Wesleyan University The Honors College

Ending the Zombie Wars: Conceivability, Possibility, and Scientific Understanding

by

Lucas Carrico
Class of 2009

A thesis submitted to the
faculty of Wesleyan University
in partial fulfillment of the requirements for the
Degree of Bachelor of Arts
with Departmental Honors in Philosophy

Middletown, Connecticut April, 2009

Table of Contents

Acknowledgements

Dedicated to Joe Rouse, Sanford Shieh, and Eclectic House.

I'd like to thank my family.

Chapter One: Possible Worlds in One and Two Dimensions

0 Introduction

Most everyone can identify some hypothetical situations as im*possible,* thus the idiom, "when pigs fly." And most everyone has some facility with some conception of *necessity*, especially those of us who have consumed large amounts of liquid prior to sitting through long movies or car rides. If, however, one is seeking to talk to *philosophers* about modality — their preferred term for possibility and necessity — one is advised to start with the work of a man named Saul Kripke. Prior to formal work by the genius philosopher and logician in the late 1950s and early 1960s, much of twentieth-century analytic philosophy was spent arguing over the status of modal logics, and therefore over the *meaning* of talk about various forms of modality. Kripke's formal work provided a complete system expressible in logical form, from which other "normal" modal logics were built, and he also provided the paradigm ontological object for understanding modality: the possible world. Although a victory for those seeking legitimacy for modal discourse, this was groundwork, or at least not the full substantive extent of what would cement Kripke's status as a philosophical legend. Instead, it was a series of talks given at Princeton in 1973 and republished in 1980 as the book *Naming and Necessity*[1] *(N&N)*, which represent the philosopher's best known original contribution to modal thought.

The first section of Chapter 1 will provide a brief introduction to the first of these two Kripkean projects — possible worlds semantics — on the way to a closer focus on of the details of his second mentioned project, which we will explore in the

[1] Kripke (1980)

second, central section of that chapter. Kripke's theory of direct reference as introduced in *N&N* must be understood as a means to two complementary ends: rejection of the then-orthodox descriptivist views of language, and groundwork for support of his belief in some forms of essentialism, the doctrine that some things have certain properties essentially, or in a metaphysically necessary respect. In particular, one striking characteristic of his account of such terms as proper names (such as "Aristotle") and general natural kind terms (like "water" or "tiger" or "cat") was that the *meanings* of those terms lay not in the linguistically defined *concepts* of even their most competent users, but instead were to be found in their *referents themselves*. He was not alone of course; the examples of water and cats and externalist arguments about meaning, language, and concepts, were exposited more canonically in the work of Hilary Putnam.[2] However, in the hands of Kripke, the effects of these arguments took a surprisingly metaphysical turn, as it was suggested that the necessity of certain *essential* features of some individuals or kinds was somehow external to the linguistic or conceptual "senses" of the terms used to describe them; it seemed that Kripke had, by way of mere thought experiments and intuitions, uncovered a dimension of necessity in the world itself.

As convincing as Kripke's arguments were taken by many to be, they did not succeed in putting the descriptivist picture of language and meaning to sleep. For one thing, *N&N* was far from systematic; unlike the logician's formal work, it did not aspire to completeness about any subject matter, and left room for reformulations galore. For another, Kripke left open the hint of an important *epistemic* dimension regarding possibility and necessity, one that could be read into even his seemingly

[2] Putnam (1973).

most metaphysically consequential examples. In the third section of Chapter 1 we will consider an account of modality propounded, endorsed, and defended by two philosophers that deserve the title "neo-descriptivist," David Chalmers and Frank Jackson. Each of the two published a hugely influential manifesto during the 1990s, and although the project differed starkly between the two — Chalmers sought to defend first-person phenomenal consciousness against physicalism in the philosophy of mind while Jackson attempted to defend the role of conceptual analysis in metaphysics and philosophy in general — each made use of a *two-dimensional* appropriation of Kripke's possible world semantics for modal discourse. What the two-dimensional framework allows is for the accommodation of Kripke's points about the *rigidity* of some terms — their ability to pick out the same thing across possible worlds — while deflating the necessity of their features as running across only one of two (conveniently, the second of two) modal dimensions. In the hands of neo-descriptivists, Kripke's metaphysical realm of modality is created *from* the linguistic, conceptual realm of necessity (this is the realm in which philosophers have made themselves at home since the time of Descartes) by way of mere *stipulation*.

This paper will seek to bring some measure of closure to major debates from the past couple decades between Kripke's anti-descriptivist essentialist soldiers and their neo-descriptivist challengers. (As we will see, the sides could just as well be the neo-descriptivist soldiers of Frege and their anti-descriptivist essentialist challengers.) To do so, we will begin by revisiting a sampling of the classical modal examples, including the famous sci-fi Twin Earth scenario and my person favorite, robot- and demon-cats. Pushing further, we will up the ante in Chapter 2 and consider Kripke's

C-fiber argument and Jackson's knowledge argument against materialism as appetizers to Chalmers and his notorious Zombie Worlds, which have provided a sort of main event for these tussles over the last fifteen or so years by tying the neo-descriptivist/anti-descriptivist debate to the outcome of a longstanding conflict in the metaphysics of mind between materialism and dualism (or as Chalmers has at times advocated, "panprotopsychism"). Following a few rounds in that ring – noting both objections to Chalmers and his responses (as well as Jackson's repudiation of his own famous argument) in Chapter 3 — will bring us to what I take to be the underlying issue of importance in all of these debates: to what extent either neo-descriptivism *or* anti-descriptivism is capable of providing a substantive metaphysical grounding for scientific laws and natural kinds.

If one considers the most excellent responses to two-dimensionalism from the last decade or so, it may seem that these disputes ought already to be considered closed — or at least inconsequential. Scott Soames wrote an entire book disputing the legitimacy of the 2-D framework for provision of adequate natural kind semantics.[3] And George Bealer's discussion of neo-descriptivism in his article on modal epistemology from the 2002 compendium *Conceivability and Possibility* ruthlessly points out contradictions from within, as well as obvious failures to accord with our basic modal intuitions.[4] But Soames and others (Christian Nimtz comes to mind[5]) have primarily sought to reassert anti-descriptivist principles from within the philosophy of *language*, and Bealer's account focuses on the linguistic needs of a correct modal *epistemology*. Chapter 4 will begin with an objection from E.J. Lowe:

[3] Soames (2005)
[4] Bealer (2002)
[5] Nimtz (2004)

why does any of this matter for *metaphysics*? If we are to truly connect with Kripke's project, we must answer that question.

My somewhat naïve metaphysical treatment of the terrain will end up highlighting recent work in the philosophy of *science*, in order to see what light the metaphysical needs of scientists can tell us about the essentialist/descriptivist tension with regard to modality and natural kinds. I will show that when one takes a framework like that pioneered by philosopher of science Marc Lange, who reduces the notions of natural *law*, natural (and all other forms of) *necessity*, and as a result natural *kinds* to invariance among counterfactuals that support stable inferential rules, one is in the position to see that both essentialism and descriptivism aspire to metaphysical systems that have as their building blocks unfaithful portrayals of scientific induction, explanation, understanding, laws, and kinds.

Taking a pragmatic, counterfactual-focused view of modal metaphysics, we will then return to our cat example. We will conclude with work done by Michael Silberstein on emergence, and see the ways in which both anti-descriptivism and neo-descriptivism serve to obscure the extent to which the answers about consciousness are still yet to be determined. That is, while Kripke and Chalmers were helpful in putting to rest descriptivism and reductionism, respectively, their descendent theories of essentialism and property dualism at this point serve to inhibit the progress of cognitive science in explaining consciousness by trying to adjudicate in the form on already existing concepts rather than waiting for those conceptual resources that accompany successful inductive strategies.

1. Kripke's Possible Worlds Semantics

1.1 Kripke Semantics, Skipping a Lot

But first, a quick crash course on Kripkean possible worlds. Of course, the level of detail one provides in this kind of account is inevitably somewhat arbitrary. After all, possible worlds semantics were first introduced as *semantics* for modal *logics.*[6]This essay is not a history of modal logic, but instead joins an ongoing debate about modal metaphysics for science and mind. Much in the way of technical details can be skipped, since those logicians in on this debate are likely to understand the underlying fine points, and those lacking the technical understanding are not likely to care. We will look at the basics:

- A Kripkean modal model (a mouthful!) is a set of possible worlds W_i, one of which is flagged as the actual world $W_@$. Each possible world is a complete history of a universe, with a set x_i of individuals that exist in that world, and a list of predicate statements (facts) that are true at that world.
- A kind of relation holds between individual worlds indicating accessibility, or as philosophers say (relative) possibility, between worlds. It is important to make note of a technical fact that the characteristics of this relation directly affect the characteristics of the set of worlds; for example, the logic S5 is a logic that has a relation that is reflexive (a world is possible relative to itself), symmetric

[6] See Kripke (1959) and (1963).

(relative possibility is a two-way street) and transitive (possibility relative to a world possible relative to another world entails possibility relative to the other world), and therefore possesses a set of possible worlds all possible relative to each other. Though there has been much debate about the suitability of S5, for the sake of simplicity we will adopt it as our choice.

- A given statement S is said to be *possible in a world W* just in case S is true for some world V possible relative to W. Similarly, S is said to be *necessary in W* just in case S is true for every world V accessible from *W*. If we are in an S5 modal structure (Kripke's completeness theorem for modal logic was proved for a variant of **S5**, and his quantified modal logic used a variant of S5), the upshot of all of this is that *S is necessary* just in case it is *true in all possible worlds*, and that *S is possible* just in case it is *true in at least one possible world*. S is said to be true *simpliciter*, in the ordinary sense, just in case it is true in the actual world $W_{@}$.

1.2 But what are they?

Even the novice reader might already notice, if vigilant, that there is a question left unanswered. While providing a way to coherently deal with possibility, Kripke's semantics beg the question of what exactly a possible world *is*, and further leave open where to fit it in one's epistemology or ontology Ones method of

answering this question and the metaphysical consequences determined by the choice are of central issue for this paper.

In the Preface to the print copy of *N&N*, Kripke writes:

> 'Possible worlds' are little more than the miniworlds of school probability writ large… 'Possible worlds' are total 'ways the world might have been', or states or histories of the *entire* world… Certainly the philosopher of 'possible worlds' must take care that his technical apparatus not push him to ask question whose meaningfulness is not supported by our original intuitions of possibility that gave the apparatus its point. Further, in practice we cannot describe a complete counterfactual course of events and have no need to do so… In practice [a practical description] involves less idealization both as to considering entire world histories and as to considering *all* possibilities.[7]

These days, there are many styles of philosophical possible worlds. They can be posited as linguistic, epistemic, or metaphysical entities (not to mention normative, temporal, etc.). Unfortunately, the most robust ontological answer to the question of possible worlds is the hardest to believe, that given by David Lewis's modal realism.[8] For those who refuse to (or just cannot) commit to believing in possible worlds as an uncountably infinite multitude of concretely existing entities, their conceptualization

[7] Kripke (1980), p. 18.
[8] See Lewis (1986).

presents difficulties that run outside the scope of this paper. One thing that we must not ignore, however, is that the above technical Kripkean version of possible worlds, taken for granted by many contemporary users, has possible worlds and individuals as the primitive objects. From these individuals and worlds and the predicate statements that are true or false of and in them, presumably properties, laws, and ordinary facts can be constructed. Nevertheless, while the anti-descriptivism/neo-descriptivism conflict tends to turn on how we match our terms and concepts with these individuals, our considerations in Chapter 4 will make it clear that *what individuals there are* in a given possible world is something that is not at all obvious or distinguishable apart from our practical interaction with our actual world.

2. Necessity, A Posteriority, and Essentialism

2.1 Descriptivism

The dominant philosophy of language (or family of philosophies, or style of philosophy) at the time of *N&N* was made up of a combination of descendent theories molded together. We can call the family of theories *descriptivist*. As Nathan Salmon points out, a distinction between the *sense* of a term, understood as all at once its cognitive significance, informative content, and method of reference-fixing, and *the denotation*, which is whatever is referred to by the term — we'll call it the term's *extension* — had been passed down from Gottlob Frege.[9] Since the sense was

[9] See Salmon (1981), pp. 9-16. Although Chalmers (2006) and Soames (2006) provide more complete stories of the descriptivist setup for Kripke's argument, Salmon's account has the advantages of both relative brevity and temporal proximity to the period when the descriptivist picture was still dominant.

separate from the external extension, so the thinking went, its meaning must be internal, and so mental content was thought to be completely internal as well. Grasp of the sense was meant to be a prerequisite for any sort of belief about the term, and was thought to completely determine the extension.

Adding to this historical strand were two others. From the time of Immanuel Kant, it had been considered common knowledge that for something to be necessarily true, it must not depend on characteristics of the world, and therefore its truth must be knowable *a priori* (without recourse to experience, as opposed to *a posteriori* knowledge that depends on some kind of nosing about in the world). On the other hand, Rudolph Carnap took Frege's notion of sense and, after wondering what cognitive content could be, found possibility and necessity to be just the right things to explain the *intension* of a term, or its meaning over and above extension. An intension could be thought of as a function from possibilities to extensions. Taken together, the Fregean conception of sense as completely reference-determining, the Kantian rationalist conception of necessity, and the Carnapian notion of intension combined into what Chalmers calls the ***Neo-Fregean Thesis***: "Two expressions 'A' and 'B' have the same intension iff 'A=B' is a priori."[10]

Two more important issues are lurking behind the scenes in this picture of pre-Kripkean philosophy. Soames points out that for the descriptivists, description or definition was primary, and thus necessity and possibility were subservient notions. Logical positivism took for granted that the position of the philosopher was to analyze meanings, and to leave the empirical fact-finding to scientists. Thus, the position or claim that some objects have certain properties necessarily, *irrespective of*

[10] See Chalmers (2006).

how they are described, would be unintelligible to a descriptivist. After all, for such a person, necessity gets its power exactly from the apriority of definition.[11] But what Kripke sought to establish, and what today's "scientific essentialists" seek to build on and defend, was that certain objects as conceptualized by scientists and followers of science, have meanings that are inherently intertwined with the empirical fact-finding itself.

2.2 Descriptional Terms, Non-descriptional Terms, and Proper Names

According to Kripke's understanding of descriptivist theory, a singular term *T* — a term with a single unique extension *x* for every world *W* in which it refers at all — denotes its referent by way of a *definite description*, which is something like a property *P* or set of properties P_i that serve as necessary and sufficient conditions for the identification of *x* with *T*. The connection between *x* and *T* is a priori; if *P* or P_i failed to be true of *x*, then it would simply not count as *T*, and conversely if not a single thing had *P* or P_i true of it, then *T* would not refer. The intension of a singular term picks out its extensions in various possible worlds, because only those individuals that fit the definite description get the endorsement of reference under the singular term. Kripke rightly points to the account as circular.

In contrast, the *direct* theory of reference that Kripke espoused treats a singular term *T* as having no meaning above and beyond its referent. To do this, the theory utilizes the notion of a *rigid designator*. A rigid designator simply assigns *T* to one variable *y* in every world *W* in which *y* exists. The ability to do so at least in principle follows from the structure of Kripke's formal semantics for quantified

[11] See (iv) and (v) in Soames (2006), p. 289.

modal logic, but the view ends up having a convincing intuitive advantage. A sort of rough sketch of how such reference could occur is also traced out, and goes something like this: an initial baptismal event acquaints one with a certain individual *o* under a certain name *n*, and the name is then passed along a causal chain to future utterers. Nothing about this assures that knowledge of the correct properties (those possessed by *o*) are possessed by those who use the name *n*, yet those users can (and often do) still refer to *o*.

As an example, consider the proper name 'Aristotle,' which belongs or belonged to (as far as we know) a famous philosopher, who could plausibly be described by the definite descriptions, "the last great philosopher of antiquity," or "the man who taught Alexander the Great."[12] Suppose, however, that historians got it wrong about the identity of Alexander's teacher, or that there lived after the time of Aristotle (but still during "antiquity") some great philosopher who was particularly bad at the whole academic politics thing and therefore slipped through the cracks of history. In either case, it seems as if the descriptivist account would pick out some *other* individual than Aristotle, while the counterfactual *Aristotle* was going about his business in the possible worlds in question without us having any way to talk of him. There is, of course, one (circular and therefore trivial) description that could do the job of picking out the right guy: "the man called Aristotle."

Kripke's sea change in analytic philosophy consists partly in the fact that his story really seems to accord much better with our intuitions regarding to whom we refer when we use proper names. It seems that at least some of the time we use names as rigid designators, rather than descriptions, in much the same way that we use

[12] Kripke (1980), p. 6, 30

indexicals like “here” or “him”. The problem for descriptivism was that Kripke did not stop there.

2.3 Twin Earth and the Necessary A Posteriori

Having begun his case for non-descriptional rigid designation of singular terms with proper names, Kripke generalizes to sentences about natural kinds (his points about artifacts and individuals are similar, but will have to be left aside as their import lies outside the scope of this paper). For each of these types of expression, the direct theory of reference along with a constitutive essentialist premise entails acceptance of a substantial essentialist conclusion, which seems to be an example of a truth that is both necessary and a posteriori. These Kripke cases serve to divorce the notions of necessity and apriority, or of what might be regarded as metaphysical and epistemic necessity, and thus to force the descriptivist to cede some metaphysical ground.

We can begin with perhaps *the* canonical philosophical example of natural kind information given to us by the hard sciences. Consider the statement, “Water=H_20.” This statement in the hands of Hilary Putnam became the famous Twin Earth Argument,[13] in which one travels to a faraway (or considers another possible) world, where there exists a substance that shows up in all the same places and plays all the same roles as H_20 does in our world, and yet has a different chemical structure, abbreviated XYZ:

[13] Putnam (1973)

> "If there were a substance, even actually, which had a completely different atomic structure from that of water, but resembled water in these respects, would we say that some water wasn't H_2O? I think not. We would say instead that just as there is a fool's gold there could be a fool's water; a substance which, though having the properties by which we originally identified water, would not in fact be water.[14]

What Kripke is asserting is that for some kinds of things, those kinds of individuals *given to us by science*, membership in that kind is determined by structural (in this case, chemical structural) similarity rather than by definition or description, even when that structural similarity is not fully grasped by all or even any competent users of the term. So a sample is water entirely and only because it shares some constitutive natural features with our paradigm case of water, which happens to be H_2O, and no matter how water-like some other substance is, if it is not H_2O then it is not water. The point is seen by scientific essentialists as intrinsic to what naturalism is meant to accomplish: an *objective* account of what things are. If a kind has a relevant property in any possible world, then it has that property at every possible world in which it exists; otherwise, it wouldn't *be the same thing*.

2.4 Tiger-Lizards and Robot Cats

[14] Kripke (1980), p. 128

For a more colorful argument along the same lines, let us consider my favorite examples. I always loved cats, but my family never could own one, since we were all allergic. But I was fascinated to read educational kids' picture books explaining that "big cats" like tigers (belonging to the subfamily *Panthera*) and garden-variety house cats and their wild relatives (the felines of subfamily *Felinae*) both belonged to the same biological family (*Felidae*).[15] How pleased I was to see a decade and a half later that Kripke had made central use of these majestic creatures!

Consider first the tiger, for which Kripke provides a dictionary definition "a large carnivorous quadrupedal feline, tawny yellow in color with blackish transverse stripes and white belly."[16] He points out, though, that unfortunate three-legged tigers (especially those former four-legged ones) would nevertheless count as tigers, and we can also consider the fact that certainly albino tigers (which are presumably white) have existed from time to time. So, membership in the tiger species does not seem to boil down to a definition. On the other hand, if we found creatures appearing to be tigers (satisfying all the regular indicators) that turned out upon closer inspection to be *reptiles* (lizards, let's say), then we would have to either deny the tiger-lizards membership in the "tiger" kind.

If tiger-lizards seem like too strange of an example, let us move on to their smaller cousins. When young, I was sometimes lucky enough to get the chance to care for my neighbors' cats while their owners were on vacation, and became quite familiar with many of the mannerisms and physical characteristics of the cute, soft, furry, capricious animals. It is likely that I thought of them as more like people than

[15] Don't worry; I was not so precocious a youngster as to actually know the *names of* those taxonomical classifications memorized, although I probably implicitly grasped much of the substance of the distinctions.
[16] Kripke (1980), p. 119.

felines; I would go so far as to say that their membership in the feline subfamily had almost nothing to do with how I thought of them, and furthermore if it had come to light that they were indeed not felines but *demons*, that might not have surprised me all that much (or at least on those days when I received unpleasant and undeserved scratches and bites for my troubles). Yet if the neighbors' pets were to turn out to be something other than animals, then surely they would not count as *cats* in the sense of the term meant by current biological science, which at this point has no taxonomical category for *demon*.

Kripke mentions, in addition to the supernatural case of demon-cats, a possible world (or set of worlds) in which cats turned out to instead to be exquisitely crafted *robots!*[17] (Let us henceforth call these hypothetical machines 'CROBOTS.') This scenario is not so easily disregarded on grounds of absurdity, due to human success with some limited robots; we cannot dismiss offhand the possibility that a far-advanced civilization has been dropping off CROBOTS over the millennia in order to spy on humans, as paranoid as that scenario would seem to be, and certainly nothing stops us from envisioning a possible world in which such is the case. Unlike the conceptual confusion that arises in considering tiger-lizards, and unlike the indeterminacy of trying to apply biological understanding to supernatural demon-cats, withholding cat-hood from CROBOTS is relatively simple: cats are animals, robots are not animals, and therefore counterfactual CROBOTS, even the exquisite CROBOTS that satisfy all of the epistemic indicators of cathood, are not cats.

Those who have read *N&N* know that these two classes of examples do not come close to exhausting the cases Kripke gives of propositions that are allegedly

[17] His word is *automata*; see Kripke (1980), pp.122-3.

necessary and yet a posteriori. To read the lectures, one might guess that Kripke's essentialism about natural kinds and origins of individuals were merely interesting expansions to what he seems most concerned with: the application of Leibniz's law to modal intuitions.[18] Also known as the indiscernibility of identicals, the principle requires that identical objects share all properties; if so, the thinking goes, then modal properties must agree for identical objects. However, in pursuing this line of thought, Kripke ventures out of bounds of the essentialist arguments above, and opens the door for strenuous objections, as well as a strong descriptivist response.

2.5 Identity and the Weak Kripke Cases

Greek mythology held that Hesperus, or the evening star, was the brother of Phosphorus, the morning star. For the Greeks, the two objects, visible in the sky at two separate times of day, were different entities. However, astronomy since ancient times has come to the conclusion that they are one and the same entity, Venus. Since Hesperus and Phosphorus are names, and therefore for Kripke are rigid designators, the statement 'Hesperus = Phosphorus' must convey a necessary truth, since both refer to Venus in every possible world. If this is analyzed by assigning the same variable v to each name, we merely get that 'Necessarily, v = v," which seems to be an example of an a priori statement if there is such a thing at all. In order to do the work needed to show its a posteriority, Kripke must spell out in greater detail what he has in mind for modal epistemology.

[18] Kripke (1980), p. 3 is evidence that significant motivation for the lectures/book came from this concern.

Discovery that Hesperus is Phosphorus is an empirical *achievement*, and therefore represents the acquisition of a posteriori knowledge for Kripke, for the same sort of reason as the water case: historically, we did not always know that it was the case that the two names referred to one entity. "Hold the boat," the vigilant reader may well be thinking at this point. If Hesperus and Phosphorus are to be thought of as two different individuals, then we must be talking about them as the evening star and morning star, or as brothers, some Greek or other *folk* objects. And in such a case, it seems much less *necessary* that they should be identical, since the appearance of either or both could have been caused by some other entities, such as two distinct stars. "Furthermore, didn't you just say that Kripke has many of these kinds of examples?" the reader might continue to gripe. "There must be some common reason for his mistake."

The source of the problem, and of Kripke's seeming inconsistency, lies in this idea of a *qualitatively identical epistemic situation*. It might have turned out, as far as the Greeks knew, that Hesperus and Phosphorus were distinct. Such a contingency can be likened to the way certain tough math problems, at least until proven one way or the other, might be true, or might be false. "Obviously, the 'might' here is purely 'epistemic' — it merely expresses our present state of ignorance, or uncertainty."[19] So an epistemic aspect of modality seems to have cropped up, and it is one that affects our intuitions about a priority, and necessity. As Soames has pointed out, whereas before we considered the knowledge of the nature of water in order to dispel Twin Earth-type intuitions in which water is some other chemical compound, now the opposite is happening and we are letting the nature of the concepts from which we

[19] Kripke (1980), p. 103.

came to know Venus expand our sense of the possible worlds involved.[20] Bealer points out that the failure to distinguish this *epistemic* meaning of modality from the nonepistemic meaning is responsible for much of the confusion that follows.[21]

We will consider one more example, to which we will return later. Suppose C-fibers are the type of nerve fibers known to accompany pain sensations. Kripke points out 'pain' gets its name from a sensation we tend to feel, the *internal* manifestation of what happens to be C-fiber firing. Epistemic situations (for Kripke at least) involving a Cartesian, disembodied pain, or C-fibers firing without being *felt* as pain cause doubts about the identity of the two phenomena, since pain is individuated as a *felt sensation*. Kripke feels that the evidence is lacking for an identity theorist unable to explain away his epistemic intuitions, and therefore his epistemic intuitions must spell real trouble for identity theorists about mind and body.

It is left a little unclear why the intuitions about epistemic possibility *do* need to be explained away. The point of the non-descriptional, direct theory of reference is to break the linguistic hold on modality, on essences of things in the world, and to instead let the world speak for itself, provide its own meaning. If *N&N* pried apart a priority from necessity, conceivability from possibility, it also left open the idea that *possible* worlds exist as a subset of *conceivable* worlds, and that furthermore relatively innocuous constitutive or individuative metaphysical theses are all that is needed to so constrain the available worlds. But in the pain case, where there is no a priori metaphysical reason to think that any of the conceivable situations are breaking some such rule, Kripke feels the full force of his conceiving, and leaves open the door

[20] Soames (2006)
[21] Bealer (2002)

for a neo-descriptivist counterrevolution. On the other hand, it is left unproven that the sensation of pain represents a natural kind concept, and therefore that the failure of C-fibers to instantiate pain with metaphysical necessity really is a failure of any sort of reasonable requirement of the concept of pain.[22] As I have said, more on this later.

3. Epistemic Two-Dimensionalism

3.1 Neo-Descriptivism

Oxford University's John Locke Lectures for 1995—the 1994-5 school year's installment in perhaps the most prestigious annual lecture series in analytic philosophy—featured Frank Jackson, who took the chance to primarily engage in metaphilosophy endorsing the by-then largely discredited but previously central analytic philosophical practice known as conceptual analysis. In order to reinstate concepts as primary in a now-orthodox anti-descriptivist metaphysical scene, he saw fit to propose a form of modal two-dimensionalism. Whereas Kripke's lectures that became *N&N* consisted three informal talks tracing a general sort of alternative to a not-explicitly spelled-out descriptivist foe, and then wandering back and forth between diverse counterexamples and potential applications, Jackson in the Locke Lectures that became *From Metaphysics to Ethics: A Defence of Conceptual Analysis*[23] *(FME)* spelled out a well-organized theoretical and then applied treatment

[22] For more on the nuances behind this superficially easy solution to the C-fiber problem, see Fine (2002).
[23] Jackson (1998)

of what might be seen as a revival of Fregean views using Kripkean parts. The work is a substantive, constructive piece of philosophy, and much of what it constructs is a response to subject material left open and unfinished in *N&N*.

To understand what Jackson wants with descriptivism and modality, one can look at his notions of "serious" metaphysics, conceptual analysis, and "folk intuitions" and how they for him point to a common method. *Serious metaphysics* for Jackson aspires to a *complete* account of the world, in something like the way that physics (as we will see later, maybe not contemporary physics) promises to explain the forces of the universe, or at the very least the forces of the universe not including the internal affairs of the human mind or its resultant social and historical products. In order to accommodate entities from a framework other than mathematical microphysics, the Jackson's *physicalist* asserts that those entities are either *reducible to* or *entailed by* the physical account. Such an entailment is usually thought of as some *supervenience* thesis like the following about tallness and individual heights:

> T supervenes on I just in case every possible world with different T has different I,

or differently,

> If I is the same in W_1 as in W_2, then T is the same in both worlds as well,

where T is the list of facts about tallness, and I is the list of heights of individuals.[24] However, for Jackson what the physicalist needs is an *in principle a priori* entailment from the base language to the target (suspect) entities. The motivation for such a requirement can be found in the fact that our concepts encode our information about the world, and therefore the kind of truth-preservation indicative of our *understanding* the world should on this view be truth-preservation by virtue of our concepts. As Jackson puts it,

> Serious metaphysics requires us to address when matters described in one vocabulary are made true by matters described in another. But how could we possibly address this question in the absence of a consideration of when it is right to describe matters in the terms of the various vocabularies? ... It is always open to us to stipulate the situations covered by the various descriptive terms, in which case we address subjects of our stipulation rather than the subjects the titles of our books and papers might naturally lead others to expect us to be addressing.[25]

Consider Kripke's cases of necessary a posteriori statements. In each case — even that dubious case of the Greek star-brothers — empirical information is unquestionably crucial, yet once it is attained and augmented by an essentialist principle of constitution (with the details filled in by science), the given conclusion

[24] See Jackson (1998), pp. 4-5.
[25] Jackson (1998), pp. 41-2.

can, in some sense, be deduced by a priori means. Jackson calls this kind of practice *conceptual analysis*, and points out that the relevant 'if… then' constitutive statements and their ilk can be manufactured independently of experience, with empirical results just serving to fill in the blanks. In formulating such conceptual delineations, Jackson thinks, we consider possible worlds *as actual*.[26] One consequence of this view, of course, is that names are treated as at once both rigid and descriptional; another is that there are must be two intensions for every term or sentence, which can fail to coincide in the case of natural kind terms.

None of this is to say that Kripke's contributions are overlooked; rather they are accommodated as important tools to advance the descriptivist program for philosophy. Jackson (with, we shall see, David Chalmers) thinks that the a posteriori, counterfactual-considering component that dominates post-Kripkean modality is a part of meaning over and above the descriptional part, but he reverses the Kripkean direction of fit by identifying the a priori component as more important for understanding. Jackson refers to meaning based intuitions about the extension of terms in *counterfactual* worlds as *C-intensions*, but thinks that these differ from the meanings of terms based on intuitions about to referents in worlds considered as actual — he calls these *A-intensions* — only in that C-intensions have rigidified themselves based on some property found in this, actual world.

In effect, then, the Kripke cases of the necessary a posteriori — and in particular the a priori essentialist constitutive principles invoked in them — are for Jackson just an (albeit distinguished) example of *"folk intuitions."*[27] By folk

[26] Jackson (1998), p. 49.
[27] It does seem more than a little strange to think of a genius logician as an example of "the folk," even when he is merely stating common sense intuitions.

intuitions, Jackson is referring to the ways in which ordinary people think about, say, properties, particularly of the modal variety. It is this notion that saves neo-descriptivism from regress into purely stipulative talk of the pre-Kripkean analytic variety of necessity. Instead, the game is now to take ordinary intuitions about the reference-fixing of vocabularies and connect different levels of description in a chain of a priori deduction. Whether such a practice is possible even in principle — not to matter in practice — is a matter open to significant debate.

It is important to note just how much work this notion of folk intuitions does for Jackson. For while essentialists would claim that those principles according to which cats are necessarily not robots are *discovered* or *achieved* by scientific experts, Jackson's account seems to reduce those principles to mere convention. Jackson's account must therefore hold that our privileging of the scientists that give us the taxonomy excluding cat-hood from robots, or tiger-hood from lizards, is *in itself* an example of folk intuition! Our societies tend to *defer* to the scientists when we use scientific natural kinds, so that when we mention tigers meaning the scientific 'tiger' kind, we must be thinking something along the lines of, "whatever creatures the scientists who study such creatures would refer to as tigers." For Jackson, it seems to be *folk intuitions all around*, except when they have to be "massaged"[28] into shape, which presumably happens when, for example, our "ordinary conceptions" fail to find any scientifically informed corollaries. Interestingly, by grounding his work in the intuitions of the folk, Jackson saves conceptualism from even social-normative accounts of meaning; he instead puts the burden on the scientists by assuming that they will give us the correct concepts to defer to in the case of natural kinds.

[28] See Jackson (1998), p. 47.

Jacskson's notions of serious metaphysics and understanding are interlinked with his commitment to a canonical form of *physicalism*, which has been passed down from the roots of the naturalist project in analytic philosophy. The idea is that, since the causal structure of the world seems to most closely lend itself to the kind of conceptual entailment we need to really understand things when described in the structural-dynamical language of mathematics microphysics, what we find out about the world had better reduce either by definition or by entailment to that level of description. We can state his thesis thus:

P_D: Every mental truth M is in principle derivable a priori from P, where P is a sentence giving a complete physical description of the world.

So long as the physicists can, at least in principle, produce a "completed physics," it is up to philosophers merely to show how all entities taken to be real are realized or instantiated by that microphysical level.

3.2 (Epistemic)Two-Dimensionalism

Jackson's provides elegant, articulate, and persuasive promotion of Fregean intuitions about sense and their connection with possible and necessary cases, but it is David Chalmers who brought the two-dimensional metaphysical take on modality to the attention of the philosophical mainstream (and beyond). His book *The Conscious Mind (TCM)* had an effect on analytic philosophy of mind akin to what Kripke accomplished in the philosophy of language and metaphysics, challenging the

establishment with counterexamples that over a decade later seem likely to persist in rejecting outright attempts at their dissolution, and shake materialists about the mind to their core. Next chapter, an attempt at spelling out Chalmers' conceivability argument against physicalism about consciousness will be made, and the chapter after that will look at some of the back-and-forth that has occurred since. In order to understand these arguments, we will use Chalmers's brand of two-dimensional modal semantics. It should be noted that there are other forms of the framework, and the scope of this paper is fairly narrow regarding two-dimensionalism in general.[29]

Chalmers follows Kripke semantics by taking as primitive a set of possible worlds.[30] These possible worlds can be thought of as being either metaphysically, conceptually, or epistemologically necessary. Although this conflation strikes some straightaway as wrongheaded (the view certainly has its share of detractors), it is a direct result of his belief that there is no *distinct metaphysical type* of possibility over and above logical possibility:

> [N]one of the cases we have seen give reason to believe that any conceivable *worlds* are impossible. Any worries about the gap between conceivability and possibility apply at the level of statements, not worlds; either we use a statement to misdescribe a conceived world… or we claim that a statement is conceivable without conceiving of a

[29] For an extensive overview of the varieties of 2-D semantics, Chalmers has a cornucopia of such material available at http://consc.net/mmm-papers.html.

[30] At least in his early formulation, possible worlds are the paradigm modal object. Chalmers later seems to prefer the more metaphysically neutral, epistemological notion of a *maximally specific way things might be*.

> world at all… So there seems to be no reason to deny that conceivability about a world implies possibility.[31]

Chalmers already looks dangerously close to denying the main point Kripke gave us: that there are some conceivable worlds that are metaphysically impossible. However, Kripke's point is accommodated by instead defining *two* relations on the set W_i of possible worlds. An expression E on the two-dimensionalist view can be said to have both a *primary* and a *secondary* intension (these also go by the names *1-* and *2-intension)*. A statement S is *primarily possible (1-possible)* if there exists some *centered world* C such that S is satisfied at C (corresponding to Jackon's A-intension); C here is defined as an ordered pair (W, V), where W is a possible world and V is a viewpoint specified by choice of individual, place and time. This is the same thing for Chalmers as considering W as actual. Likewise, 1-necessity can be understood similarly to hold when the primary intension of S holds for all centered worlds C_i. On the other hand, we can define the secondary intension and secondary possibility and necessity according to our familiar Kripkean notions of modality, so that while the 1-intension is evaluated over ordered pairs (counteractual worlds with designated centers), the 2-intension picks out extensions by way of possible worlds considered as counterfactual.

What this amounts to is that terms in the two-dimensionalist system have two meanings like in Jackson's system, roughly corresponding to the Fregean sense and the Kripkean extension. When applied to the case of tigers — a (singular) natural

[31] Chalmers (1996), p. 68.

kind term and therefore a rigid designator — the primary and secondary intensions and their necessities might be spelled out as follows:

> It is a priori (epistemically necessary) that tigers are the large, striped, creatures of our acquaintance.
>
> It is (metaphysically) necessary that tigers are not lizards and are animals.

The first statement expresses an identity about tigers in terms of their *epistemic* or *inferential role.* Tigers are 1-present in every world in which a given creature is qualitatively the same as tigers are in the actual world, including worlds in which they are some kind of lizards or even robots. But they are only 2-present in those worlds where creatures have the biological structure of *Panthera*, and therefore are nowhere 2-present as any of the constituents of the *Reptilia* class. It seems, then, that the two dimensions of meaning, although each necessary, do not always *coincide*. This, the two-dimensionalist must hold, is what Kripke intuited and then made explicit in *N&N*. That the descriptivist is able to explain away the Kripke cases while containing in their notion of the primary modal notions something like the Fregean sense is extremely impressive in its own right, although various foundational issues with the 2-D framework do arise, some of which we will consider later in this paper. Moreover, acceptance of this framework allows for expression of some startling ideas in the metaphysics of mind.

Chapter Two: Conceivability, Zombies, and Reductive Physicalism

4. Pre-Zombie Anti-Physicalist Arguments

4.1 Pain and C-Fibers in 2-D

The two-dimensional neo-descriptivist apparatus can be seen as inspired in large part by the mind-body argument sketched at the end of *N&N*. Recall that Kripke raises two concerns: the appearance of pesky epistemic possibilities in which either pain or C-fiber firing is present without the other and the failure of any sort of metaphysical principle for rejection of such possibilities. The two-dimensionalist picture of the correspondence provides a structured way of thinking about the conceivable situations. If one takes pain to be a rigid designator, as Kripke does, it also seems to be essential to pain that it is a *sensation* or *state*, so unlike the firing of C-fibers, pain does not have a causal or functional role that fixes it modally in any special way, and therefore its primary and secondary intensions can be expected to coincide. It follows that there really *are* possible worlds in which either pain or C-fiber firings happen without their concurrence, and therefore pain and the phenomenal are not identical with C-fiber firings and the physical.[32]

While the Kripke case bears a resemblance to Chalmers's arguments to follow, there are significant points of dissimilarity. For one thing, Kripke himself circa *N&N* has no commitment to a modal structure other than basic metaphysical modality; his epistemic intuitions in the Kripke cases are do not show as metaphysical

[32] Chalmers (1996), p. 19, 146; see also Soames (2006), p. 304-7.

impossible per se but instead just seem to misdescribe the possibilities in question. However, both his and the zombie arguments seek to prove the distinctness of the phenomenal from the physical, using conceivable world counterexamples to the physicalist orthodoxy. Another dissimilarity is that while Chalmers and Jackson want agreement of primary intensions in the form of logical, conceptual, or epistemic supervenience or entailment of one by the other, the Kripkean merely seeks to disprove the mind-body identity theorist and so only considers whether or not the two are *identical*. If all mental states were identical with physical states—if some token-token identity thesis were true—then the supervenience of the mental on the physical would surely be satisfied, but the failure of any given instance of identification does not in and of itself disprove the supervenience thesis in question, as all the physicalist needs to show the neo-descriptivist is that one can in principle provide a priori reasoning for changes in the way mental or phenomenal things are from the way that physical things are.

4.2 Jackson and the Knowledge Argument

There is another line of argument against physicalism known as the *knowledge argument*, which is might be considered even more important than Kripke-style identity cases as a precursor to the two-dimensional argument we will be focusing on. The two famous knowledge arguments are attributed to Thomas Nagel and Jackson; the first considers the subjectivity or phenomenal quality of experience, while the second considers the informational epistemic gap between accounts of experience and experience itself.[33] We consider Jackson's argument in this paper, as

[33] Nagel (1974), Jackson (1986).

reconsidered in a recent article entitled, "Mind and Illusion".[34] Our use of Jackson's argument rather than Nagel's reflects both the similarity of Chalmers's argument to that posed by Nagel and the fact that Jackson, after active participation in the neo- vs. anti-descriptivist debate, changed his mind on the validity of his own original claims as counterevidence to physicalism. So for those readers who take Jackson's argument to be less than fully convincing: fear not, for he now agrees. The alternative picture he paints does not (or so he thinks) work wholesale against two-dimensionalism or neo-descriptivism; in fact, his formulation of a response is motivated by a desire to save neo-descriptivism from the threat of dualism. We will consider his misgivings about the knowledge argument at the beginning of next chapter, in order to set the bar for the objections against Chalmers. First, let us look at Jackson's argument and consider in what respects it *is* convincing.

Imagine if you will a room that is entirely black and white. Every surface is grayscale, and all light sources provide white light; there are no windows in the room. Insert into this thought experiment an incredibly brilliant and knowledgeable scientist, Mary, to whom the same black-and-white rules apply; she has permanently been painted white from head to toe, and is dressed only in black. The details of the situation are bizarre and largely irrelevant, since what is important to the account is that Mary has never had what we might call color experiences; everything that Mary experiences is either black, white, or gray. On the other hand, she knows all there is to know about the physical natures of objects, the neurophysiology of humans, and how they meet in human color perceptions, as she has had ample time to study black and white books and the black and white television screen in her otherwise boring room.

[34] Jackson (2003).

Still—and this is the intuition that matters—it does not seem as if Mary has enough to *understand* or *know* "what it is like" to see colors. That is to say, some information is missing from the physical account of color experiences, and that information cannot even in principle be derived from the physical account. Since the traditional formulation of physicalism requires that everything follow a priori from the physical, the failure of deducibility in the case of color experiences seems to carry strong count-physicalist intuitions.

5. Conceivability and Possibility of Zombies

5.1 Various Conceivables

To appreciate the full force and sophistication of Chalmers's modal argument against physicalism and, more generally, materialism, one must first understand the distinctions he makes between various forms of possibility, and to which forms he claims conceptual access. After all, as Kripke pointed out, we should not expect conceivability to tell us everything there is to know about what is possible. The question Chalmers must answer is what, if anything, we *can* learn from what we find to be conceivable. In order for his answer to be intelligible at all, we must accept three binary dichotomies he introduces for making explicit the kind of conceivability he has in mind. Whereas Kripke seems to utilize about one and a half types of conceivability — the oftentimes-aposteriori-informed type corresponding to metaphysically possible worlds, and the sometimes-deficient type corresponding to

epistemic or evidential situations — Chalmers recognizes eight categories; he distinguishes between primary and secondary, positive and negative, and prima facie and ideal conceivability.

We should by now be ready for Chalmers's use of primary and secondary conceivability. Primary conceivability, roughly put, is that space of our intuitions which have as their objects primary or 1-possibilities. To conceive of S in the primary sense is to conceive of something *as actual*, by imagining a centered possible world in which S is the case. Conceiving in the primary epistemic space is a priori, as it considers how the world could have *turned out* to be, and sets aside the facts of our actual world.[35] In contrast, secondary conceivability refers to conceiving of states of affairs as counterfactual, and is thus a posteriori in that it is informed by the way things actually are. This 2-dimension, of which Kripke makes much use for evidence of a posteriori necessity, is closed to scenarios that contradict known necessary truths, such as that the chemical structure of water is as it is. So while it is taken to be primarily conceivable that water possesses a chemical structure other than H_2O — after all, on Twin Earth 'water' is XYZ — it is not secondarily conceivable because the 2-intension is fixed by the actual state of affairs.

But the failure of conceivability to reach possibility is not always due to a failure to keep distinct the two modal dimensions. For one thing, Chalmers wants to distinguish what he calls *prima facie* conceivability — conceivability on "first appearances" — from *ideal* conceivability.[36] This helps give conceiving a respectability that, say, imagination might lack; the idea is that while one can have

[35] Yablo (2002) offers a good discussion of this treatment of the "primary" modal dimension.
[36] Chalmers (2002), p. 147.

mistaken intuitions about what is possible — whether in the counteractual or counterfactual sense — ideal rational reflection would presumably work out the mistakes and leave something more legitimate. Of course, humans with all of their neuropsychological quirks may not seem like very good candidates for ideal conceivers, but that is a topic we will leave for later. The other distinction Chalmers considers salient is between *positive* conceivability, which takes the form of some sort of representation, construction, or other faculty resembling imagination, and *negative* conceivability, which just represents a lack of intuition *against* possibility. This distinction concerns the quality of evidence; one can *see* that something is possible, or one can merely fail to see that it is impossible and therefore *suppose* that it is possible. While the three distinctions might not be the only ways of organizing notions of conceivability, and are far from uncontroversial, they suffice to make clear what Chalmers has in mind with the important first premise of his argument against physicalism (and other forms of materialism).

5.2 Zombie Twins and Zombie Worlds

> The most obvious way (although not the only way) to investigate the logical supervenience of consciousness is to consider the logical possibility of a *zombie*: someone or something physically identical to me (or to any other conscious being), but lacking conscious experiences altogether.[37]

[37] Chalmers (1996), p.94.

So goes the passage from *The Conscious Mind: In Search of a Fundamental Theory* that opens our debate. Chalmers has a deep, enduring intuition that the consciousness we have as feeling, first-person, experiencing subjective persons, could conceivably fail to be present in complete physical duplicates of us. The argument takes many forms; one can imagine a zombie twin of oneself, with the "lights out," or that everyone but oneself in the world is a zombie. But it is important to notice from the outset what kind of conceivability Chalmers describes himself as employing. He is no dummy about current science, and knows that consciousness tends to accompany physical human brains very reliably and regularly, so that consciousness could even be said to be physically or naturally necessary, but the epistemic possibilities he has in mind consider worlds as actual, so that they do not have to be worlds that share the same natural laws the actual world. That is to say, Chalmers has intuitions that zombies and indeed entire zombie worlds are *primarily conceivable*, even though it remains to be shown whether they are secondarily so. It just so happens that the neo-descriptivist account physicalism includes a priori entailment or deducibility of phenomenal (and all other) truths about the world from a complete list of all physical truths, and 1-intensions are taken to correspond to what is accessible a priori. So the 1-possibility of zombies spells (for Chalmers, definitive) trouble for physicalism, making room for his particular brand of dualism or at least some non-materialist alternative. We will examine what it is about two-dimensional metaphysics that makes this the case in the next subsection.

There are gaps that must be filled, however, on the way even to 1-possibility. Chalmers begins with a seemingly mild thesis about what might suffice to get us there, which we will call $CP_{WMR\ (Weak\ Modal\ Rationalism)}$:

CP_{WMR}: Ideal primary positive conceivability entails primary possibility.[38]

An obvious problem with this is that even if Chalmers by some miracle of metaphysical intuition could conceive with unbridled veridical clarity of a zombie world (we'll from now on adopt the convention Z-world and extend to analogous uses), there is no way that mere mortals could share in his certainty. And it is to his credit that he does not attempt to provide a candidate for an ideal positive account beyond his assertions of the depth of certainty he has in his Z-intuitions. However, he does contend that the following, stronger thesis will do, and as well that we can extend his weaker *modal rationalism* with relative ease to accommodate the stronger version.

CP_{SMR}: Ideal primary negative conceivability entails primary possibility.[39]

If he is right about the SMR conceivability-possibility thesis, the burden of proof has shifted; Chalmers has outsourced the ideality of his conceiving. For if no one can

[38] This thesis shows up as (1) in Chalmers (2002), p. 171.
[39] Ibid.

provide a rational argument for why a Z-world fails to be 1-possible, then that—on the basis of CP_{SMR}—is evidence for its possibility in fact. In effect, under strong modal rationalism, the *persistence* of Chalmers's views in the face of rational debate lends them their strongest support.

In showing why negative conceivability might suffice, Chalmers finds that candidates for membership in the "twilight zone" between positive and negative conceivability tend to belong to two categories: inscrutabilities and open inconceivabilities.[40] Both categories can be explained in terms of a *canonical description* or *complete qualitative description* D that is assumed to in principle accompany each centered, counteractual world (so that for each world W and viewpoint V, there is in principle a description D). Inscrutabilities are then truths that cannot be derived or deduced from a complete account of the way the world is or could be. On the other hand, open inconceivabilities are truths that cannot be conceived of but on the other hand also cannot be ruled out (or at least not on an a priori basis). Chalmers postulates that both classes are empty. His reason for ruling out inscrutabilities relies on his candidates for actual and zombie canonical descriptions $D_@$ and D_Z. On the other hand, open inconceivabilities are less easily explained away. All that Chalmers can offer against such undiscoverable twilight-truths is to point out that just so long as they do not include *impossibilities*, they do not endanger the CP_{SMR} thesis.

With some form of modal rationalism in hand, we can give a logical analysis of Chalmers's argument:[41]

---

[40] Chalmers (2002), p. 174.

[41] I found http://consc.net/papers/2dargument.html to be at least as helpful as Chalmers (2002) in formulating brief but clear version of the zombie argument.

(i) A Z-world W_Z—where a complete physical description P is true and some phenomenal truth Q fails to hold—is ideally, negatively 1-conceivable, and therefore ideally, positively 1-conceivable.

(ii) If a world W is ideally, positively 1-conceivable, then W is 1-possible.

(iii) If W_Z is 1-possible, then materialism fails.

(iv) *Conclusion*: From (i)-(iii), materialism is false.

Notice the characterization of Chalmers's opponent as a "materialist." This term extends past what Jackson and Chalmers take to be the physicalist's commitment to a priori deducibility between levels of description, as we will see from some so called "Type-B" opponents. One can (and these days, most do!) object to the zombie argument without endorsing such a claim, and without identifying oneself as a physicalist.

5.4 A Neo-Descriptivist Taxonomy of Materialists

Ever since *TCM*, Chalmers has done an exemplary job of anticipating the objections likely to be raised against his conceivability arguments. In particular he identifies several varieties of materialism from which to expect counterattacks. Chalmers distinguishes between what he calls *type-A materialists* (T_A henceforth), who dispute premise (i), and *type-B* (T_B) *materialists* who are willing to accept at least the 1-conceivability of Z-worlds while still maintaining a monist, physical-privileging ontology. Other types of materialist arguments are brought up by

Chalmers, but they are summarily disregarded. Chief among these is the *type-C* (T_C) materialists, who maintain that the epistemic gap between conceivability and possibility is in principle bridgeable, even if the limitations of human epistemic access to modal properties make such a task practically impossible (or impossible for the foreseeable future). As we will see next chapter, Chalmers sees them as defending an "inherently unstable" point of view that is best seen as collapsing into some other type of commitment. Either there is no real epistemic or explanatory problem — the type-A stance that Chalmers derides as an outright denial of the very fact that a "hard problem of consciousness" presents itself — or physics as so far developed does not *need* to *conceptually or logically* entail the phenomenal facts (exactly the point characteristic of the type-B view) — or else there must be some kind of *intrinsic* rather than just *structural-dynamic* physical properties responsible for the correspondence between the physical and the phenomenal. This last way of thinking leads to either *type-D dualism* (T_D), *type-E dualism* (T_E), or *type-F monism* (T_F), which correspond to (interactionist) dualism, epiphenomenalism, and panpsychism (or even the even stranger panprotopsychism) .[42] Briefly, T_D acknowledges that the mental causally interacts with the physical even while being metaphysically disctinct, T_E sees the mind as causally ineffective, and T_F has phenomenal properties or some intrinsic precursor playing a fundamental role in all reality.

Of the two straw men that Chalmers sees as offering metaphysically robust (but for him counterintuitive alternatives), a type-A position T_A is most immediately in need of response. This is because the general way these arguments go is to deny the non-reductionists their argument. In the next section we will consider a recent T_A

[42] Chalmers (2003), p. 23.

or T_C reconstrual that Jackson offers of his original knowledge argument, which has the upshot of massaging away the supposed problem that Mary raised. As per usual in $T_{A/C}$, and as Chalmers is careful to point out about the form that those arguments tend to take, Jackson's account *rejects* a compelling intuition about the phenomenal in order to *defend* some kind of functionalism about experience. Then we will consider the ways such potential twilight examples as mathematical statements or skeptical doubts about Chalmers's ability as a conceiver have led to much nuance, qualification, and even backsliding in the non-reductive two-dimensionalist account of consciousness without ever definitively putting the argument to rest by way of a compelling, consistent, coherent, and defensible rebuttal of the zombie intuitions.

More promising in making clear the widely held metaphysical modal intuitions against the zombie argument are the family of concerns that Chalmers places in the T_B category. These arguments tend to assume more various guises than T_A or T_C, yet have the comparative advantage of a common rallying point: at least some Kripke cases of necessary a posteriori metaphysical truths. Chalmers, of course, alleges that the neo-descriptivist account suffices to adequately accommodate not just Kripke's separation of the epistemic and the metaphysical dimensions of natural kind meanings but also of the logician's concerns with about epistemic situations, while giving no reason that relevant a posteriori knowledge could not inform a "neutral semantics" which would at the limit succeed in providing, say, water with coincidental 1- and 2-intensions, but which—the dualist must maintain—cannot succeed in coinciding intensions for phenomenal experience short of drastically

reimagining either our ontology about minds or the fundamental nature of our physics.

To foreshadow where the field of battle will be found between the modal anti-descriptivist and the neo-descriptivist, let us briefly consider what the T_B member, supporter, or defender must require of a Z-world argument. For a T_B account of consciousness to be made true at some W, the physical facts must merely *2-entail* the phenomenal truths at W. So the 1-possibility of W_Z is not *prima facie* enough, without additional possibility-link premise (iia.) and the substitution of an a posteriori version (iiip.) for premise (iii).

(i) A Z-world W_Z—where a complete physical description P is true and some phenomenal truth Q fails to hold—is ideally, negatively 1-conceivable, and therefore ideally, positively 1-conceivable.

(ii) If a world W is ideally, positively 1-conceivable, then W is 1-possible.

a. If W is 1-possible, then W is 2-possible.

(iiip.) ***If W_Z is 2-possible, then materialism fails.***

(iv) *Conclusion*: From (i)-(iii), materialism is false.

We will see that while Chalmers treats the two additional premises as following relatively harmlessly from one fundamental assumption of the neo-descriptivist apparatus—a deflationist view about what has been called *strong, deep, or metaphysical* (not to be confused with mere *natural*) necessity—that assumption sits

on shaky ground, since a priori deduction is not even the rule in physics, much less the rest of science.

Chapter Three: Chalmers and His Objectors

6. Jackson v. Jackson: The Knowledge Argument Reversed

6.1 A Late-Blooming Physicalist

Frank Jackson for decades sounded the descriptivist battle-cry against the sloppy blind faith of the physicalist majority in philosophy of mind. Methodologically, the situation in analytic philosophy as a whole looked soft; it was supposed that the successes in contemporary natural sciences would continue indefinitely. Ideas holding that everything in our universe could be assimilated by a scientifically-informed conception of *nature*, with mathematical physics sufficing as a solid grounding, were popular to the point of delusion. Swept under the proverbial rug were uncomfortable metaphysical explanatory gaps regarding such entities as phenomenal, subjective experiences or intentional states, just as in the social sciences the ontological status of economies and cultures and language communities seemed distinctly non-physical in certain important respects. Within that milieu, Jackson's knowledge argument against physicalism can be seen as representing a certain amount of rigor from physicalism. It wasn't necessarily that physicalists were wrong in fact, but instead that they were less than thorough in the practice of providing justification for their beliefs in the face of serious intuitions against those beliefs.

Thus, the import of Jackson's recent reversal regarding the outcome of tensions arising between scientific promises for physicalism and the epistemological

intuitions encapsulated in his account of Mary's room can be seen as sort of explanatory victory for physicalism.

> Most contemporary philosophers given a choice between going with science and going with intuitions, go with science. Although I once dissented from the majority, I have capitulated and now see the interesting issue as being where the arguments from the intuitions against physicalism—the arguments that seem so compelling—go wrong.[43]

Jackson qua defender of the knowledge argument is not here admitting defeat; rather, Jackson all along has seen the debate as involving two strong arguments, and here is looking for a resolution on the side of science. It is important that we recognize what Jackson's argument is decidedly not: an argument against the required-in-principle a priori deducibility characteristic of his neo-descriptivism; as far as he is concerned, that metaphysical apparatus is well-grounded and well-defended. In addition, the argument does provide *positive* instructions about at least the general way to conceive of the knowledge argument in order to leave open the truth of Jackson's physicalist thesis P_D, and an attempt at tracing how the greatest challenge to P_D — consciousness — might be approached by human epistemic means.

Matching against the materialist straw men set up by Chalmers, one sees that the first of Jackson's two commitments entail that his account be of the $T_{A/C}$ variety, as it is committed to the lack of any permanent epistemic gap between the physical

[43] Jackson (2003), p. 1.

and the phenomenal. And as a positive conceivability argument for physicalism, it could be argued that the following account is deserving of as much evidential recognition as does the original anti-physicalist knowledge argument about Mary. If one were to be able to find so convincing a counterargument for all anti-derivability arguments, one could save the physicalist entirely. But that is not what Jackson is doing here. By providing what is really a sophisticated T_C argument from representationalism, Jackson undermines the entire $T_{D/E/F}$ set of concerns about conscious experience but leaves alone the metaphysical apparatus forcing their application in the first place.

6.2 Mary Re-represented

We now return to Mary in her room. There are truths that the anti-physicalist "tub thumping" from Jackson and others have previously located outside her reach. One of these is *acquaintance* with redness. However, Jackson thinks, such concerns about her lack of adequate acquaintance-based knowledge have mistaken the *intensional* or *representational* awareness of *experiencing the representation of something as red* with the *instantiated, relational* property of *redness*. Whether talking about the redness of objects—which we can call "redness"—or redness in our visual field—which we can follow Jackson in calling "redness*",

> [K]nowing what it is like is knowing about redness or redness*, and the knowledge argument is an argument to the conclusion that Mary does not know about redness or redness* -- that is, about the property

> we are, according to the picture, acquainted with when we sense red… Intensionalism tells us that there is no such property.[44]

What the common property of redness refers to by Jackson's lights is not the property of being red, but instead some shared sense of in what epistemic states it would count to see red and shared experience of those representations. So Mary is not precluded from finding out what *redness is* prior to her release.

Instead, what the representationalist has left to show is how Mary can know what it is like to *sense* red prior to her in fact sensing it. But Jackson, following representationalists like John McDowell, sees sensory experience as an inherently *conceptual* activity.[45] Here is where, (perversely, I think), that opens the door for his A-intensions to come in and save the day. If there is a semantically neutral way to spell out the conditions of application for the concepts represented by experience, then it is in principle possible to deduce the same truths from physical facts as from representational experience. All that Mary really gains upon release is *direct acquaintance* with the representational states that constitute experiences of redness, in that she can now remember what it is like to have them. But if representational states consist of, say, inextricable and immediate richness and the playing of a functional role, then such information must be available to Mary, for it is surely structural-dynamic based on what we now know about the differential perceptual field. As Jackson says, "we get the phenomenology for free."[46]

---

[44] Ibid., p. 15.
[45] McDowell (1994) provides a clear neo-Kantian story of how and why this must be the case.
[46] Jackson (2003), p. 26.

6.3 A Disappointing Success

On the one hand, the representationalist counterexample to anti-physicalism is tremendously successful. Most other counter-arguments fail to come close to instilling the same level of doubt about the strength or clarity of Chalmers's zombie intuitions. Furthermore, none of the objections offers a similarly positive and competing account of phenomenal experience as anchored directly in the physical facts. The strength of Jackson's argument comes from an ingenious synthesis of more or less the same kind of neo-descriptivist metaphysical commitments that Chalmers holds with a thoroughly conceptual account of experience. In general, objectors to the conceivability of Z-worlds endorse either some cases of non-descriptional content (natural kinds) or some cases of non-conceptual experience. Jackson's judicious withholding of both allow him to provide the level of detail necessary for bringing serious doubt to bear on the epistemic access to zombie intuitions.

Still, there is a sense in which Jackson has missed the point with his counterargument. By locating the important meaning in question in A-intensions there is a sense in which he is not exactly following science at all. Natural kind terms given by science seem to inform not just an epistemology about objectivity, but metaphysics as well; science tells us how the world is organized, what is in some sense allowed and what things there are. If a priori access really is what is required for us to make sense of the world, the question still remains how we are to escape McDowell's miserable image of "frictionless spinning in the void."[47]

McDowell gets his friction—his external constraint on the human "space of reasons"—by way of the necessarily acquisitive process of *second nature*, or our

[47] McDowell (1994).

Bildung (upbringing, playing the role of self-realization). However, McDowell in so naming our membership in the rational human community is careful to accord respect to the (first) nature of our world, which is a realm of causal rather than rational rules. It is likely that an account like McDowell's is too beholden to common sense notions of "nature" to do the job of answering the nitpicking needs of the modal metaphysical issues we are concerned with. But by placing epistemology in the driver's seat, Jackson conflates the categories of causality and rationality, and therefore makes problematic (or at least leaves unexplained) our *access* to objectivity except as mediated by conceptual and logical analysis. So while he has provided evidence *against* one explanatory gap, Jackson qua the neo-descriptivist keeps alive the metaphysical commitments that make explanatory gaps a problem in the first place, and thus leaves it open to Chalmers to merely reinforce his intuitions that something is left out (the phenomenal, first-person "what it is like") by such evidence.

7. Objections to the Conceivability-Possibility Theses

7.0 Welcome to the Zombie Wars

There are two main targets available for those who dispute Chalmers's Z-arguments: the CP theses, and the 2-D modal neo-descriptivist framework. The first class is composed mostly of Chalmers's Type-A and Type-C straw men, to whom the CP theses are for various reasons less than fully convincing, but so-called Type-B materialists and others like them have also seen fit to challenge various steps in the

CP links. We'll consider these cases first, both because they require answering in order for the Z-arguments to get off of the ground at all, and because Chalmers has developed definitive replies to all of them. In observing the results of the $T_{A/C}$ battles, a way of replying even to Jackson's ingenious representationalist argument can be found. It seems as if Chalmers has earned himself the lasting impact that he has had.

We consider the so-called T_B arguments separately, in Section 8. It will become apparent that these objections are mostly different in kind from the materialist arguments considered before, in that they to varying degrees dispute the argumentative framework on which Chalmers relies, rather than just denying him his premises. These objections rely either on a posteriori necessity, or on doubts about a priori deducibility as a requirement for metaphysics; Chalmers characterizes these objections either as stemming from problematic conceptions of "metaphysical"[48] or "strong"[49] necessities. I will conclude by suggesting that, rather than merely refuting Chalmers's CP theses, intuitions along this line should cause us to rethink the neo-descriptivist system in general as unsuitable for science-sensitive metaphysics. What we are left with is the task of reformulating talk of modality in order to better avoid the kind of tensions that have preoccupied and confounded analytic philosophers and cognitive scientists since the publishing of *TCM* thirteen years ago.

7.1 Representationalism, Functionalism, and the Twilight Zone

[48] This is the term used in Chalmers (1996), and his problem with it is that it differs from his identification of *logical* and *metaphysical* necessity as describing basically the same things. We'll see that this stems from his view of modality as built upon *epistemic* primitives.

[49] Much of Chalmers (2002) is spent disputing the legitimacy of this kind of notion; an updated treatment shows up in Chalmers, "The Two-Dimensional Argument Against Materialism," which can be found at consc.net/papers/2dargument.html.

To what extent does Chalmers leave himself open for refutation? Let's consider again the general form of the arguments he gives:

(i) A Z-world W_Z—where a complete physical description D_P is true and some phenomenal truth Q fails to hold—is ideally, negatively 1-conceivable, and therefore ideally, positively 1-conceivable.

(ii) If a world W is ideally, positively 1-conceivable, then W is 1-possible.

(iii) If W_Z is 1-possible, then materialism fails.

(iv) *Conclusion*: From (i)-(iii), materialism is false.

Since (iii) corresponds to the a priori physicalist thesis P_D, and since neo-descriptivism seems to require P_D, we will for the time being leave the premise alone, revisiting it in §8. Likewise, one can trace a sort of informal proof of (ii) from what we know about Chalmers's modal framework: Suppose you can give a complete qualitative description D_Z of W_Z, not too stringent a requirement of an ideal positive conception. Then, as long as D_Z includes the indexical truths V_Z situating you in that world, it seems as if you have a description of a centered possible world, which for Chalmers corresponds to 1-possibility. There doesn't seem to be any room for mistake at this point; if the world is not in fact 1-possible, then there must be an incoherence in D_Z, in which case you have not practiced ideal rational reflection. One could use a version of Chalmers's notion of *open inconceivabilities* to hold that there is an inconceivable yet true property of W_Z that therefore escapes determination by any D. However, in order to defeat the Z-arguments, this truth M_Z would have to

either entail W_Z's impossibility—which seems paradoxical—or entail Q, in which case (i) would be amended that W_Z has D_P true, and Q mysteriously true. The retreat to M_Z is self-defeating in a sense, turning a mere explanatory gap in the case of consciousness into a completely unsolvable mystery by way of an ad hoc act of faith. It seems, then, that as in the case of (iii), one must dispute Chalmers's conception of modality in order to find (ii) questionable, which is again the point of §8.

With (ii) and (iii) safe for the time being, we can now consider the various objections that have been offered to (i). Preliminarily, we notice that (i) itself can be broken down into the following steps, each with its corresponding justificatory source:

(1) W_Z, which shares the complete physical truth D_P but not some phenomenal truth Q with the actual world $W_@$, is prima facie 1-conceivable. (Taken for granted by Chalmers and everyone else who shares his intuitions)

(2) W_Z is ideally negatively primarily 1-conceivable. (Supported by ongoing refutation of counterarguments against W_Z's conceivability, as well as a failure of materialists to provide Q's a priori deduction from D_P)

(3) W_Z is ideally positively 1-conceivable. (Lack of any relevant inscrutabilities or open inconceivabilities in the ideal epistemic case)

Three categories of objection become available within (i)'s sub-arguments: **(1)** an objector can question Z-conceivability prima facie. This is one of the typical T_A routes, which we will talk about below; **(2)** proponents of materialism can voice their optimism that between prima facie and ideal conception, there will be found the

necessary requisite connection between D_P and Q. For Chalmers calls this T_C materialism does not succeed without reduction to one of either the T_A or T_B views or capitulation to the $T_{D/E/F}$ cause he champions ; **(3)** the connection between negative and positive conceivability can be questioned. These objections are curious in that they deal with the nature of what D_Z can establish, but do not promise much in the way of positive reasons to believe in materialism on their own.

In the second category, we find Chalmers's T_C straw men. What is common to these accounts is the notion that Chalmers—this has consequences as well for any other such human Z-conceivers—is not conceiving hardly or correctly enough. Perhaps the most eloquent and concise of these arguments comes from Sara Worley.[50] Worley points out that on the two-dimensionalist account, 1-possibility is constituted ultimately by conceptual *coherence*, and Chalmers has certainly offered a coherent enough account of Z-worlds that incoherence given his story is not obviously available. However, Worley finds that Chalmers does not offer good reason to believe that such incoherence could not be found, given a complete account of the possible world in question. If the complete qualitative description that Chalmers has in mind for each world in determining 1-intensions is even something that it makes sense to think about, our lack of such a description prevents us from predicting the satisfaction or non-satisfaction of coherence. That is to say, if we are speaking precisely, it is not at all obvious from the mere *appearance* of coherent conceivability of W_Z that W_Z is in fact (if there is a fact of the matter) coherently conceivable.

We have already seen Jackson's positive version of such an account. Because of a failure to use the correct concepts in describing phenomenal experience —

[50] Worley (2003).

representational concepts, Jackson thinks that neo-descriptivist anti-physicalist accounts misunderstand consciousness as being conceptually contingent. However, when one experience as functionally and epistemically representational, one can understand how in principle one could understand all of the *facts* about experience. What this boils down to for Jackson is that Chalmers's *"what it is like*" should really be thought of as "*what representational states are immediate*." Adequate physical and functional specification of the roles and contents of the representational states, as well as of the relation of immediacy, gives one the facts of the matter. That accounts of perception, for instance, tend not to be this fine-grained does not pose a problem for the T_C stance; humans are of course limited in practice, but what matters is the availability of an adequately descriptive account *in principle*.

Category (3) objections make use of the fact that the main ammunition available to Chalmers is the lack of decisive counterexamples to the Z-arguments. He challenges the materialist to provide a non-problematic reason to believe that zombies or Z-worlds are *not* conceivable. The longer they take, the smugger he is entitled to be, and so attempts have been made to erode that smugness by offering doubts that the negative-positive gap can be ignored. Examples of inscrutabilities might be mathematical truths, which seem a priori after their proof but are conceptually indeterminate in the meantime (and are thus open inconceivablities). There might be certain Gödel sentences, for example, that are not a priori deducible but nevertheless are determinately true.[51] Other candidates for membership in Chalmers's *twilight zone* between negative and positive conceivability are macroscopic, vague, moral, and metaphysical statements. As Ned Block and Robert Stalnaker argue, it is not

[51] Chalmers (2002), p.

always possible to provide a priori deduction from a more fundamental to a higher order of description, even in established science.[52] That is, one can know the molecular structures that accompany genes, without knowing the functional laws of how genes function. So D_P should not be expected to settle the Q question one way or another. Likewise, vague truths seem to sometimes escape determination by a complete description. As far as moral and metaphysical truths go, some think that one could plausibly infer a truth M that is either moral or metaphysical as well as its negation ~M from a canonical description D.[53]

Although Chalmers sees (3) objections as serious, and makes numerous efforts to defeat them, as a class they cannot be said to belong to $T_{A,B,C}$. They do not tend to determine a definite form of materialism, so much as a set of doubts about the work that the complete description D can do for Chalmers in solidifying his conceptions. The currency they cash in on is well-described by Stephen Yablo, who pointed out that conceivability can fall prey to issues of *determinacy*, so that we can say that a statement S and its negation can both be left indeterminate, so that D implies both ~det(S) and ~det(~S).[54] A strong reading of (3) objections, with the corresponding assertion of a robust set of twilight zone truths, might entail the following kind of concern: it is *indeterminate* whether or not zombies are conceivable, because there is something about the notion of consciousness that is left up to interpretation. From such a suggestion, one could come up with three readings:

[52] Block and Stalnaker (1999).
[53] Yablo (2002).
[54] See Yablo (1993).

T_A: Consciousness isn't really anything above and beyond functional or physical truths, and therefore the very idea of D_P without Q trades on a mistake about the term consciousness.

T_B: Consciousness, as a natural process, isn't reducible to description in the a priori deducible way that neo-descriptivists want, but that doesn't change the fact that it follows necessarily from physical truths in the correct way for materialism to hold.

T_C: Chalmers thinks that Z-worlds are conceivable because he does not yet have the correct concepts with which to carry out the a priori deduction of consciousness from the material.

The T_B reading we will discuss next chapter, while the T_C reading seems like just a specific case of Worley's point above, in which the proper future physical concepts will entail the incoherence or inconceivability of Z-scenarios after all. On the other hand, looking at the T_A version of the vagueness objection alerts us to the privileged status of (1) and (3- T_A) objections. Usually referred to as *eliminativism* about consciousness, T_A stances can be hard to argue with, since they amount to a refusal either to admit to sharing the intuitions that Chalmers feels, or to acknowledge the contentfulness or correctness of one of the concepts on which the conceiving is based. Under eliminativism, consciousness simply does not exist, at least in the manner in which Chalmers sees it; even if it does appear to exist, this can written off as an illusion due to the conflict between the way in which we gather data about phenomenal consciousness — through first-person experience of it — and the way in

which we gather physical and functional data — through third-person publicly-accessible observation.

Of course, most materialists of even the T_A brand do not choose to disown talk of consciousness *per se*. Instead, what is more common is a retreat to what Chalmers calls "the easy problem": the functional or behavioral roles that are often referred to when describing consciousness. An analytic functionalist might suppose that the awareness, access to representational content, or certain kinds of beliefs or dispositions — this last view is the behaviorist special case — are all that there is *to* consciousness. Part of the appeal of this strategy is that it seems as if scientists stand every chance of explaining the functional aspects of consciousness, and so if there is nothing over and above the functional, then there is no problem whatsoever, just puzzles to be solved in providing the functional explanations.

Finally, one can adopt a different reading of the (3) form of objection to (i), which I'll call **Type-W** for "Wait and see." One can choose to withhold assent to ideal positive Z-conceivability without committing oneself to the current or future ability to conceptualize or explain consciousness using the resources of materialism. In order for this argument to work, it must free itself from the physicalist constraint of a priori deduction; Chapter 4 will seek to accomplish this. Furthermore, the Type-W view must remain agnostic regarding whether the concepts with which consciousness is to eventually explained with turn out to be material, mental, or of some as of yet unknown kind. Of course, materialists have largely stayed away from this kind of viewpoint not only because of their commitments to deduction or even because of T_B commitments to essentialism, but instead for fear of giving Chalmers what he really

wants, which is an openness to expanding our conception of the physical beyond our current structural-dynamical resources that makes naturalists queasy. At the end of this paper I will briefly return to a kind of Type-W view.

7.2 Pro-CP Responses from Chalmers

For Chalmers, T_A materialists are the most straightforward to dispatch with. Most of us are just not willing to just let go of our intuitions of consciousness; those intuitions are manifest in our experiences. Of blunt denialists about consciousness, he can ask, "are there any compelling *arguments* for the claim that on reflection, explaining functions explains everything?"[55] To arguments comparing consciousness to previously mysterious processes like "life" that turned out do dissolve into functional investigations, Chalmers retorts that the analogies do not fit, since we have a different sort of acquaintance with consciousness, and that ongoing successes in exploring the cognitive functional roles referred to under the name 'consciousness' do not seem to dissolve the first person phenomenal data in an adequate way.

If representation is thought to play the role of intermediary between the physical and the phenomenal, Chalmers thinks that it must assume one of two insufficient shapes. *Functional* representation, which tends to be explained in functional terms, serves only to explain behavior, rather than phenomenal consciousness. On the other hand, *phenomenal* representation, in which representation is equates with the having of conscious experiences, tends to leave unexplained how and why functions entail that kind of representation as opposed to any other. One can explain the *beliefs* we have about consciousness, but this just

[55] Chalmers (2003)

leaves out the first-person justification that we feel for those beliefs and so seems counterintuitive. Furthermore, it seems like my zombie twin would too believe that he was consciousness, and so without the first-person evidence being brought to bear in justifying those beliefs, such an account cannot be good enough.

As far as T_C materialist accounts go, Chalmers is willing to admit of their appeal, but thinks that their distinction from the other five accounts is unstable. Either consciousness is a functional concept (T_A), materialism does not need implication from physics (T_B), or since our current structural-dynamical style of physics does not provide the resources to explain consciousness, future success will inevitably result in our expansion of physics ($T_{D/E/F}$).

8. **Objections to Neo-Descriptivist Two-Dimensionalism**

8.1 Type-B Materialist Objections

Let's turn to those positions which are held by T_B objectors and their ilk. The objections raised by a T_B account are meant to cut into two basic tenets of Chalmers's two-dimensional neo-descriptivist framework, corresponding to rejection of either (ii) or (iii) from §7.1. As we shall see, the playing field changes in new ways when either the movement from conceivability to possibility or 1-possibility to 2-possibility of zombies is challenged. Note that the two distinctions are not identical; while some seek to defeat the move from modal epistemology to modal metaphysics from within

Chalmers's framework, others see the framework itself as a problem (I am one of the latter).

Starting from within the framework, a T_B materialist could hold that, just as is true of the Kripke cases, the primary and secondary intensions of consciousness or of mental terms or states do not agree, so that while the primary possibility of a zombie is fine, it does not assure secondary possibility. Alternatively, it could be held that while Z-worlds in question may *satisfy* the physical description D_P, they do not *verify* it. That is to say, some *deep* or *intrinsic* physical properties are being left by superficial qualitative agreement with D_P.

However, those examples seem stretched or contrived in ways that they need not be if one just denies Chalmers his neo-descriptivist framework. This can be done by nixing the requirement of *deducibility* as opposed to mere *metaphysical entailment* of consciousness by the physical. Block and Stalnaker accomplish this by showing the lack of conceptual analysis often available for high-level (macro-) properties. That is, two-dimensionalism presupposes an unfair account of content in natural terms.[56] Rather, background empirical knowledge serves to situate the use of natural kind terms. Bealer objects that it is unnecessary and counterproductive to suppose that a term can be used two different modal purposes; he conjectures that two-dimensionalism only succeeds when it is not needed.[57]

By far the most driven attack against neo-descriptivist two-dimensionalism has come from Soames. In *Reference and Description (R&D)* Soames declares that in interpreting natural kind terms as rigidified *descriptions,* two-dimensionalism has

[56] See Block and Stalnaker (1999).
[57] Bealer (2002)

missed the intuitions that Kripke laid bare, in which names and natural kind terms refer directly. But neither are natural kind terms *indexicals*, since they do not merely *point* to objects but also invoke the *essences* of those objects in counterfactual situations. He correctly infers that by making intuitions about primary intensions person- and circumstance-specific, Chalmers and Jackson must allow that different people when using the same natural kind terms might not be able to share the same (primary) meaning.

8.2 Two-Dimensionalist Defenses

Responses to T_B positions over the years have helped shape the development of the accounts that Chalmers gives of his own positions, but those positions have really just hardened and perhaps retreated a little. For instance, Chalmers is not willing to let the primary and secondary intensions of conscious states differ, since personal intuitions just are the relevant evidence for consciousness. In addition, by assuming that a complete qualitative physical description of the world would be only structural-dynamical in the fashion of mechanistic or computational physicalism, he can shoo talk of *intrinsic* (or *emergent)* physical properties into the T_F (or T_A) category.

Epistemological or linguistic complaints about two-dimensionalism such as those offered by Bealer or Soames have been disappointing. For Chalmers has been able to just claim that they misunderstand him: primary intensions only include the implicit ability to identify extensions in possible cases, not full-fledged conceptual fluency, and the two intensions only represent the full content of a term, rather than

two competing uses, he holds. In fact, Soames has gotten called out for himself espousing a kind of two-dimensionalism due to his own convoluted metaphysical and epistemological account of modality.[58]

The fact that Chalmers responded to Block and Stalnaker with Jackson helping says much about the perceived force of the objections the a priori deductivism intrinsic to the two-dimensional apparatus.[59] If macro-properties truly do not follow from micro-facts, then the force of the Z-arguments loses its weight. Furthermore, Chalmers has admitted since *TCM* was published that the need for primary possibility or a priori deducibility would fall in the face of either of two sorts of objections: taking *identities* as modal primitives, or espousing some *deep*, *strong*, metaphysical form of necessity (these can be seen as two sides of the same essentialist coin). How one can properly introduce such notions without resorting to deductivism or some kind of linguistic conceptualism on the one hand, or unfounded conventionalism or mysterianism on the other has provided a philosophical puzzle, and therefore convinced Chalmers that the choices he has identified must be the only choices. Chapters 4 and 5 will suggest that by grounding our modal talk in our practical successes through scientific practices, we can anchor necessity not in anything deep or strong beyond our own efforts but nevertheless deflate the apparent gridlock between anti-descriptivism and neo-descriptivism.

[58] See for instance Chalmers, http://consc.net/papers/soames2d.pdf. It's just one episode in a face-to-face squabble so I'll leave it out of the bibliography.
[59] See Chalmers and Jackson (2001).

Chapter Four: Metaphysics and Natural Laws

9. Who Cares?

9.1 Lowe's Objection

Last chapter's survey of the Zombie Wars was intended to highlight a few key points, on the way to our central point of contention. One thing we saw is that if one accepts the terms of Chalmers's argument, he ends up fairly successful, in that no exceedingly damning objections have shown up to defeat the zombie cases. It is for that reason that even to this day some still mention the controversy as ongoing. Another thing that we saw is that, in objecting to the terms of the zombie arguments, the most intuitively reasonable option is to deny Chalmers the extension of his CP-principle to the secondary dimension of possibility; furthermore, such a denial needs either a positing of a deep, strong, metaphysical type of necessity, or a rejection of a priori deduction as a requirement for the kind of necessity indicative of metaphysical identity and supervenience; these two choices come out to very much the same thing. And, since a rejection of linguistic or conceptual analysis as the only acceptable grounds for metaphysical identity and supervenience rejects the core tenet of neo-descriptivism, we have made our way via the zombie wars to the center of the proverbial anti-descriptivist/neo-descriptivist battlefield, with the choice between dualism and materialism at stake.

But perhaps we are misled about the seeming importance of this battle; E.J. Lowe seems to think so. In his review of Soames's *R&D*, what I will henceforth call Lowe's Objection (LO) asserts that the debate between Soames and scientific essentialists on one side and Chalmers and Jackson on the other over the legitimacy of neo-descriptivism for natural kind semantics may really be a debate over *just* semantics, without further defense of the metaphysical import of Kripkean cases in the first place.[60] Given what we just saw in the last paragraph, his objection would mean that the difference between materialists and panpsychists (those who believe that the mental is *funda*mental in reality) could just be a difference in how they *talk about* things! Going back to the Kripke cases, however, it turns out that he is very close to being correct.

Consider the *N&N*-inspired CROBOTS. The reason that we were to determine that the hypothetical CROBOTS would not in fact be cats is that in our actual world is that cats happen to be animals (more specifically, the *Felinae* biological subfamily is a member of the *Animalia* kingdom), and since it is constitutive of biological natural kinds that their members are picked out by their biological structure in every possible world in which they exist, the non-animal CROBOTS do not qualify as cats. However, while Kripke and Putnam were making waves with their direct theory of reference and externalist theory of meaning, Salmon was producing *Reference and Essence,* which points out that the type of constitutive principles used in these example—that if a natural kind has a certain natural structure in one possible world, then it must have it in every world—is itself an analytic premise required to get the

[60] Lowe (2007)

essentialist arguments off the ground.[61] LO urges us to go back to this often-overlooked point before we get ahead of ourselves and try to start adjudicating one way or the other between the anti-descriptivist essentialist and neo-descriptivist two-dimensionalist versions of the Kripke cases.

9.2 Apparent Metaphysical Contingency

To give LO its proper due, we will first note some obvious qualms one could have (and many do have) with too blasé an adoption of constitutive premises as holding with *metaphysically necessity*. In contrast with "old school" notions of metaphysical possibility (or necessity) as logical compatibility (or entailment), Kit Fine points out that contemporary philosophers use it more often to mean "the sense of necessity that obtains in virtue of the identity of things (broadly conceived)."[62] However, when one opens up one's metaphysical modality—as Kripke surely did—beyond tight questions of logic, one runs into difficulties in determining just how broadly *identity* should be so conceived.

For instance, numerous arguments over the years have shown that metaphysical possibility must outstretch physical necessity. To see this, just consider a possible world in which one or more physical *constants* (such as the gravitational constant or the speed of light) differ in quantity from the actual world, while the physical *laws* have the same structure and interrelationships. If such a world is to be deemed metaphysically *impossible*, it is not clear for what reason; after all, all of Kripke's natural kinds could still be individuated, were the speed of light to differ.

---

[61] See Salmon (1981).
[62] Fine (2002)

But then this principle can be applied to Kripke cases as well, and the easiest way to see that is to consider the case of tigers.

What is the constitutive biological structure of tigers? Presumably it has to do with their evolutionary history and certain mechanisms of their development, or some such characteristics. But their evolutionary history could presumably have differed in possible worlds with even the exact same physical laws as ours, such that different mechanisms could have evolved; after all, mutations happen within species all the time. Furthermore, consider what we took to be the one sure metaphysical necessity of big and small cats alike: that they are animals, rather than non-living robots. Is there, however, any *metaphysical* necessities that determine whether a thing is living or not? More worryingly, are there really any *necessities* at all to individuate a given species, given that evolution is always at work? These kinds of questions abound relative to the essentialist picture, and are not particular to biological species. As a chemical case, we can see that there may be no determinate matter of fact of exactly at what point one element becomes another as a result of alpha-particle bombardment, yet remember that molecular structure is supposed to be *constitutive* of chemical kinds. So LO turns out to be not such a trivial stumbling block for the essentialists, and therefore their Kripke cases may be of no use after all.[63]

9.3 Anti-Descriptivism, Essentialism, and Respect for Science

Yet there is a strong intuition among anti-descriptivists that the Kripke cases *are* important for metaphysics. LO prompted a response from Soames himself, who

[63] Lowe's own system makes use crucial use of universals, rather than necessities, as modal primitives. While worthy of study, it is certainly no less cumbersome than systems of either the essentialists or the descriptivists. See Lowe (2006).

not surprisingly offers a quite excellent case for why it should be *no* damnation of anti-descriptivism that metaphysical consequences do not arise by themselves from the non-metaphysical premises of Kripke's direct theory of reference.

> Lowe's failure to find the philosophical significance of semantic anti-descriptvism comes from looking in the wrong place. Its importance lies in *expanding* the range of metaphysical hypotheses to be taken seriously, not in limiting debate by proving metaphysical theorems from non-metaphysical premises.[64]

This fits in adequately with what seems to be a central facet of Soames's work and the driving force behind his commitment to the continuation of the anti-descriptivist project: conserving the significant opportunity Kripke afforded essentialism to make itself legitimate after years of positivist and Quinean oppression.

However, the serious attention given neo-descriptivism over the last decade and a half suggests failure by anti-descriptivists to fully realize the potential of their opportunities for expression by fully convincing all involved of the tenability of the basic essentialist position, at least so far as it is supposed to apply to consciousness. While there are a number of diligent and talented champions offering positive, constructive accounts of what goes by the terms "new" or "scientific" essentialism,[65] and while later we will consider objections to these accounts, I think that most of such accounts share a principle found in the end of Soames's reply to Lowe.

---

[64] Soames (2007), p. 34.
[65] Ellis (2002) and Bird (2007) come to mind.

> [W]e assume that it is an essential property of a substance that instances of it have the molecular structure that they do…Being true, [Water = H_20] is, therefore, necessary. Since knowing the proposition it expresses requires knowing of a certain substance that its instances have a particular chemical structure, [Water = H_20] is knowable only a posteriori. Note, there is no attempt here to derive a metaphysical truth about essence from non-metaphysical premises. On the contrary, the point of the exercise is to show how compelling, and widely accepted, examples of the necessary a posteriori can be *explained* using plausible essentialist assumptions. The point of [*R&D*] is that they can't be explained without them.[66]

I take these last sentences to be the two main arguments common to scientific essentialism, which work in pincer-like concert:

E1: Essentialist constitutive premises about natural kinds are plausible, and offer explanations of Kripke cases and other similar intuitions about natural kinds.

E2: Other explanations either are not plausible or cannot explain (as well) such intuitions.

In the next section, I will consider an account that disputes E2, and therefore offers (I think) a plausible alternative to essentialism.

[66] Soames (2007), p. 36.

10. Stability and Counterfactual-Based Necessity

10.1 What are Natural Laws?

Philosopher of science Marc Lange tends to depart from the majority of professional philosophers specializing in the metaphysics of science in that he tends to start with the *uses* of (rather than intuitions about) metaphysically meaningful notions from *within* science, and work up pictures that fit those uses. So, in considering so-called laws of nature, which an essentialist would just hold are only the statements that hold with metaphysical necessity given to us by science, Lange instead solves the traditional problem of distinguishing laws by noticing that

> If a scientist takes some claim to be a law statement, then she uses it to perform various functions that she does not regard accidental generalizations as able to perform. Although philosophers have yet to state these functions precisely, they have long believed that these functions involve counterfactual conditionals, scientific explanations, and inductive confirmation.[67]

None of this is the least bit controversial, but what happens next surely is. Lange has built a definition of laws and necessities on the relationships he sees as holding between counterfactuals, laws, and scientific explanations, using induction as a guide to what such relationships require.

[67] Lange (1995), p. 435.

To enter his account we'll begin by considering an objection offered by Alex Rosenberg, that *functional* explanations in biology are really more like placeholders for complete explanations to come in the form of molecular biology (which can be worked up from physics), history, and what he takes to be *the* natural law of biology, the principle of natural selection.[68] In a sense, Rosenberg is correct; things could always have been different as far as, say, the biological structure of tigers is concerned, while still keeping the physical and evolutionary laws the same. Lange points out that this view stems from a something like a physical supervenience thesis, which Lange calls Nomic Preservation:

> *Nomic Preservation (NP)*: *g* is physically necessary if and only if *g* would still have held had *p* obtained, for every *p* that is logically consistent with every physical necessity.[69]

I think that such a principle is implicit in descriptivism and in physicalism. For in order to a priori deduce all facts from the history of the physical facts together with the physical laws, we need such preservation. And the anti-descriptive physicalists really only seem to differ from the neo-descriptivists in that they do not always agree that the relevant necessitation can be reduced to the conceptual level. But Lange is interested in what happens if one takes the schema and applies it to functional biology, by reformulating *NP*, which he sees could result in two forms:

---

[68] Rosenberg (2001).
[69] Lange (2004), p. 97.

> *g* is one of the laws *of functional biology* (or a logical consequence of those laws) if and only if *g* would still have held had *p* obtained, for every *p* that is logically consistent with
>
> *NP'* : the laws *of physics*
>
> *NP''* : the laws *of functional biology*[70]

It is clear that the received view only considers *NP'* to be acceptable if the laws are to be considered physically necessary, and it is this consideration that compels Rosenberg to assert that indeed there are no laws of functional biology apart from the physical-historical explanations they promise. But Lange, pointing out that functional biology does seem to presuppose such laws, formulates law-hood in such a way as to allow autonomy; instead of building up from physical-causal historical necessitation, he relies on a notion of holistic *stability*.

10.2 Stability

For laws to support counterfactuals in the correct way, Lange has established, they must "all still have held under any counterfactual supposition that is logically consistent with the laws. No accident is always preserved under all of these suppositions."[71] To formalize what he means, Lange uses a notion of a *stable set.*[72] A stable set is a logically closed (i.e. contains all of its logical consequents) set made up of true statements such that each member is consistent with each counterfactual

[70] Ibid.
[71] Ibid., p. 98.
[72] This notion and others of Lange's receive excellent, technical discussions in Lange (2000), and those discussions have been advanced to some extent in subsequent works such as Lange (2005), Lange (2007); Lange (2004) provides a level of detail appropriate for the needs this paper, as we are withholding from confronting the logic directly.

situation or supposition consistent with all of the rest of the members of the set. It is (trivially) true, then, that the "set of all truths" is stable. But for any other, non-maximal set of statements to be stable, or as Lange puts it to be "invariant under as broad a range of counterfactual suppositions as they *could* logically be,"[73] expresses the way in which they can be said to hold of *necessity*, he thinks.

Grounding talk of necessity in invariance with regard to subjunctive antecedent suppositions themselves rather than the other way around is a radical departure from the modal metaphysical orthodoxy, and is unquestionably controversial among metaphysicians. In particular, the question of "truth-makers," which is to say whether the counterfactuals are responsible for the truth of laws, or vice versa, is unclear from preliminary discussion, whereas essentialists can fill the role of truth-maker with the essences of things. Yet the account does accommodate a great deal; for instance, consider so-called "logical necessities." Since these statements hold *no matter what* (non-contradictory) counterfactual supposition is entertained, they can be identified as just the smallest stable set.[74] Indeed, by varying the range of counterfactuals to be considered, Lange's method promises to provide accounts of every genuine form of necessity.

10.3 The Autonomy of Scientific Disciplines

What does the notion of stability promise about functional biology, and of the biological structural properties that can be said to necessarily hold of tigers? Lange points out that the *contents* of a stable set in question help determine *which*

[73] Lange (2004), p.99.
[74] Lange (2007)

counterfactuals that set must be invariant with regards to. For instance, consider the example of species. Just so long as the functional rules about what counts as a reptile or lizard are coherent with a rule delineating tiger-hood, for example, we can imagine a taxonomic framework that does disallow tiger-lizards. That such a law is not strictly *physically* or *evolutionarily* necessary need not bother us, since our set does not prima facie need to include *all* natural laws in order to possess stability (at least over a relevant time period; more on this will follow in §11.1. Although laws can show up in different levels of detail corresponding to different stable sets, the sets *do* form a hierarchy of containment. At the same time, explanations between levels need not be deductive or even supervening.[75]

Macro-explanations from disciplines like functional biology can offer things that the microphysical rules leave out, by operating at a coarser-grained level of counterfactual consideration. Just as that the physical constants have the values they do may not be *cosomologically necessary,* but is considered *physically necessary* for most purposes, that tigers have the biological features by which they are individuated can be considered *functionally necessary* without being *evolutionarily necessary*. Consider an example attributed to Putnam: that a square peg does not fit into a certain round hole differs not "whether the peg consists of molecules, or continuous rigid substance, or whatever."[76] That a continuous rigid substance itself is physically *impossible* – and therefore *metaphysically* impossible according to scientific essentialism – is of no importance to the explanation; the peg just does not fit because of size and shape! If one starts looking for this kind of practically-oriented style of

[75] Lange offers a proof of this in Lange (2005), p. 422.
[76] Putnam (1975), quoted by Lange (2004), p. 109.

explanation in science, it turns out that it is the rule, rather than the exception; this should be evidence against both deductivism *and* reductive physicalism.

Chapter Five: Scientific Understanding and Physicalism

11. Salience and Inductive Strategies

11.1 Salience and Best Inductive Strategies

Left out in the above is an explanation of how the range of counterfactual invariance is to be determined. The answer is something like this: which counterfactuals to be considered are just those counterfactuals that a given scientific research program finds *salient*, which is to say it depends on the interests of the scientists in that field. A functional biological account of species can admit of laws because functional biologists are less concerned with evolution over long time scales. Salience, or the perceived prominence of certain indicators, can also explain why it is that some statements with many exceptions (such as the albinism of big cats) can be accorded a degree of necessity, while others must remain generalizations:

> [W]e are more willing to say 'The lion is tawny', while knowing that white lions occur occasionally, than to say 'The Witch's Hat mushroom is nonpoisonous', while knowing that poisonous Witch's Hat mushrooms occur occasionally, because our tolerance for eating poisonous mushrooms is lower than our tolerance for making inaccurate predictions of a lion's color.[77]

[77] Lange (1995), p. 440.

By itself, this idea of salience promises advantages over descriptivist and essentialist views in that it allows one to see in what way scientists and society at large are partially responsible for the formulation of laws, where Jackson's example was content to leave the meaning of terms as "folk intuitions." However, it does not yet explain how our rational understanding is to get its "friction" in the McDowellian sense.

It is not just the concerns of scientists that determine what is given endorsement of necessity. Statements receive inductive confirmation when they correctly predict future cases, and so laws can be thought of as the results of *inductive strategies.*

> A law statement is a reliable inferential rule yielded directly by an inductive strategy in the set of inductive strategies that is best for us to have been pursuing. Whether a set of inductive strategies is the best for us to have been pursuing depends on the range of evidence that has been accessible to us.[78]

This is important for two reasons. First of all, it is an *accomplishment* to achieve inferential reliability, especially given the interests driving research, and therefore we one can provide grounding for endorsement of a kind of "metaphysical" necessity; relative to the kind of data with which we interact, a given law is necessary in that it is reliable relevant to our practices. But more importantly, thinking along these lines gives reason why metaphysical modality would do *better* to be grounded in the

[78] Lange (1995), p. 447.

unfamiliar *counterfactual*s which Lange employs: scientists are in general *adept* at isolating and testing the relevant counterfactual situations with which to arrive at *reliable inference rules*. Put another way, Lange's conception of law-likeness actually stands a chance at building on something that scientists can provide, a novel contribution to the literature to be sure.

11.2 Essentialism and Practical Worlds

Well, okay, you maybe thinking, but what does all of that tell us about CROBOTS? It may be hard to think of cats as having *essences* in the same sense as that meant by anti-descriptivists if the kinds of practical scientific concerns voiced above preclude any objective (in that old-school naïve sense of "free of subjective human concerns") account of the natural kinds. My response is this: let's instead call the properties individuated by Langean laws *schmessences*. The hypothetical CROBOTS which we encountered in §2.4 are still necessarily non-cats when relying on schmessences as guides because it is *schmessential* to felines in any biological taxonomy that exists, that they are part of either an evolutionary lineage or at least a roughly determinate (perhaps not natural, that is up for debate) kind that we fuzzily call "living." So, any apparent cat that turns out instead to be a CROBOT, no matter how exquisitely crafted a CROBOT it might be, will fail to be a cat (here we are assuming for simplicity's sake that CROBOTS are missing enough "signs of life" that it is not indeterminate whether or not they can be deemed animals themselves).

Although we may have dispatched with the CROBOTS, essentialists are likely not to feel safe reading this. For "*schmessentialism"* takes away the metaphysical

entitlement that anti-descriptivists rely on *essences* to provide. Instead of solving the "location" problem for natural kinds with in either the descriptivist in principle a priori deducibility (whatever that is) or essentialist trust in deep necessity or dispositions somehow *discovered* to hold in objects, we can use informed judgments of salience combined with the best inductive strategies available given the limited epistemic resources available to scientists, so that rather than appealing to *essences* of natural kinds left wide open to charges of mere *convention*, the law-governed objects and properties of *nature* can be thought of as *constituted* in part by the requirements of the practical needs of scientific research projects (which are always funded, run, and executed by actual rather than merely possible people). That these projects themselves admit of a certain degree of fallibility should be found certainly no more worrisome than that the intuitions of Jackson's "folk" need a certain amount of "limited massaging";[79] conveniently, Jackson's version has philosophers practicing conceptual analysis in armchairs doing the massaging whereas the pragmatic version relies on such dingy professionals as engineers, technologists, and other undeniably aposteriori-oriented practitioners to maintain a correct image of the world in which we live. So much the better, I say, for *schmessentialism*.

[79] See Jackson (1998), p. 47.

12. Understanding Physicalism

12.1 Deductivist Physicalism and the Unity of Science

Now that we have explored a pragmatic re-imagining of the relationship between counterfactual situations, necessities, and scientific explanation, we are in the position to see a priori deduction for what it really is: a holdover from 1950's logical positivism. As Nancy Cartwright has alluded to, most proponents of the Unity of Science movement spearheaded by Vienna Circle member Otto Von Neurath (the social scientist and civil servant of the group) shared with him the hope of unifying the scientific disciplines while dismissing his concerns for the practical import of philosophy.[80] In particular, exuberance for the prospects of logical analysis exemplified by the work of Rudolph Carnap led to the emergence of a sort of layer-cake view of the sciences, which has survived throughout the 20th century in analytic philosophy.

This is due, I hold, in large part to a misunderstanding about scientific understanding. If, following Frege, we hold the sense and therefore the informative content of language to be inside the head, then it does seem to follow that the formation of a "serious" metaphysics will require some kind of (at least "in principle") *entailment* in a conceptual or logical sense between various facts about the world. In this light, Jackson's definition of physicalism from *FME*, that

P_D: Every mental truth M is in principle derivable a priori from P, where P is a sentence giving a complete physical description of the world,

[80] See Cartwright (1999), pp.5-7.

seems entirely reasonable, and is even required. But two related problems seem to me insurmountable to this position. The first problem is that, analogous to what Wilfred Sellars showed about the lack of any non-conceptual "Given" in which to ground talk of intentionality,[81] there is no microphysical "Given" from which construct a bottom-up logical lattice of scientific deduction. And even if there *were* to be some "language of God" in which the unified *completed physics* upon which Jackson's picture relies could predict the entire universe's happenings, the second problem is that, as Lange has shown, such an account would miss out on the important explanatory efficacy that can be gleaned from macro-explanatory reliable inferences employing even microphysically *impossible* antecedents.

12.2 Silberstein on Emergence

One can take the lessons from §9.2 combined with misgivings about a priori deductivism, and attempt to reformulate the physicalist thesis. Some have tried to defend physical*ism* without recourse to the circular definition of explanation by the physical by defining physicalism as the thesis that the fundamental laws are physical and *not* mental, and defining the *mental* as a kind of implicit natural kind.[82] Although Lange's account of necessity can accommodate some laws as fundamental,[83] it does not appear that they can bear the metaphysical weight required, since they do not entail the results of the macro-laws (the meta-laws might, but that's another story…).

[81] Sellars (1997)
[82] See Nimtz and
[83] See Lange (2007)

I will now turn to an interesting (and radical) take on consciousness as a kind of *emergence.*

Silberstein uses the term 'emergence' to claim that "a whole is 'something more than the sum of its parts,' or has properties that cannot be understood in terms of the properties of the parts."[84] Thus, it is the historical competitor to reductionism, either reduction of the ontological "nothing-but" sort or the epistemological "component parts plus relations" preference for understanding or both. Emergentism makes a claim surely odious to "serious metaphysicians" like Jackson, that

> the best understanding of complex systems must be sought at the level of the structure, behavior and laws of the whole system and that science may require a plurality of theories (different theories for different domains) to acquire the greatest predictive/explanatory power and the deepest understanding.[85]

Silberstein recognizes two flavors of ontological emergence that should by now be familiar to us: if something cannot be eliminated from our ontology (a case of *non-elimination*) then we must accommodate it as *real*, and if something resists identification with a lower-level thing (a case of *non-identity*) then it must be acknowledged as distinct. However, he also introduces two additional shapes that emergence can take: *mereological emergence* or *holism* holds of objects that have properties not determined by those of their constituent parts, while *nomological*

[84] Silberstein (2001)
[85] Ibid, pp. 67-8.

emergence applies to entities governed by laws that do not follow from the laws governing their constituent parts.[86] It seems then that Lange is an advocate of at least nomological emergence in science. We will see next that Silberstein gives reason to believe that at least mereological emergence is quite common in contemporary physics.

Two examples from current physical understanding seem to support mereological emergence: quantum entanglement and symmetry breaking. According to quantum mechanics, it appears that properties of some systems do not break down to the properties of the fundamental parts. In fact, contrary to supervenience, the compound system can determine the state of the constituent parts in a *top-down* manner! Symmetry breaking, on the other hand, is the interaction with the Higgs field in the Standard Model (arguably as close as we have to a theory of everything) in order to explain differences between fundamental particles. If our theories of the smallest building blocks of physical matter have such level-crossings, why is it that we are so committed to either the reduction of or a causally efficacious role for consciousness?

The supposed choice between physicalism and property dualism now seems like a mistake. On the one hand, physicalism is actually less intuitive than most materialist philosophers pretend; it withholds any causal power from mental actions, which seems at least practically foolhardy. There is no overwhelming evidence for physicalism, and on the other hand, physics when considered with its own examples of emergence does not seem on track to satisfying the deductive-hierarchical or

[86] Ibid, p. 72; Silberstein addresses epistemological emergence as well, but we are for the purposes of this paper concerned primarily with the metaphysical and therefore ontological consequences of emergence.

completeness formulations that Jackson ascribes to it. And on the other hand, Chalmers's argument against the emergence of the conscious from the non-conscious really seems to come down to two premises — that theoretical deduction fails and that it is the hallmark of scientific understanding. But really, it does seem like much of a hallmark, when one considers points made by Lange or Cartwright[87] showing the failure of supervenience in the scientific hierarchy.

> Just to be clear, my point here is not that I prefer physicalism over fundamental dualism or panexperientialism when it comes to an account of the fundamental ingredients of the world. Rather, barring full-blooded idealism, I see no *empirically useful* difference between any of these three views when it comes to the nature of the fundamental ingredients.[88]

I think in honor of David Chalmers we ought to call this kind of view Type-W monism.

13. Conclusion

Getting to the kind of "Type-W" view about consciousness espoused by Silberstein and (really only a few) others has taken a lot of work. This is because of philosophical baggage attached to notions like natural laws – originally thought to be

[87] See for instance Cartwright (1999).
[88] Silbertein (2001), p. 90.

legislated or enforced by a deity – and necessity, as well as antiquated conceptions of how science works. Given that metaphysics exists along a sort of historically developing continuum, works by Kripke, Jackson, and Chalmers each deserve credit for helping to clear out obsolete ideas and make way for new interpretations. However, assumptions left from each provide impediments to the kind of scientifically respectable metaphysics that is being furthered work like that done by Lange and Silberstein, work rooted more strongly in and showing more promise for potential contributions to the empirical successes and practical realities of science as practiced in the actual world.

Bibliography

Bird, Alexander (2007). *Nature's Metaphysics: Laws and Properties*. Oxford: Oxford University Press.

Block, Ned, and Stalnaker, Robert (1999). "Conceptual Analysis, Dualism, and the Explanatory Gap." *Philosophical Review* 108 (1): 1-46.

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

- (2002). "Does Conceivability Entail Possibility?" in *Conceivability and Possibility,* Gendler, Tamar Szabó and Hawthorne, John eds. Oxford: Clarendon Press.

- (2003). "Consciousness and its Place in Nature," in *The Blackwell Guide to the Philosophy of Mind*, Stich, Stephen and Warfield, Ted eds. Oxford: Blackwell Publishing.

- (2006). "The Foundations of Two-Dimensional Semantics" in *Two-Dimensional Semantics*, García-Carpintero, Manuel and Macià, Josep eds. Oxford: Clarendon Press.

Chalmers, David, and Jackson, Frank (2001). "Conceptual Analysis and Reductive Explanation." *Philosophical Review* 110 (3): 315-360.

Ellis, Brian (2003). *The Philosophy of Nature: A Guide to the New Essentialism*. Montreal: McGill-Queen's University Press.

Fine, Kit (2002). "The Varieties of Necessity" in *Conceivability and Possibility*, Gendler, Tamar Szabó and Hawthorne, John eds. Oxford: Clarendon Press.

Jackson, Frank (1986). "What Mary Didn't Know." *Journal of Philosophy*, 83: 291-295.

- (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press.
- (2003). "Mind and Illusion," in *Minds and Persons: Royal Institute of Philosophy Supplement 53*, O'Hear, Anthony, ed. Cambridge: Cambridge University Press.

Kripke, Saul (1959). "A Completeness Theorem in Modal Logic." *Journal of Symbolic Logic*, 24: 1-14.

- (1963). "Semantical Considerations on Modal Logic." *Acta Philosophica Fennica*, 16: 83-94.
- (1980). *Naming and Necessity.* Cambridge: Harvard University Press.

Lange, Marc (1995). "Are There Natural Laws Concerning Particular Biological Species?" *Journal of Philosophy*, 92 (8): 430-451.

- (2000). *Natural Laws in Scientific Practice*. Oxford: Oxford University Press.
- (2004). "The Autonomy of Functional Biology: A Reply to Rosenberg." *Biology and Philosophy* 19 (1): 93-109.
- (2005). "Laws and Their Stability." *Synthese* 144 (3). 415-432.

- (2007). “Laws and Meta-Laws of Nature,” *Harvard Review of Philosophy*, 15: 21-36.

Lowe, E.J. (2006) *The Four-Color Ontology: A Metaphysical Foundation for Natural Science*. Oxford: Clarendon Press.

- (2007). “Does the Descriptivist/Anti-Descriptivist Debate Have Any Philosophical Significance?” *Philosophical Books*, 48 (1): 27-33.

McDowell, John (1994). *Mind and World.* Cambridge: Harvard University Press.

Nagel, Thomas (1974). “On What It’s Like to be a Bat.” *The Philosophical Review*, 83: 435-450.

Putnam, Hilary (1973). “Meaning and Reference.” *Journal of Philosophy*, 70 (8): 699-711.

- (1995). “Philosophy and Our Mental Life,” in *Philosophical Papers, Volume 2: Mind, Language, and Reality*. Cambridge: Cambridge University Press.

Rosenberg, Alex (2001). “How is Biological Explanation Possible?” *British Journal for the Philosophy of Science*, 52 (4): 735-760.

Salmon, Nathan (1981). *Reference and Essence*. Princeton: Princeton University Press.

Silberstein, Michael (2001). “Converging on Emergence,” in *The Emergence of Consciousness*, Freeman, Anthony, ed. Imprint Academic.

Soames, Scott (2005). *Reference and Description*. Princeton: Princeton University Press.

- (2006). “The Philosophical Significance of the Kripkean *Necessary Aposteriori*”, *Philosophical Issues*, 16: 288-309.
- (2007). “The Substance and Significance of the Dispute over Two-Dimensionalism.” *Philosophical Books*, 48 (1): 34-49.

Worley, Sara (2003). “Conceivability, Possibility, and Physicalism.” *Analysis*, 63 (1): 15-23.

Yablo, Stephen (1993). “Is Conceivability a Guide to Possibility?” *Philosophy and Phenomenological Research* 53 (1): 1-42.

- (2002). “Coulda, Woulda, Shoulda,” in *Conceivability and Possibility*, Gendler, Tamar Szabó and Hawthorne, John eds. Oxford: Clarendon Press.