomenal beliefs are unjustified.

Chalmers rejects both (7) and (9). He rejects premise (7) on the grounds that zombies don't have phenomenal beliefs. According to him, such beliefs constitutively involve phenomenal qualities, and since zombies lack phenomenal qualities, they don't have any phenomenal beliefs. *A fortiori*, they don't have the same phenomenal beliefs as their conscious twins (2003, p. 294). But my epistemological argument can't be criticized on similar grounds, because it uses the neutral terminology ("phenomenal judgment") that Chalmers himself introduced.

Chalmers rejects premise (9) on the grounds that, according to him, the justification of a phenomenal belief doesn't depend on the mechanism (or, as I put it, the "faculty") by means of which it was formed (*Ibid.*). There is, according to him, a more "intimate epistemic relation" between a subject and her experience. One wonders, of course, what this more intimate relation is. For Chalmers, "the very fact that I have a red experience now provides justification for my belief that I am having a red experience.

its causal consequences; and so, according to a causal theory of knowledge, we could have no knowledge of the qualitative character of pain; but given that we *do* have knowledge of the qualitative character of pain ...absent qualia are not possible" (p. 380). Block criticizes the argument on the grounds that the possibility of absent qualia does *not* imply that the qualitative character of a mental state makes no difference to its causal consequences. He asks us to consider a hydraulic computing machine. "Now, for some such device, an 'absent fluid' hypothesis may be true: that is, there may be a functionally identical device that lacks fluid, for example, because it works electrically. ...But no one in his right mind would argue that since the absent fluid hypothesis is true, the presence or absence of fluid would make no difference to the operation of the hydraulic device" (pp. 382-383). Block invites us to think of qualia as like the fluid in his hydraulic computer: "the fluid is crucial to the working of the device, even though there could be a functionally identical device that lacks fluid" (*Ibid.*). I bring this up in order to point out that (1)-(6) aren't susceptible to Block's criticism.

... My experiences are *part* of my epistemic situation, and simply having them gives me evidence for some of my beliefs" (1996, p. 196). Chalmers labels this direct justification-conferring relationship "acquaintance". "To have an experience, and consequently to be acquainted with the experience, is to stand in a relationship to it more primitive than belief: it provides *evidence* for our beliefs, but it does not in itself constitute belief" (p. 197).

All of this might seem reasonable. But it's one thing for experience to "provide" evidence for a phenomenal judgment; it's quite another for it to be sufficient for the justification of that judgment. After all, one's phenomenal judgments may not be based on the evidence that one's experience provides. Imagine someone with Reverse Anton's Syndrome—a remarkable condition that causes people with normal visual faculties to think they're actually blind (Block 2001). Patients suffering from RAS make radically false and unjustified judgments about the phenomenology of their visual experiences, despite the fact that they happen to be acquainted with their visual experiences (that is, their visual experiences provide a rich source of evidence for beliefs about them). So one doesn't want to commit oneself to the unattractive thesis that our experiences suffice to justify our phenomenal judgments. Of course, Chalmers is aware of this. He rightly qualifies his position to allow for unjustified phenomenal judgments.

> Because beliefs about experiences lie at a distance from experiences, they can be formed for all sorts of reasons, and sometimes unjustified beliefs will be formed. If one is distracted, for example, one may make judgments about one's experiences that are quite false. The claim is not that having an experience is the *only* factor that may be relevant to the justification or lack of justification of a belief about experience. The claim is simply that it is *a* factor ... and provides a potential source of justification that is not present when the experience is absent (p. 197).

The point that Chalmers is trying to make here can be paraphrased as follows. I have more evidence for my phenomenal judgments now than I do after God zombifies me. But the question that deserves to be answered is whether my phenomenal judgments are presently *based* on that richer body of evidence. The mere fact that I'm acquainted with my experiences (that simply having them provides evidence for my judgments about them) isn't, as Chalmers acknowledges, the only factor in determining whether my phenomenal judgments are based on my present body of evidence. What else,

then, *is* involved in determining the basis of my phenomenal judgments? Surprisingly, Chalmers does not say. So it's unclear whether the 'acquaintance-based' epistemology of phenomenal judgment really conflicts with either (9) in Shoemaker's argument or (2) in my argument. Presumably, the other factor in determining whether a phenomenal judgment is properly based on the evidence that one's experience provides is whether the formation of that judgment was mediated in the appropriate sort of way by one's faculty of introspection. It's via such mediation that the epistemological distance between our experiences and the judgments we make about them is bridged. Consider, again, someone with RAS. Why are her judgments about the phenomenology of her visual experience unjustified? Because the faculty that mediates phenomenal judgment formation isn't related to her visual experience in the appropriate sort of way. I claim that (2) reasonably specifies a minimal condition on what mediation in the appropriate sort of way involves. And acquaintance-based epistemology, according to which one is acquainted with an experience only if "it provides *evidence* for our beliefs", doesn't challenge this claim.

At one point, Chalmers briefly considers an argument that seems to come even closer to (1)-(6) than Shoemaker's. His response is surprisingly dismissive.

> It might be objected, "But the [phenomenal] belief would still have been formed even if the experience had been absent!" To this, the answer is, "So what?" In *this* case, I have *evidence* for my belief, namely my immediate acquaintance with experience. In a different case, that evidence is absent. To note that in a different case the belief might have been formed in the absence of the evidence is not to say that the evidence does not justify the belief in this case (1996, p. 198).

Chalmers is right in one respect; to say that a belief might have been formed in a situation very like the actual one, but in which the evidence is absent, is not to say that the evidence does not justify *the proposition* believed. The evidence may well do so. But he's wrong in an even more important respect; to say that a belief might have been formed in a situation very like the actual one, but in which the evidence is absent and in which the belief is formed via the same faculty as it actually is, *is* to say that your belief *state* is not appropriately based on the evidence. And the relevant question, as I tried to emphasize earlier, is whether your cognitive state is based on the evidence that your experience provides.

I can think of one more response to the central argument on behalf of the dualist, but it's even less promising than the objections we've looked at so far.[14] The dualist might concede the intuitive judgment that the sudden presence of consciousness in my zombie doppelgänger would result in a novel and unique cognitive reaction, but object to the inference that I drew from this judgment, *viz.* that he and I didn't have the same functional organization all along. Why might this inference be targeted? Well, the dualist might think that the sudden presence of consciousness would alter my doppelgänger's functional organization. If God were to bestow consciousness on him, the causal connections between his internal states, possible environmental inputs, and subsequent behavioral outputs would change so that our respective functional organizations would differ. So it might seem as though the fact that he would react to the divine stimulus in a way that I wouldn't is perfectly consistent with the supposition that my twin and I were, prior to God's meddling, functionally identical.

This response concedes far too much to be effective. Suppose along with the dualist that my doppelgänger's functional organization *would* be altered by the sudden presence of consciousness. As long as the dualist admits that mine wouldn't be altered, my inference is warranted. For if my twin and I were physical duplicates to begin with, then the very same stimulus should alter his functional organization only if it alters my functional organization. Since his functional organization would be altered and mine wouldn't, it follows that he and I weren't functionally identical to begin with. Perhaps an analogy will drive the point home. Imagine two computers, and suppose for the sake of argument that they're physically identical. Now suppose that if one of them were to be infected by a virus, its functional organization would be altered, but if the other were infected by the very same virus, its functional organization wouldn't be altered. Are the computers functionally identical? Perhaps at some highly abstract level of description they are—a level of description at which almost anything is functionally identical to almost anything—but surely not at a level of description that specifies their programming. Since physical identity implies functional identity, the computers mustn't be physical duplicates.

[14] I discuss it only because it was frequently brought up in conversations about my central argument.

## 4.5 Closing remarks

The central argument of this chapter is quite simple. Its simplicity is a considerable strength, because it allows for very few escape routes. And, as we've seen, none of them succeed. I suspect that some readers will be disposed to think, 'Surely nothing so grand as the conclusion that zombies are inconceivable can be derived from such a simple chain of reasoning.' Perhaps. But the challenge is to cultivate one's inchoate suspicion into a concrete objection. I see no promising way of doing so.

# References


[1] Alexander, Iona and Cowey, Alan. 2010. 'Edges, Colour and Awareness in Blindsight'. *Consciousness and Cognition* 19: 520-533.

[2] Atiyah, Michael. 1988. *Collected Works, Volume I.* Oxford: Oxford University Press.

[3] Baker, Alan. 2009. 'Mathematical Accidents and the End of Explanation'. In O. Bueno and Ø. Linnebo, eds., *New Waves in the Philosophy of Mathematics*. New York: Palgrave Macmillan.

[4] Bar-On, Dorit. 1995. "'Meaning" Reconstructed'. *Pacific Philosophical Quarterly* 76: 83-116.

[5] Block, Ned. 1978. 'Troubles with Functionalism'. *Minnesota Studies in the Philosophy of Science* 9: 261-325.

[6] ———. 1980. 'Are Absent Qualia Impossible?' *Philosophical Review* 89: 257-74.

[7] ———. 2001. 'How Not to Find the Neural Correlate of Consciousness'. Reprinted in N. Block, *Consciousness, Function, and Representation: Collected Papers, Volume 1.* Cambridge: MIT Press.

[8] ———. 2002. 'The Harder Problem of Consciousness'. *The Journal of Philosophy* 99: 1-35. Reprinted in N. Block, *Consciousness, Function, and Representation: Collected Papers, Volume 1.* Cambridge: MIT Press. Page references to this volume.

[9] Block, Ned and Stalnaker, Robert. 'Conceptual Analysis, Dualism, and The Explanatory Gap'. *Philosophical Review* 108: 1-46.

[10] Briggs, Helen. 2008. 'Blind Man Navigates Maze'. BBC News. http://news.bbc.co.uk/1/hi/7794783.stm

[11] Buchanan, Ray. 2010. 'A Puzzle About Meaning and Communication'. *Noûs* 44: 340-371.

[12] Buchanan, Ray and Ostertag, Gary. 2005. 'Has the Problem of Incompleteness Rested on a Mistake?' *Mind* 114: 889-912.

[13] Burge, Tyler. 'On Knowledge and Convention'. *Philosophical Review* 84: 249-255.

[14] ———. 'Individualism and the Mental'. In P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy, Volume IV: Studies in Metaphysics.*

[15] ———.1998. 'Frege on Knowing the Foundations'. *Mind* 107: 305-347.

[16] Campbell, John. 2002. *Reference and Consciousness.* Oxford: Clarendon Press.

[17] Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press.

[18] ———. 2002. 'Does Conceivability Entail Possibility?' In T. Gendler and J. Hawthorne, eds., *Conceivability and Possibility.* New York: Oxford University Press.

[19] ———. 2003. 'The Content and Epistemology of Phenomenal Belief'. Originally published in Q. Smith and A. Jokic, eds., *Consciousness: New Philosophical Perspectives.* New York: Oxford University Press. Reprinted in D. Chalmers, *The Character of Consciousness.* New York: Oxford University Press. Page references to this volume.

[20] ———. 2010. *The Character of Consciousness.* Oxford University Press.

[21] Chalmers, David and Jackson, Frank. 2001. 'Conceptual Analysis and Reductive Explanation'. *Philosophical Review* 110: 315-161.

[22] Chomsky, Noam. 1980. *Rules and Representations.* New York: Columbia University Press.

[23] ———. 1986. *Knowledge of Language.* New York: Praeger.

[24] ———. 2000. *New Horizons in the Study of Language and Mind.* Cambridge, UK: Cambridge University Press.

[25] Colasent, Rita and Griffith, Penny L. 1998. 'Autism and Literacy: Looking into the Classroom with Rabbit Stories'. *The Reading Teacher* 51: 414-420.

[26] Crick, Francis and Koch, Christof. 1990. 'Toward a Neurobiological Theory of Consciousness'. *Seminars in the Neurosciences* 2: 263-275.

[27] Davidson, Donald. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

[28] Davis, Philip J. 1981. 'Are There Coincidences in Mathematics?' *American Mathematical Monthly* 88: 311-320.

[29] Dehaene, Stanislas. 2009. *Reading in the Brain: The New Science of How we Read*. New York: Penguin Books.

[30] Derbyshire, John. 2004. *Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics*. London: Plume.

[31] Devitt, Michael. 2006. *Ignorance of Language*. Oxford University Press.

[32] Dummett, Michael. *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press.

[33] Edgington, Dorothy. 2004. 'Two Kinds of Possibility'. *Aristotelian Society Supplementary Volume* 78: 1-22.

[34] Elkies, Noam. http://www.math.harvard.edu/ elkies/ferm.html

[35] Freedman, David; Pisani, Robert; and Purves, Roger. 1998. *Statistics* (3rd edition). New York: W.W. Norton & Co.

[36] Field, Hartry. 1986. 'Stalnaker on Intentionality'. *Pacific Philosophical Quarterly* 67: 98-112.

[37] Fodor, Jerry. 1983. *The Modularity of Mind*. Cambridge: MIT Press.

[38] ———. 1989. 'Why Should the Mind be Modular?' In A. George, ed., *Reflections on Chomsky*. Oxford: Blackwell.

[39] Frege, Gottlob. *The Foundations of Arithmetic*. Translated by J.L. Austin. Evanston: Northwestern University Press.

[40] George, Alexander. 1990. 'Whose Language is it Anyway? Some Notes on Idiolects'. *Philosophical Quarterly* 40: 274-298.

[41] Gowers, William Timothy. 2007. 'Mathematics, Memory, and Mental Arithmetic'. In M. Leng, A. Paseau, and M. Potter, eds., *Mathematical Knowledge*. Oxford: Oxford University Press.

[42] Grice, Paul. 1957. 'Meaning'. *Philosophical Review* 66: 377-388.

[43] ———. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

[44] Hart, H.L.A and Honoré, Tony. 1959. *Causation in the Law*. Oxford University Press.

[45] Hawthorne, John and Magidor, Ofra. 2009. 'Assertion, Context, and Epistemic Accessibility'. *Mind* 118: 377-397.

[46] Heck, Richard. 2002. 'Do Demonstratives Have Senses?' *Philosophers' Imprint*.

[47] ———. 2006. 'Idiolects'. In J. J. Thomson and A. Byrne, eds., *Content and Modality*. Oxford University Press.

[48] Hirsch, Eli. 2010. 'Kripke's Argument Against Materialism'. In R. C. Koons and G. Bealer, eds., *The Waning of Materialism*. New York: Oxford University Press.

[49] Hohwy, Jakob. 2011. 'Phenomenal Variability and Introspective Reliability'. *Mind and Language* 26: 261-286.

[50] Howell, Robert. 'The Two-dimensionalist Reductio'. *Pacific Philosophical Quarterly* 89: 348-358.

[51] Jackson, Frank. 1982. 'Epiphenomenal Qualia'. *Philosophical Quarterly* 32: 127-136.

[52] Kripke, Saul. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

[53] Kratzer, Angelika. 1981. 'The Notional Category of Modality'. In H.J. Eikmeyer and H. Rieser, eds., *Words, Worlds, and Contexts*. New York: Walter de Gruyter & Co.

[54] Lange, Marc. 2010. 'What Are Mathematical Coincidences (and Why Does It Matter)?' *Mind* 119: 307-340.

[55] Laurence, Stephen. 1996. 'A Chomskian Alternative to Convention-Based Semantics'. *Mind* 105: 269-301.

[56] ———. 1998. 'Convention-Based Semantics and the Development of Language'. In P. Carruthers and J. Boucher, eds., *Language and Thought: Interdisciplinary Themes*. Cambridge, UK: Cambridge University Press.

[57] ———. 2003. 'Is Linguistics a Branch of Psychology?' In A. Barber, ed., *The Epistemology of Language*. Oxford University Press.

[58] Lewis, David. 1969. *Convention*. Camrbidge, MA: Harvard University Press.

[59] ———. 1983. 'Languages and Language'. In K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science: Volume VII*. Minneapolis: University of Minnesota Press. Reprinted in D. Lewis, *Philosophical Papers: Volume I*. Oxford University Press. Year and page references to this volume.

[60] Levine, Joe. 1983. 'Materialism and Qualia: The Explanatory Gap'. *Pacific Philosophical Quarterly* 64: 354-361.

[61] ———. 1993. 'On Leaving Out What It's Like'. In M. Davies and G. Humphreys, eds., *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell.

[62] Maor, Eli. 1998. *e: The Story of a Number*. Princeton: Princeton University Press.

[63] McGinn, Colin. 1989. 'Can We Solve the Mind-body Problem?' *Mind* 98: 349-366.

[64] Nagel, Thomas. 1974. 'What Is It Like to be a Bat?' *Philosophical Review* 4: 435-450.

[65] Nation, Kate; Clarke, Paula; Wright, Barry; Williams, Christine. 2006. 'Patterns of Reading Ability in Children with Autism Spectrum Disorder'. *Journal of Autism and Developmental Disorders* 36: 911-919.

[66] Neale, Stephen. 2004. 'This, That, and the Other'. In A. Bezuidenhout and M. Reimer, eds., *Descriptions: Semantic and Pragmatic Perspectives*. Oxford: Oxford University Press.

[67] Potter, Michael. 1993. 'Inaccessible Truths and Infinite Coincidences'. In J. Czermak, ed., *Philosophy of Mathematics: Proceedings of the 15th International Wittgenstein Symposium*. Vienna: Verlag Hölder-Pichler-Tempsky.

[68] Putnam, Hilary. 1962. 'The Analytic and the Synthetic'. In H. Feigl and G. Maxwell, eds., *Minnesota Studies in the Philosophy of Science* 3. Minneapolis: University of Minnesota Press.

[69] Sacks, Oliver. 2010. 'A Man of Letters: Why Was the Morning Paper Suddenly in a Different Language?' *The New Yorker*. June 28.

[70] Schiffer, Stephen. 1995. 'Descriptions, Indexicals, and Belief Reports: Some Dilemmas (But Not the Ones You Expect)'. Reprinted in G. Ostertag, ed., *Definite Descriptions: A Reader*. Cambridge, MA: MIT Press.

[71] ———. 'Stalnaker's Problem of Intentionality'. *Pacific Philosophical Quarterly* 67: 87-97.

[72] Searle, John. 1987. 'Indeterminacy, Empiricism, and the First Person'. *Journal of Philosophy* 84: 123-146.

[73] Shoemaker, Sydney. 1975. 'Functionalism and Qulia'. *Philosophical Studies* 27: 291-315.

[74] Soames, Scott. 1984. 'Linguistics and Psychology'. *Linguistics and Philosophy* 7: 155-179.

[75] ———. 2006. 'Understanding Assertion'. In J. J. Thomson and A. Byrne, eds., *Content and Modality* Oxford: Oxford University Press.

[76] Stalnaker, Robert. 1979. 'Assertion'. In P. Cole, ed., *Syntax and Semantics 9*. New York: New York Academic Press.

[77] ———. 1984. *Inquiry*. Cambridge, MA: MIT Press.

[78] ———1986. 'Replies to Schiffer's "Stalnaker's Problem of Intentionality" and Field's "Stalnaker on Intentionality"'. *Pacific Philosophical Quarterly* 67: 113-123.

[79] ———. 2002. 'What Is It Like to be a Zombie?' In T. Gendler and J. Hawthorne, eds., *Conceivability and Possibility*. New York: Oxford University Press.

[80] ———. 2006. 'Replies'. In J.J. Thomson and A. Byrne, eds., *Content and Modality*. New York: Oxford University Press.

[81] ———. 2009. 'On Hawthorne and Magidor on Assertion, Context and Epistemic Accessibility'. *Mind* 118: 399-409.

[82] Stoljar, Daniel. 2005. 'Physicalism and Phenomenal Concepts'. *Mind and Language* 20: 296-302.

[83] ———. 2006. 'Actors and Zombies'. In J.J. Thomson and A. Byrne, eds., *Content and Modality*. New York: Oxford University Press.

[84] Tappenden, Jamie. 2008a. 'Mathematical Concepts and Definitions'. In P. Mancosu, ed., *The Philosophy of Mathematical Practice*. Oxford University Press.

[85] ———. 2008b. 'Mathematical Concepts: Fruitfulness and Naturalness'. In P. Mancosu, ed., *The Philosophy of Mathematical Practice*.

[86] Weiskrantz, Larry. 1998. 'Consciousness and commentaries'. In *Towards a Science of Consciousness II–The second Tucson discussions and debates*. Cambridge, MA: MIT Press, pp. 371-377.

[87] Whitaker, Craig. 1990. 'Ask Marilyn'. *Parade Magazine*. 9 September.

[88] Yablo, Stephen. 1993. 'Is Conceivability a Guide to Possibility?' *Philosophy and Phenomenological Research* 53: 1-42.

[89] ———. 2000. 'Textbook Kripkeanism and the Open Texture of Concepts'. *Pacific Philosophical Quarterly* 81: 98-122.

[90] ———. 2008. *Thoughts*. Oxford University Press.

[91] Young, Jake. 2009. 'Intact Visual Navigation in a Blindsighted Patient'. http://scienceblogs.com/purepedantry/