# Emergence and Cosmic Hermeneutics

by

Jeffrey Watson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2013 by the
Graduate Supervisory Committee:

Bernard W. Kobes, Chair
Nestor Pinillos
Terence Horgan
Steven Reynolds

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

Emergentism offers a promising compromise in the philosophy of mind between Cartesian substance dualism and reductivistic physicalism. The ontological emergentist holds that conscious mental phenomena supervene on physical phenomena, but that they have a nature over and above the physical. However, emergentist views have been subjected to a variety of powerful objections: they are alleged to be self-contradictory, incompatible with mental causation, justified by unreliable intuitions, and in conflict with our contemporary scientific understanding of the world. I defend the emergentist position against these objections. I clarify the concepts of supervenience and of ontological novelty in a way that ensures the emergentist position is coherent, while remaining distinct from physicalism and traditional dualism. Making note of the equivocal way in which the concept of sufficiency is used in Jaegwon Kim's arguments against emergent mental causation, I argue that downward causation does not entail widespread overdetermination. I argue that considerations of ideal *a priori* deducibility from some physical base, or "Cosmic Hermeneutics", will not themselves provide answers to where the cuts in the structure of nature lie. Instead, I propose reconsidering the question of Cosmic Hermeneutics in terms of which cognitive resources would be required for the ideal reasoner to perform the deduction. Lastly, I respond to the objection that emergence in the philosophy of mind is in conflict with our contemporary scientific understanding of the world. I suggest that a kind of weak ontological emergence is a viable form of explanation in many fields, and discuss current applications of emergence in biology, sociology, and the study of complex systems.

To Jason Watson, my brother

ACKNOWLEDGMENTS

This work owes its existence, from initial inspiration through final completion, to the guidance of my dissertation advisor, Bernard W. Kobes. I am sincerely grateful for his direction, suggestions, criticism, encouragement, supervision, and dedication over and above anything which could be expected. To the extent novel ideas are present in this work, they predominantly emerged through my conversations with him.

I am also very grateful for conversations on this work with the other members of my committee, Terry Horgan, Steven Reynolds, and N. Ángel Pinillos, and for their comments and feedback on these ideas.

This work was greatly enhanced by conversations while visiting the Australian National University in March of 2013, especially discussions with David Chalmers and Daniel Stoljar, as well as comments on topics in this work from Philip Pettit, Jonathan Simon, Frank Jackson, David Wiens, Colin Klein, Wolfgang Schwartz, Leon Leontyev, Alex Sandgren, Alma Barner, Daniel Gregory, Colin Klein, and those in the audience at the ANU Philosophy Society. I am very grateful to all of them for lending me an ear.

I should also mention my appreciation for the contributions of members of the "Emergence and Cosmic Hermeneutics" seminar at Arizona State University in the Spring of 2012, including Thomas Fournier and Paul Henne; the audience at the 2013 Central Division meeting of the American Philosophical Association, especially comments by Jennifer Matey; and conversations over the years with Billy Dunaway, Sarah Bernstein, Joshua Mugg, Andrew Bailey, Christian Ryan Lee, Joe Hedger,

Gerald Marsh, and many others, which pointed me to recent literature which I would have overlooked otherwise.

This work was made possible in part by a doctoral research grant from the Arizona Board of Regents. It was assisted by grants for travel provided by the American Philosophical Association, and by Arizona State University through the Graduate and Professional Students Association, the Graduate College, and the School of Historical, Philosophical, and Religious Studies.

Finally, on a personal note, I also must acknowledge my appreciation for the support of my parents, for the loyalty of my friend David, and, most especially, for the love and immense patience of my wife Johanna, who made what might otherwise have been a very stressful time joyful.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

# INTRODUCTION: FOUR CHALLENGES TO ONTOLOGICAL EMERGENCE

## §1. What is Emergence?

### 1.1 The Ontological Landscape

Desert landscapes have a certain austere beauty to them. Some people may have the temperament to appreciate this beauty more than others. For instance, Willard Van Orman Quine had this sort of aesthetic sense, at least for *ontological* deserts.[1] There is a certain simplicity and purity to a barren ontology, every feature an aggregation of the same particular grains of sand. Unburdened by commitments to multiple kinds of beings, physicalism enjoys freedom under the wide open skies. Unpretentious and humble, its perspective remains firmly rooted to the earth.

Compare this to a different sort of landscape: the view one gets from the top of a mountain, looking down from the snowy heights onto the dry valleys below. From this view, the world seems to be a place of radical differences: above and below, the higher realm and the lower, deep within the caves beneath the mountain, and up on top of it, nearly touching the sun. For a certain temperament, the view from on top of a mountain can be very exciting. A person who imports this aesthetic taste into metaphysics will begin trading in an ontology of oppositions. Perhaps it flatters the thinker too much, but dualism feels more uplifting than the desert floor.

---

FOOTNOTES

[1] For example, Quine confesses this "taste for desert landscapes", in *Pursuit of Truth* (1980) as well as in "On What there Is" (1968)

A third temperament is inclined neither towards radical division nor barren simplicity, but finds its focus fixed upon the many layers of in-between. At the bottom level lies the desert, and snow capped peaks arise at the top, but much of what's interesting in the world occurs at the intermediate: the forests, the plains, the grassy hills. Reality is more like a *Sky Island* – a range of mountains surrounded by lowlands, where the landscape begins with desert cacti, transitions upward into the bushy chaparral, extends higher into forests of pine and oak, and near the summit packs in thick forests of Douglas fir. Here one finds drastically different kinds of things emerging at different levels of elevation, each thriving only at its own level and at no other. In a Sky Island, many different kinds of things exist, and they exist atop each other, in many different layers. This ontological landscape seems, to those so inclined by this particular temperament, both the more accurate to the world we find ourselves in, and the more beautiful.

The emergentist Samuel Alexander identified this temperament with a kind of "natural piety".[2] The emergentist sees reductivistic materialism on the one hand, sees substance dualism on the other hand, and notices how much logical space lies in the vast distance between the two positions – and how many real things might fit into that space.

[2] "The natural piety I am going to speak of is that of the scientific investigator, by which he accepts with loyalty the mysteries which he cannot explain in nature and has no right to try to explain. I may describe it as a habit of knowing when to stop in asking questions of nature." Alexander (1922).

## 1.2 Emergence

We have good reasons to believe that the world is, in some sense, physical. All of these things we observe– rocks, rivers, light bulbs, lungs, stars, and so on – is made up of the same kind of *stuff* that the rest of this stuff we observe is: the stuff studied by physics. Perhaps numbers and sets aren't physical, because they are abstract and necessary. But if something is located in space and time and has physical effects, then it seems suspicious to claim that it has no physical connection to the rest of this physical world. How can it be located somewhere if it isn't *physically* located somewhere? How can it have effects on physical things without standing in *physical* relations to those things? How can it interact with the physical if it's independent of it? So, we have *prima facie* reasons to believe that all contingent, concrete particulars (events, property instances, states of affairs, objects, kinds, and other phenomena) are either physical or, at least, *entirely depend* upon the physical.

We also have good reasons to believe there are certain phenomena which are located in space and time and causally efficacious, and yet that these phenomena are *not* wholly physical. If something is a physical kind of thing, then it ought to be possible, at least in principle, to fully describe it and explain it in physical terms. Consider the phenomenally[3] conscious experience you are having right now. At best,

[3] I am using the adjective "phenomenal", the adverb "phenomenally" , and the noun "phenomenology" to indicate the sort of conscious experience which I as a subject am having right now, smelling scent of perfume and listening to the birds outside, as distinct from the ordinary sense of "conscious" in which we mean that

this might be *partly* explained by a statement about neurons firing in a brain. For one thing, there are neurons firing in billions of other brains around the planet elsewhere – but you are not having an experience over there. For another thing, nothing about neurons firing, as opposed to millions of other complex physical processes like cellular division or radioactive decay, makes it obvious that your experiences should appear where neurons are firing and nothing else.

Similarly, no statement about how a material object *is* makes it obvious why one *ought* to believe the truth, make valid inferences, or avoid causing harm. No non-intentional description of the phonology and syntax and social use of a language seems enough to explain why the words in it are *about* something. Phenomenology, normativity, and intentionality are real parts of our world, and yet they seem to require more explanation than physics could conceivably offer. This is why we have *prima facie* reasons to believe that these things are, at least partly*, independent* of the physical world – in the absence of some further explanation.

So, we have good reasons to think that everything entirely depends on the physical world, and yet we also have good reasons to think that some things are partly independent of the physical world. Yet, for anything to entirely depend on

---

someone is aware and responsive to us. A comatose person might be unresponsive, and thus "unconscious", yet having a rich variety of phenomenally conscious experiences; a body might display reflex-responses, yet there is nothing it is like to be it. Wherever I use the word "conscious" or "experience", I am referring to phenomenally conscious experiences. See Block (1995) for discussion of this distinction.

something else and yet be partly independent of it looks like a clear contradiction. How do we resolve this problem?

Physicalist (or materialist[4]) responses to the problem aim to preserve the dependence of everything on the physical by providing some sort of explanation for why the conscious mind (or normativity, or intentionality) isn't really independent after all. For instance, a physicalist might hold that the mind isn't *real*: an eliminativist position. A physicalist may hold that the mind is real but *identical* to some material thing: a reductivist position. More moderately, a physicalist might hold that the mind is independent in the sense that our concepts for it are independent of our physical concepts, so there could be no explanatory reduction from our theory of the mind to any physical theory – but that every token mental state is identical to some token physical state. This position represents a sort of non-reductive physicalism.[5]

---

[4] Following recent conventions, I will often use the term "physicalism" rather than "materialism", insofar as "materialism" suggests the claim that everything is made of matter whereas "physicalism" only suggests that physics is a theory of everything. Nonetheless, "physicalism" should be understood as the historical successor of the traditional materialism of Democritus, Epicurus, Lucretius, etc.

[5] For simplicity, I am grouping reductivist physicalism with type-identity theories on the one hand, and non-reductive physicalism with token-identity theories on the other hand. I recognize that there may be further logical space in between these views, or in different combinations of these views, but the eliminativist / reductivist / non-reductivist schema suffices for my purposes here.

Traditional dualist[6] responses aim to preserve the independence of the mental by providing some sort of explanation for why it merely seems to depend on the material. For example, a dualist might hold that the mind is an immaterial substance, but that material and immaterial substances interact. More moderately, a dualist might hold that the mind's dependence upon the material isn't really *complete*, but only partial and contingent.

Emergentism is an attempt to explain how both claims could be reconciled: the claim that everything depends on the physical *in one respect*, and the claim that minds are independent from the physical *in some other respect*. For the emergentist, in some situations novel phenomena[7] *arise out of* a set of basal conditions while

---

[6] Both Cartesian and non-Cartesian substance dualism will qualify as "traditional dualism". A kind of property dualism might also qualify as "traditional dualism", if one asserts that something further is needed for the existence of phenomenal property instances over and above instances of physical properties. I prefer to distinguish varieties of dualism in terms of the degree of independence they assign the mental from the physical, rather than in terms of what they are a dualism *of*.

[7] Throughout, when I use the term "phenomena" as a noun with respect to the sorts of things which may be emergent, I mean it to include properties, events, states of affairs, entities, kinds, natures, and/or whatever sorts of items might be included in the reader's ontology whose existence can be dependent on other phenomena – in other words, anything but substances as traditionally construed. Someone who holds that there are no events need not hold that there are emergent events; someone who holds that there are *only* events can hold that events are

remaining distinct from these basal conditions. For example, conscious properties might be said to emerge out of neuro-biological properties – they are dependent upon these neuro-biological properties, but they are *novel* properties which are not themselves found in neuroscience or biology.

### 1.3 Cosmic Hermeneutics

What is "novelty" supposed to mean? The early British Emergentist C. D. Broad (1925) articulated a test for metaphysical "novelty" which might be called *Cosmic Hermeneutics*[8]. Suppose that an ideal reasoner has knowledge of all of the underlying properties of this material world – the fundamental physical laws and kinds, the locations of various particles and their respective masses and charges, and so on. Would that reasoner then be able to deduce *a priori* the truths of a higher-level domain given only the knowledge of these fundamental physical facts? Broad held that if the higher-level domain is deducible, then it's not emergent; if the higher-level domain *isn't* deducible, then it is emergent. Since conscious properties can't be deduced *a priori* from physical properties even by an ideal reasoner, according to Broad, conscious properties qualify as ontologically "novel".

---

emergent. In this, I am breaking somewhat with a contemporary emphasis on emergent *properties*, which seems to me to confuse what is distinctive about emergentism with the distinctives of property dualism. The term "phenomena" (used of the variety of objects in the world which might possibly be experienced) is *not* here being defined in terms of what is actually or possibly "phenomenal" (see footnote 3).

[8] This neologism can be credited to Horgan (1983)

In this work, I aim to articulate and defend a form of emergentism which (i) preserves the physicalist claim that in some respect all spacio-temporally located properties and entities completely depend upon the physical world, yet (ii) does not sacrifice the dualist claim that phenomenal properties and entities, though spacio-temporally located and causally efficacious, are yet in some respect independent of the material world. My view shares similarities with that of Almog (2009), and with the view which Horgan (2009) named "Moorean Minimal Emergentism": the view that phenomenal properties supervene with metaphysical necessity upon physical properties, but that the supervenience relationships are fundamental and explanatorily basic. Unlike historical emergentism, the form of emergentism I will defend denies that novel *forces* can emerge above and beyond those at the physical level.[9]

## 1.4 Four Objections

In this work, I will respond to four objections which have arisen against ontological emergence.

First, there are concerns about the coherence of the claim that physical basal conditions are sufficient for emergent phenomena to arise, and yet the emergent phenomena remain distinct from their basal conditions, a "genuine novelty". On one interpretation of the sufficiency claim, the physical base must metaphysically necessitate the emergent phenomenon. However, this seems to conflict with the

[9] McLaughlin (1992) takes it as definitive of the British Emergentism of Mill, Broad and Alexander that emergentists believed in novel "configurational forces", although O'Connor (2012) disputes that Alexander agreed with Mill on this point.

traditional interpretation of "distinct from", on which emergent phenomena are not ideally deducible from their basal conditions. Kim (2010) has argued that the Emergentist's criterion of *ideal a priori non-deducibility* amounts to nothing more than a lack of metaphysical necessitation.[10] So, emergence seems to be self-contradictory.

Second, it is objected that emergence cannot account for the causal powers of mental events: for instance, that my belief that I will be called upon if I raise my arm, and my desire to be called upon, can together produce the action of raising my arm. Kim (2010) has argued that, if mental events are able to cause other mental events, or physical events, then mental events must be identical to underlying physical events – or else, physics is not causally closed. The emergentist, unlike the substance dualist, does not want to deny the causal closure of physics. However, according to Kim, this commits the emergentist to the pervasive, regular, and inexplicable *overdetermination* of some physical events by both mental events and other physical events. Overdetermination of this sort does not seem plausible.

Third, there are concerns about the validity of Cosmic Hermeneutics as a criterion for emergence. The emergentist is making an inference from an *epistemic*

---

[10] Consider that, if something is a necessary truth, then it should be deducible *a priori* given any sound and complete derivation system. Presumably, any ideal reasoner will have a sound and complete derivation system. So, according to Kim (2010), an emergentist can't coherently describe the dependence of the emergent on the physical as involving metaphysical necessity, and instead must accept at most *nomological* supervenience.

*gap* in what an ideal reasoner can infer about consciousness (or normativity, or intentionality) on the basis of the physical facts to a *metaphysical gap* between the nature of the conscious (or normative, or intentional) and the physical. But Kripke (1980) demonstrated that some necessary truths are *a posteriori* (and some contingent truths are *a priori*). So, without further bridge premises, one can't derive conclusions about what's metaphysical necessary from premises about what's epistemically *a priori*. While Chalmers and Jackson (2001) have argued that two-dimensional semantics provides a hurdle over the barrier, Soames (2005) and Block and Stalnaker (1999) have cast doubt on their plan. The relationship between cosmic hermeneutics and emergence remains unclear.

Fourth, all emergentist positions are subject to the concern that non-material properties and kinds have a certain "suspicious smell" to them – they are unlike the sorts of properties we encounter in the natural sciences. Any fundamental "laws of emergence" linking higher-level phenomena to their basal conditions would be "nomological danglers", laws without precedent or parallel in physics, chemistry, or biology. The very "novelty" of emergent properties is reason to be dubious about their existence.

This work will be very limited in terms of providing any positive argument for emergentism. Rather, it will focus on defending emergentism from these four objections. I aim to:

(i) Develop a coherent account of emergence, on which emergent properties are those which supervene with manipulative necessity on their basal conditions, but whose natures are not grounded in the natures of their basal conditions. (Chapter 2)

(ii) Explain how emergent phenomena can have causal powers, distinct from those of their basal conditions, without leading to widespread overdetermination or violations of the causal closure of physics. (Chapter 3)

(iii) Re-connect this account of ontological emergence with the project of cosmic hermeneutics, by defining emergent phenomena as those which aren't deducible from their subvening base by means of analytic *a priori* inferences alone. (Chapter 4)

(iv) Suggest that, far from being an *ad hoc* response to the case of consciousness alone, a weak form of emergence might describe a wide variety of phenomena in the natural and social sciences, boosting the explanatory power of the concept of emergence. (Chapters 5 – 7).

In the remainder of this introductory chapter, I offer a summary of some of the history of Emergentist thought, clarify the distinctions between *epistemic* and *ontological* emergence, reflect on the relationship between Emergentism and "naturalism", and introduce some of the major concepts which will appear in subsequent chapters.

## §2. History

### 2.1 The Idea of Emergence

It is not unusual to think of the world as being made up of varying layers of complexity. Sometimes people say things like "that is a nice dog". Then, someone less tactful will say "but really, it's just a bunch of organs and tissue moving around," or someone who wants to sound especially smart will add, "but really, it's just a bunch of atoms floating around and bumping into each other." Sometimes a loftier individual goes in the other direction: "but really, it's just playing a small

function in the planetary ecosystem", or even "but really, that dog is nothing when you think of the big picture, the stars and the galaxies." One might think of the microphysical as one layer, the biological as another, the whole organism as another, the social or ecological as another, and the galactic as another layer still.

Sometimes, when people see large groups of things, nothing particular strikes them. A collection of grains of rice gathered into a heap is just that: a collection of grains of rice. This is ordinary composition: the whole is simply the sum of its parts. However, some of the time, people begin to worry that something more than ordinary composition may occur when things are grouped together. People sometimes admonish those in the midst of investigating the intricate details of a complicated problem, "don't miss the forest for the trees!" or "the whole may be more than the sum of its parts!"

None of this is serious metaphysics. Still, this is more or less where the idea of emergence gets its start. Out of a system which is very complex on one level, something new appears at a higher level which seems very different from the system on the lower level – it emerges.

## 2.2 Ancient Emergentism

In Plato's *Phaedo*, Simmias proposes an alternative to Socrates's account of the relationship between the soul and the body. Simmias suggests as a metaphor the relationship between a harmony and a tuned lyre:

> One might say that the harmony is invisible and incorporeal, and very beautiful and divine in the well attuned lyre, but the lyre itself and its strings are bodies, and corporeal and composite and earthy and akin to that

> which is mortal. Now if someone shatters the lyre or cuts and breaks the strings, what if he should maintain by the same argument you employed, that the harmony could not have perished and must still exist? …. And I fancy, Socrates, that it must have occurred to your own mind that we believe the soul to be something after this fashion…. the soul is a mixture and a harmony of … elements, when they are well and properly mixed.[11]

On Simmias's harmony theory, while the soul might be distinguished from the body as a harmony is distinguished from a lyre, nonetheless the soul wholly depends for its existence upon there being a body, much as a harmony wholly depends for its existence on there being a lyre. This is in contrast to Socrates's theory of the soul, on which the soul does not depend upon the body and cannot be affected by changes in it. Simmias's theory can be identified as a kind of *supervenience* thesis: there is no possibility of the change in the soul (or the harmony) without a change in the body (or the lyre). Simmias's theory also suggests a kind of *epiphenomenalism*, since arguably it's the lyre and not the harmony which does all of the causal work.

Victor Caston (1997) interprets Aristotle's philosophy of mind in *On the Soul* as an emergentist response to the harmony theory. Aristotle adopts the supervenience thesis, but denies that the soul is epiphenomenal: the mind has the power to cause things.[12] In accepting supervenience, while insisting that downward

---

[11] Plato, *Phaedo*, 85 - 88

[12] Caston (1997), 327

causation occurs, Aristotle's position on the soul is reminiscent of contemporary emergentism.

Caston also interprets Galen as an early emergentist. Galen too endorsed the supervenience thesis of the harmony theory, without accepting its epiphenomenalism. Instead, Galen distinguished between effects which resulted from a mere compounding of causes, and effects which resulted from something more than a compounding of causes – cases in which a "novel characteristic" appeared. [13]

### 2.3 British Emergentism

Galen's notion of the distinction between a mere compounding or aggregation of causes, and a "heterogeneous" mixing of causes, found its way into modern philosophy through the work of John Stuart Mill. In his *System of Logic* (1843), Mill distinguished between *homopathic* effects and laws and *heteropathic* effects and laws. A homopathic effect obeyed Mill's law of the "composition of causes." This law stated roughly that two or more causal forces exerted on the same object would produce a predictable result by means of something like vector addition. A heteropathic effect, on the other hand, was an effect which was not the result of mere addition of two causal forces, and which involved an effect of another type, or at another *level*, than the causes in the base.

Mill's idea was adopted by George Henry Lewes (1874), who coined the term *resultant* to refer to Mill's notion of a homopathic effect, and *emergent* to refer to Mill's notion of a heteropathic effect. From Lewes, the concept of "Emergence" found

[13] ibid.

its way to the philosopher Samuel Alexander (*Space, Time, and Deity*, 1920), and from there to the psychologist C. Lloyd Morgan (*Emergent Evolution*, 1923).

Morgan's emergentism was *historical* and *diachronic*. Morgan believed that at different points in history, a system at one level would necessarily reach a certain degree of complexity and then evolve into something at a higher level. The higher level system would be a genuine historical novelty in the universe, a new kind of thing, not merely a compounding of the things that came before it. For example, life would have been emergent on Morgan's view – a new thing over and above what happened at the chemical level, yet which inevitably evolved out of the chemical level at a certain point of complexity.

C. D. Broad (1925) moved emergentism out of the domain of historical cosmology, and instead used it to analyze the *synchronic* relationships between "higher level" and "lower level" kinds and properties, within the practice of scientific explanation. Broad aimed to strike a middle of the road view between Substance Vitalism and Substance Dualism on the one hand, and "Mechanism" about life or the mind on the other. Unlike the Substance Vitalists, emergentists did not posit the existence of any non-physical vital substance or entelechy behind biology. Unlike Cartesian Dualists, emergentists did not argue for the existence of independent mental substances. Broad held that the only substance was a physical substance, and that all things depended on the physical for their existence. However, he rejected the mechanistic picture on which all of the higher-level properties (life, consciousness, and so on) were fully predictable from a micro-physical description of the world.

Broad held that higher-order phenomena like life and consciousness emerged from lower-level physics as a result of contingent "trans-ordinal" laws, without the intervention of supernatural entelechies or Cartesian substances. However, these "trans-ordinal" laws were not predictable, *even in principle*, on the basis of physics, except by actually observing how things occurred. Contrast this with the mechanistic picture, on which the world could be compared to a kind of cosmic clock – anyone who knows exactly what is happening inside the clock will be able to predict the motions of the hands of the clock, without needing to know how to tell the time. For Broad, while it might be possible *after knowing* the laws of biology or psychology to match them up with the various microphysical events from which they emerge, not even an infinitely powerful mind (a "mathematical archangel") could deduce the whole of biology or psychology simply by looking at microphysics. Biological and psychological properties were *genuine novelties*. Cosmic hermeneutics failed for the case of biology and psychology, and this indicated that these domains were emergent.

Unfortunately, according to McLaughlin (1992), many of the British emergentists took their paradigm example from chemistry. Consider how the properties of table salt (NaCl) have nothing in common with the properties of sodium (Na) or of chlorine (Cl). This sounds like a compelling example of emergence – the properties of NaCl are unpredictable even in principle from the properties of Na or of Cl. Alexander, Broad, and many other emergentists relied on this example from chemistry to build plausibility for emergence at a higher level. However, shortly after they wrote in the 1920's, the discovery of quantum mechanical laws provided a comprehensive and satisfying explanation of chemical properties in terms

of microphysical ones – laws which successfully predicted, among other things, that table salt should have the chemical properties it does. The discovery of DNA shortly after, offered an explanation of supposedly "emergent" vital properties in biology, like reproduction.

Although Broad was not committed to emergentism about chemistry or vitality[14], and acknowledged that chemistry or life might very plausibly turn out to be mechanistic, the association of Broad and other emergentists with this view lead to the dismissal of emergence as an outdated hypothesis.

The gravestone of any idea arrives when its language is co-opted by its opponents. The gravestone for British Emergentism was the redefinition of "emergence" within the Deductive-Nomological account of explanation of Hempel and Oppenheim (1948) and Ernest Nagel (1961). On this framework, *in principle* irreducibility or unpredictability from fundamental physics wasn't a reason to call something a "new" or "novel" property – it was a reason to dump it from our ontology entirely. "Emergence" now indicated a purely epistemic category, rather than an ontological one. Those higher-level phenomenal were emergent which, given our current state of knowledge, could not yet be reduced to fundamental physics by means of bridge laws.

[14] Broad (1923) sees "no *a priori* impossibility in a mechanistic biology or chemistry", unlike the case of "secondary qualities". (72). Chemistry is reducible to physics "at least in part" (54), and emergence holds at most "in certain cases" (55), "so far as we can tell" (63).

## §3. Distinctions

### 3.1 Epistemic and Ontological Emergence

Surprisingly, over the last few decades, "Emergence" has gradually returned as an increasingly popular topic both in the sciences and in the philosophy of mind. That said, it is worth distinguishing two different kinds of "Emergence" here: *epistemic* emergence, which occurs when a higher-level system is not predictable from a lower-level system, at least given the limitations of our present knowledge or perspective, and *ontological emergence*, which occurs when a higher-level system is something over and above a lower-level system.

The epistemic sense of emergence is used in the science whenever a phenomenon does not seem to be readily predictable from what happens on the level of its parts. In ecology, for example, ecosystems may be argued to be emergent in the sense that no amount of data gathering about the parts will permit a prediction about the behavior of the whole. Likewise, in the social sciences, patterns at the social level may produce prediction-enabling social laws, but these laws defy any form of modeling or predicting from the lower level. For example, market cycles, may not be predictable from the spending behavior of individuals; the actions of nations or governments may be predictable based on macro-scale political, social, and economic data, but not from a summary of the properties of the individuals in them. Simulations in the sciences may produce novel and interesting patterns which do not seem to be predictable *except by performing the simulation* over again. Nonetheless, none of these cases of epistemic emergence entails the ontological emergence of these properties.

It is in the philosophy of mind that the stronger notion of ontological emergence tends to be discussed. Here, while it is acknowledged that chemical and biological properties have found reductive explanations in microphysics, there is still great resistance to the idea that conscious or phenomenal properties will find reductive explanations in microphysics. If anything is emergent, consciousness seems to be a good candidate. To many people, it does not seem as though *even in principle* that any microphysical description of reality will in itself enable a deduction of *what it is like* to experience something.[15]

### 3.2 Weak and Strong Emergence

The distinction between epistemic and ontological emergence is often confused with a separate distinction, between weak (or weaker) forms of emergence and strong (or stronger) forms of emergence. "Weak emergence" is taken to mean "merely epistemic", and "strong emergence" is taken to mean "ontological." For instance, Chalmers (2006), offers this definition:

> We can say that a high-level phenomenon is strongly emergent with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are not deducible even in principle from truths in the low-level domain . . . We can say that a high-

[15] It might be deducible that I am having thoughts, or that I have beliefs, if 'thought' and 'belief' are stripped of phenomenal quality and given a functional definition. But it seems difficult to see how my phenomenal experience of things, if it is real, could be deduced.

level phenomenon is weakly emergent with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are unexpected given the principles governing the low-level domain.

Bedau (2008) develops a very different three-part hierarchy of nominal emergence, weak emergence, and strong emergence:

> The simplest and barest notion of an emergent property, which I term mere *nominal emergence,* is simply this notion of a macro property that is the kind of property that cannot be a micro property…. *Strong emergence* adds the requirement that emergent properties are supervenient properties with irreducible causal powers…. Assume that *P* is a nominally emergent property possessed by some locally reducible system *S*. Then *P* is *weakly emergent* if and only if *P* is derivable from all of *S*'s micro facts, but only by simulation…. Properties come in various degrees of derivability without simulation, so there is a spectrum of more or less weak emergence. (159-160) [16]

In this work, I will be defending *ontological* rather than merely epistemic emergence, and my primary interest and source of examples will be the emergence of phenomenal consciousness. However, I will challenge the view that emergence need be *merely* epistemic in the social sciences, natural sciences, or computational

[16] Emphasis added.

sciences. Instead, I will suggest that in these sciences, there may be such a thing as *weak* ontological emergence. For example, while consciousness may be strongly ontologically emergent, the function of the heart in pumping blood may be considered weakly ontologically emergent. On my account, whether emergence is stronger or weaker will depend on what cognitive resources would be needed for an ideal reasoner to deduce the emergent phenomena.

## §4. Naturalism

### 4.1 What is Naturalism?

Emergence is inconsistent with physicalism. At the same time, emergence is supposed to in some sense be a naturalistic project, unlike traditional dualism – emergent phenomena are supposed to "arise from" nature, rather than result from external impositions upon nature. So, is emergence a form of naturalism, or a form of non-naturalism? It depends what "naturalism" is supposed to mean, and whether it is supposed to be regarded as synonymous with "physicalism".

Etymology offers little help. We get "physical" from the Latin *physica*, or "the study of nature", which comes from the Greek *φύσις*, or "nature", which comes from the root verb *φύειν* "to bring forth, produce, make to grow." This, in turn, descends from the proto-Indo-European root **bhū-*, which is the root of many words associated with being and becoming, including "be". The Latin *natura*, from which we get "nature" and "natural", comes from the verb *nasci*, "to be born", which descends from the proto-Indo-European **g'ene*, which is the root of a number of words involving birth, including our modern English "generate", "gene", and "genus".[17] Perhaps one

[17] OED Online. March 2013. Oxford University Press.

can gather a very metaphorical picture from this of the natural as that which is brought forth, produced, grown, or born *of its own accord*, without some external intervention or manipulation. We say a landscape is "natural" if it isn't by crisscrossed by power lines, and that a child's behavior is "natural at that age" if there is no need to explain it as resulting from influences outside the order of nature. This definition of the "natural", as "that which is not produced by outside intervention", is broader than the typical deferential definition of the "physical" as "that subject matter studied by physicists".[18]

We can distinguish methodological naturalism and epistemological naturalism from metaphysical naturalism. I understand *methodological naturalism* to be the claim that *the mode of investigation typical of the natural sciences can provide a theoretical understanding of the world*.[19] Whether emergentism is consistent with this sort of naturalism for a domain will depend upon whether one regards the study of that emergent domain as a *bona fide* natural science, or whether only physics qualifies as the truly natural science. I understand *epistemological naturalism* to be ambiguous between either (i) the claim that, between competing scientific theories, *it is rational to believe only the one which is the best by scientific standards*[20], a claim which is consistent with emergentism, and

---

[18] Stoljar (2009)

[19] That is, *to the extent such an understanding is possible*. (Stoljar, 2009). It is assumed that some but not all changes in the scientific "mode of investigation" are possible within the constraints of the definition.

[20] Colyvan (2009)

(ii) the claim that *there is no a priori knowledge or justification*[21], which is inconsistent with emergentism to the extent emergence is defined in terms of the success or failure of cosmic hermeneutics.

The more pressing question is whether emergentism is consistent with metaphysical naturalism. I believe that there are four senses of the term "natural" in which ontological emergentism *is* a form of metaphysical naturalism, despite being a form of non-physicalist dualism. There are two senses in which emergence is *not* a form of naturalism, largely in virtue of its being non-physicalistic. In this section, I will first consider the use of "naturalism" as a claim about the metaphysical nature (or essence) of a particular domain. I will then consider the use of "naturalism" as a claim about the cosmos as a whole.

### 4.2 As a claim about the Nature of a Domain

If "naturalism" is used merely as a synonym for "physicalism", understood as the claim that the nature of everything in a domain is exhausted by its physical nature, or that everything is fundamentally physical, then emergentism is not naturalism.

On the other hand, "naturalism" may be used to indicate that everything in a domain has a nature, or essence, which is partly physical. Here, what I call "strong emergentism" is not consistent with naturalism, but what I call "weak emergentism" is consistent with naturalism. For example, a strong emergentist about consciousness will hold that consciousness is not essentially physical, even in part – consciousness could possibly have arisen from non-physical ectoplasm rather than

[21] Devitt (2005)

physical brains. On the other hand, weak emergentism about biological functions, which I will defend in Chapter 6, does qualify as a brand of naturalism. Part of the nature of a biological function is that it is physically realized. While biological functions are *multiply* realizable, they are not realizable in non-physical ways: there could be no ectoplasmic hearts or lungs.

Lastly, in a weaker way, "naturalism" may be used to indicate that everything in a domain supervenes upon the natural. In this sense, emergentism is a form of naturalism. For instance, an emergentist about biological functions agrees that bio-functional properties supervene upon the properties of their underlying physical states with metaphysical necessity. An emergentist about phenomenal consciousness holds that conscious states supervene upon physical states with some form of necessity, whether metaphysical, nomological, or manipulative[22] necessity. However, unlike the physicalist, the emergentist holds that this supervenience is not ontologically explainable in terms of the grounding of emergent phenomena in their physical basal conditions.

### 4.3 As a claim about the Nature of the Cosmos

One might understand "naturalism" differently, as a claim about the cosmos as a whole. Here, "metaphysical naturalism" is most commonly understood as the view that all that exists are those properties, relations, and entities which are required for complete material and causal explanations of phenomena – that is, the

[22] A term I will Chapter 2.

sorts of explanations which constitute the methodology of physics.[23] In this sense, non-reductive realism in any domain is in conflict with metaphysical naturalism.[24] Someone who is a naturalist in this sense may, of course, admit truth conditions for the claims in higher domains – he or she can simply be a nominalist rather than a realist. Alternatively, a naturalist may hold that these domains are in fact ontologically reducible to material and causal explanations without loss of explanatory content. However, the emergentist about a domain must be a realist and ontological non-reductivist about it. So, in this sense, the emergentist is not a metaphysical naturalist.

However, in a weaker sense, "naturalism" may be understood as a claim that denies that there are non-natural substances or forces which interact with the natural cosmos – for instance, there are no forces within the cosmos above and beyond the fundamental physical forces. Here, the minimal sort of emergentism which I defend (as opposed to historical British Emergentism) qualifies as naturalistic, insofar as it denies that there are emergent forces closely analogous to physical forces in any domain. Likewise, emergentism rejects the existence of non-natural substances.

---

[23] Logical and mathematical properties may be admitted, insofar as they are necessary for the methodology of physics.

[24] For that matter, in this sense, it's arguable that realism about the "natures" of things as mentioned above, including the claim that everything has a physical nature, is in conflict with naturalism.

Lastly, "naturalism" may be used metaphorically to give an ontological picture on which the cosmos as a whole contains its own principles of change, and no external intervention from outside of the natural realm are needed to produce change in it. (Upon making nature, God successfully completed the work, and needn't keep intervening to fix it.) In this metaphorical sense, emergentism is consistent with naturalism. To some extent, it is this metaphor which the emergentist appeals to in holding that emergent properties "arise from" their basal conditions and do not involve non-natural interventions, forces, or substances.

Thus, the form of emergentism which I am defending is inconsistent with some forms of naturalism, but consistent with others – particularly, it is consistent with the senses of "naturalism" which rule out all of the so-called "spooky" things in haunted houses and horror movies. One might say that emergentist theories involve "cranes" rather than "sky-hooks".[25]

## §5. Overview of Subsequent Chapters

The remainder of this work will be occupied with providing a response to the four objections noted earlier: (i) the objection that emergence is either incoherent or collapses into non-reductive physicalism or traditional dualism, (ii) the objection that emergence entails widespread causal overdetermination, (iii) the objection that it is incoherent to define emergence in terms of the impossibility of the ideal *a priori* deducibility of emergent phenomena from their basal conditions, or cosmic hermeneutics, and (iv) the objection that emergence lacks explanatory power because it applies only to the special case of consciousness.

[25] Dennett (1996)

In Chapter 2, I will attempt to clarify the definition of ontological emergence, and to provide a coherent concept which is distinct from non-reductive physicalism and from traditional dualism. On the one hand, physical phenomena are supposed to be *sufficient* for emergent phenomena to arise; on the other hand, emergent phenomena are supposed to remain *distinct* from physical phenomena. To define the relevant notion of sufficiency, I will introduce the concept of *manipulative supervenience*, on which there is no possible manipulation of a supervenient phenomenon through anything except for its basal conditions. To define the relevant notion of distinctness, I will attempt to connect emergence to the recent discussion in philosophy on grounding and ontological explanation. On my definition, emergent phenomena are those whose natures are not grounded in the natures in their subvening base, and are thus ontologically fundamental.

Kit Fine (1995) and others have introduced well-motivated distinctions between the sort of dependence involved in supervenience, and the sort of dependence involved in grounding. I believe that the emergentist's acceptance of the total dependence of the phenomenal on the physical can be expressed as an *existential* dependence claim, corresponding to manipulative supervenience, whereas the emergentist's acceptance of the partial independence of the phenomenal from the physical can be expressed as a claim about the *ontological* independence of the nature of the phenomenal from the physical, corresponding to an absence of grounding.

In Chapter 3, I will argue that emergentists need not be overly concerned about the causal overdetermination objections presented by Jaegwon Kim, on which mental-to-physical downward causation entails that some physical events have two

sufficient causes – one physical, and one mental. However, I will argue that the notion of sufficiency in "sufficient cause" is ambiguous between *completive* and *deductive* notions, and that overdetermination only applies in the case where an event has two *deductively* sufficient causes. The sort of emergence I am defending denies that there are emergent downward causal *forces* (which would be deductively sufficient causes), while accepting that there is a form of genuine downward causation on which mental events are completively sufficient for their effects.

In Chapter 4, I will provide a history and analysis of the project of cosmic hermeneutics as it relates to emergence. There are a variety of "scrutability"[26] or deducibility relations which might hold for an ideal reasoner between sets of base-level phenomena and higher-level phenomena. I will argue that, with respect to considering whether a phenomenon is emergent or not, one ought to be concerned with *ontological* rather than propositional or sentential scrutability, and with *analytic* rather than *a priori* or *simulative* scrutability. Contrary to most current accounts, I will argue that an emergent property may be ideally *a priori* scrutable given its basal conditions, provided that an ideal reasoner would have to rely on cognitive resources over and above those involved in analysis.

Cosmic Hermeneutics tends to be discussed as a binary, yes/no sort of question: is a higher-level fact deducible *a priori* from the fundamental facts or is it not? Instead, I consider Cosmic Hermeneutics as a question of *which a priori resources* would or wouldn't be needed in order for the ideal reasoner to deduce the higher-level facts. On the one hand, it seems unlikely to me that many higher-level

[26] A term from Chalmers (2012)

domains are deducible by means of the purely analytic reasoning procedures from some subset of the physical facts. On the other hand, these fields could be *a priori* deducible from their basal conditions. It is not entirely implausible that, were an ideal reasoner given all of the necessary truths as *a priori* premises, even consciousness might be deducible *a priori*. However, *a priori* deducibility by means of any *non-analytic* and *normatively-guided* mental simulation procedures does not establish any kind of dependence of the *natures* of high-level domains on their lower-level basal conditions.

In Chapter 5, I suggest that my approach towards emergence in the preceding chapters offers a resolution to the concern that emergent laws are "nomological danglers" with no precedent outside of consciousness. Instead of being a strange and mysterious phenomenon that only occurs at the highest level, I will argue that emergence can be seen as pervasive, regular, and occurring in many forms and at many levels. I believe this can resolve a number of concerns about dualism raised by the work of J. J. C. Smart.

In Chapters 6 and 7, I offer a variety of examples of emergence in the natural and social sciences, hoping to reinforce my argument in Chapter 5. In Chapter 6, I argue that biological functions which supervene upon evolutionary histories of an organism should be regarded as weakly ontologically emergent. In Chapter 7, I argue that a similar form of weak ontological emergence might plausibly be extended to computational and natural patterns, various functional properties, phenomena in the social sciences, abstract and normative truths, and standards of representational correctness and accuracy.

These chapters build upon one another and refer to each other to some extent, but I have tried to write them with adequate summaries of the material in previous chapters to make it possible to approach any given chapter on its own.

I have tried to err on the side of explaining too much rather than explaining too little. On occasion my discussion of emergence has required diversions into topics like the analytic/synthetic distinction, the varieties of grounding relation, the risks of teleological explanations, the purposes of ontology, the varieties of supervenience, debates over two-dimensional semantics, the project of radical interpretation, and many other topics which might not on the surface appear to be about emergence. A reader who is already familiar with these topics may be able to skim these sections without substantial loss.

William James may have been correct that, to some degree, divergent views and intuitions in philosophy are a reflection of temperament. If this is so, then my own temperament inclines me to seek reconciliation between divergent views and to try to satisfy conflicting intuitions, and this desire will guide much of my defense of emergence.

Chapter 2

# MAKING EMERGENCE COHERENT

## §1 Introduction

Traditionally expressed, emergentism is the position that some phenomena "arise from" a collection of lower-level basal conditions, but these phenomena nonetheless remain "genuine novelties". Conscious properties and entities, for example, are sometimes said to "emerge from" some underlying set of physical properties and entities, and yet to amount to something "over and above" those physical properties or entities.[27] An emergentist would say that the world being the way it is physically is *sufficient for* the phenomenal (or mental)[28], but the phenomenal remains *distinct from* the physical.

These notions of sufficiency ("arising from") and distinctness ("genuine novelty") need clarification. One common complaint about emergentism is that,

---

[27] I do not think the emergent properties are limited to the conscious properties (see chapter 5), but insofar as consciousness presents us with the strongest and clearest case for emergence, and the goal of this chapter is to get clear on what emergence is, I will discuss emergentism in this chapter as if it were only relevant to the hard problem of consciousness within the philosophy of mind.

[28] I grant that, insofar as there are mental states which are not phenomenally conscious, these phenomenal states may admit of a sort of reductive explanation which consciousness does not admit of. For brevity, I will use "mental *x*" in this chapter as though it were synonymous with "phenomenally conscious phenomenal *x*", even though I do not believe the two are synonymous.

whenever one tries to clarify these two claims, then it seems that either (i) emergentism collapses into traditional dualism, or (ii) emergentism collapses into non-reductive physicalism, or else (iii) emergentism turns out to be incoherent and self-contradictory. After all, non-reductive physicalists and traditional dualists will both accept *some* sense in which the physical is sufficient for the phenomenal, and yet the phenomenal remains distinct from the mental.

For example, emergentists since Broad (1925) have understood the distinctness claim to mean that the facts about emergent phenomena are *not deducible* from the facts about their basal conditions. Non-deducibility would not be ontologically interesting if it were merely a reflection of our contingent cognitive limitations. For this reason, it's often held that ontologically emergent phenomena must not be deducible *a priori* even by an ideal reasoner or "mathematical archangel".

However, this formulation of the distinctness claim invites conflict with a certain interpretation of the sufficiency claim. Sufficiency may be interpreted as involving the *supervenience* of emergent properties on their basal properties. Supervenience comes with different brands of necessity. Suppose that the existence of the subvening properties *metaphysically* necessitates the existence of the supervening properties. This raises a problem, as Kim (2010) notes: why wouldn't the facts about the supervening properties be deducible by an ideal reasoner? Either the ideal reasoner has access to all of the necessary truths – meaning that the failure of ideal deducibility amounts to a failure of necessitation, and thus a failure

of supervenience[29] – or else the ideal reasoner lacks access to some *a posteriori* necessity, in which case emergentism collapses into *a posteriori* physicalism.[30]

Van Cleve (1990) suggests that the emergentist should be understood as accepting supervenience with *nomological* necessity rather than metaphysical necessity. On this view, it is metaphysically possible for the same basal conditions to occur without the emergent phenomena (hence, ideal deducibility fails), but there is some law of nature which is sufficient for the emergent phenomena. These laws of nature must *not* be part of the ideal reasoner's deduction base, since the emergent phenomena would then be deducible from the conjunction of the facts about the basal conditions and the laws of nature.[31] So, the emergentist must hold that the relevant "trans-ordinal" laws linking basal states to emergent states remain contingent and non-deducible even given the other laws of nature. These trans-ordinal laws are not *physical* laws of nature, but laws "added" to physical nature which make emergence happen. Besides raising Smart's (1959) worries about "nomological danglers", this seems to undercut much of the appeal of emergentism, and risks collapsing emergentism into traditional dualism. It suggests that the physical world being exactly as it is and having the laws which it does is no longer enough for emergence.[32] Something further and non-physical must be present in our

[29] The argument appears in Chapter 4 of Kim (2010): " 'Supervenient and Yet Not Deducible': Is there a Coherent concept of Ontological Emergence?", 85-104

[30] Byrne (1993)

[31] Kim (2010), Ch. 4, makes this complaint.

[32] Braddon-Mitchell (2007).

world to explain why emergence happens in our world and not in a physical duplicate: one must add contingent trans-ordinal laws, or mental substances, or so on.

To put the problem briefly: if emergent properties supervene with metaphysical necessity on the physical, then emergentism seems like repackaged non-reductive physicalism; if emergent properties only supervene on the physical given the addition of contingent trans-ordinal laws, then emergentism seems like repackaged traditional dualism.

In this chapter, I *offer* two strategies for the emergentist to express both the distinctness and sufficiency claims coherently, in a way which distinguishes emergentism both from traditional dualism and non-reductive physicalism. I will begin in Part A by exploring the notion of *distinctness*, the sense if an emergent phenomenon's being a "genuine novelty". I will attempt to relate this to recent discussions on grounding and fundamentality in metaphysics. In Part B, I will explore the notion of *sufficiency* and develop an alternative form of supervenience which better matches what the emergentist is saying. On my account, there are *two* distinct kinds of explanation at play in the debate: *ontological* explanation, or "grounding", and *existential* explanation, or "supervenience."

## PART A

## §2. Emergence and Ontological Explanation

### 2.1 Sufficiency and Distinction

Traditional dualists, ontological emergentists, and non-reductive physicalists all acknowledge that there is an "epistemic gap" between our physical concepts and our phenomenal concepts. The question which divides these positions is whether and

in what way this epistemic gap translates into an *ontological* gap.[33] All sides acknowledge that phenomenal properties are distinct from physical properties: the question is what precisely this distinctness consists in.[34] We'll assume a physicalist cannot accept *strong modal distinctness*: that is, that it is possible that the same physical properties should occur without the same phenomenal properties.[35] The traditional dualist *does* accept strong modal distinctness. If an emergentist denies strong modal distinctness, then something else must distinguish emergentism from physicalism.

What will do the trick? A physicalist can accept *weak* modal distinctness (or "multiple realizability"): the same phenomenal properties might occur, yet be realized by different physical properties. A physicalist can also accept that physical properties are *numerically* distinct from phenomenal properties; provided, that the right sort of further explanation must hold between the phenomenal properties and the physical properties.[36]

---

[33] Levine (1993)

[34] The following discussion of types of distinctness is adapted from Stoljar (forthcoming).

[35] ibid.

[36] For an argument for why supervenience (that is, a denial of strong modal distinctness) isn't enough for physicalism without some further explanation of the supervenience, see Horgan (1993)

What sort of further explanation is a physicalist committed to? Let's call the relevant sort of explanation an ontological explanation.[37] We can say that the physicalist holds that there is an ontological explanation for phenomenal properties in physical terms, that phenomenal properties ontologically depend upon physical properties, and that phenomenal and physical properties are not ontologically distinct.

If the emergentist can deny that this sort of ontological explanation holds between physical and phenomenal properties, then the emergentist can reject physicalism while maintaining an equally strong sense in which physical facts are sufficient for mental ones. The prescription for the emergentist who rejects strong modal distinctness is to hold that phenomenal and physical properties are nonetheless ontologically distinct insofar as the phenomenal does not ontologically depend upon the physical or admit of an ontological explanation in physical terms.

Of course, we have yet to define what an "ontological explanation" is supposed to be. One might be skeptical about the concept at this point and hold that ontological explanation *just is* modal explanation. In this section, I'll argue that (i)

---

[37] The terms "ontological explanation" and "metaphysical explanation" are used inconsistently within current literature in a variety of contrastive ways, typically with one supposed to refer to some sort of modal explanation or supervenience, and the other supposed to refer to a "metametaphysical" grounding or fundamentality relation or to the traditional concept of essence or nature. I will be using "ontological" for the latter and "metaphysical", "existential" or "modal" for the former. For the reverse usage, see Barnes (forthcoming).

there are a family of concepts of "grounding" in the present literature which diverge from modal explanation and which support a distinctive kind of ontological explanation, and that (ii) non-reductive physicalists need this notion in order to formulate the minimal commitments of physicalism, leaving the door open for emergentists to take advantage of it.

## 2.2 Grounding as Ontological Explanation

Ontological explanation is a non-causal form of explanation. A causal explanation tells us how a thing came to be as it is; an ontological explanation tells us why a thing is the sort of thing that it is. For instance, suppose a burglar steals my television. A causal explanation of why the burglar stole my television will involve the burglar's desire for dishonest gain, my door being unlocked, his patiently observing my being away from home, and the various physical movements involved in the burglar's transporting the television into his minivan. An ontological explanation of the burglar's stealing my television will begin with an explanation of what precisely *stealing* consists in: that is, the nature of private property in Western society, what sorts of transfers are legitimate and which aren't, the various facts which explain why the television qualifies as mine, and so on – with each of these in turn requiring some ontological explanation in their own right. These ontological explanations appeal to some basic notion of "ontological dependence"[38], which has come to be known as "grounding."

[38] Note that various sources refer to the intended relation instead as "metaphysical dependence", and reserve "ontological dependence" for a kind of modal dependence.

Fabrice Correia (2008) distinguishes three forms of "ontological dependence." One of these is the dependence of the *existence* of one thing upon another[39], and a second is the dependence of the *essence* of one thing upon another[40]. The third, which is relevant to my purposes here, is a kind of non-causal explanatory dependence, on which one thing explains why another is the case. Putative examples of this include: (i) Sam is ill or 2=5, because Sam is ill; (ii) Something is human because Sam is human, and also because Kevin is human, etc.; (iii) This vase is colored because it is red; (iv) This ham sandwich exists because the slice of ham is between the two pieces of bread; and (v) The event that was Sam's walking yesterday exists because Sam was walking yesterday.[41] For simplicity, this is the relation I mean to indicate by "ontological dependence" or "grounding."

Notably, the recent literature on grounding represents a revival of a sort of neo-Aristotelianism, and a move away from the debates between Carnap and Quine which occupied much of the 20th century. The central question in metaphysics shifts from *what* things exist to *how* things exist: that is, what grounds what.[42] Jonathan Shaffer writes:

> The Quinean and Aristotelian tasks involve structurally distinct conceptions of the target of metaphysical inquiry. For the Quinean, the target is flat. The

---

[39] I identify this with supervenience – see section 3 of this chapter.

[40] I identify this as a variety of grounding – see 2.6.1 of this chapter.

[41] All examples taken from Correia (2008), 1022

[42] Corkum (2008)

task is to solve for E = the set (or class, or plurality) of entities. There is no structure to E. For any alleged entity, the flat conception offers two classificatory options: either the entity is in E, or not.

For the neo-Aristotelian, the target is ordered. The task is to solve for the pair <F, G> of fundamental entities and grounding relations, which generate the hierarchy of being. For any alleged entity, the ordered conception offers not two but four major classificatory options: either the entity is in F, in G, in neither but generated from F through G, or else in the rubbish bin of the non-existent. (If the entity is in the third class, then there will be further sub-options as to how the entity is grounded.)[43]

For example, metaphysical questions in the Quinean mode will be whether numbers exist, whether hurricanes exist, whether wars exist, and so on. Metaphysical questions in the neo-Aristotelian mode will instead take for granted that such things exist, but instead ask whether these things are grounded in some other facts about the world, or whether they are not grounded in anything further: that is, whether they are fundamental. Hurricanes and wars are clear cases of entities grounded in other entities.[44] Numbers are a case under debate: a Platonist will hold that numbers are fundamental, whereas on alternative views numbers will qualify as existing but as grounded in something else. Grounding relationships are

[43] "On What Grounds What", 354

[44] We might instead speak of the property of being a hurricane as being grounded in other properties, or of the facts about hurricanes as being grounded in other facts.

supposed to help explain *why* certain necessary truths are so. It is a necessary truth that 2+2=4. If set theory is what grounds ordinary arithmetic, then 2+2=4 is explained by the facts about set theory. If something else grounds ordinary arithmetic, then 2+2=4 is not really explained by set theory, even though it is necessitated by it.[45] Many of the far more ordinary facts about the world are believed to hold *in virtue of* some more basic set of facts.[46] For example, the fact that Smith is a polygamist holds in virtue of the fact that Smith has more than one spouse, and the fact that Smith has a spouse holds in virtue of some complex set of social facts

[45] Psychologism thus is the view that the facts about a domain are grounded in certain psychological facts. In contrast, one might hold that psychological facts necessitate that 2+2=4 without being the grounds of 2+2=4.

[46] It is helpful to work in the mindset of a correspondence theory of truth. Sentences (basic units of language) express propositions. Propositions consist of various concepts structured together. Propositions represent the world and can be either true or false of it. Facts are those representable parts of the world in virtue of which propositions are true. (We might as easily speak of "states of affairs".) The proposition <salt dissolves in water> is true iff the fact [salt dissolves in water] exists. Facts have a corresponding internal structure: facts consist of combinations of objects and properties: [salt dissolves in water] consists of two elements: the object 'salt', and the property 'dissolving-in-water'. Facts ground propositions, and facts are grounded in their elements or other facts.

about marriage and the public actions of Smith and his relevant spouse. This "in virtue of" relation can be identified with the notion of grounding.[47]

Suppose we accept a version of Occam's razor on which we wish to minimize our ontological commitments. For the Quinean, this means minimizing the set of objects and properties which we assert positively exist. For the neo-Aristotelian, on the other hand, this only means minimizing the set of *fundamental* objects and properties. So long as a property is grounded in a property which is more fundamental, it doesn't come with any ontological cost.

Other relations similar to grounding have been the source of recent discussion. One central motivation is concern about reference and indeterminacy: why it is that "pigs" refers to pigs and not to parts of pigs, pigs + random molecules, or something even more bizarre.[48] The more fundamental properties are supposed to attract the reference of our terms. For Ted Sider, the question is whether or not something is "structural" (that is, a relation which "carves nature at its joints").[49]

---

[47] So, *x ontologically explains y = y is true in virtue of x = x grounds y = y ontologically depends on x*

[48] For the origin of this discussion, see Putnam (1981) and Lewis (1984). Consider how "pigs have legs" could be made equally true of the world if every instance of "pigs" referred to eggs, and every instance of "legs" referred to yolks, with certain other changes to our vocabulary. In some way this would be an unusually "gerrymandered" interpretation of our language, so the more "natural" properties instead shape the reference of terms.

[49] Sider (2011)

For David Lewis, there is an important distinction between properties which are "perfectly natural" and those which aren't.[50] In the truthmaker literature associated with David Armstrong and others, the question is about what the minimal truthmakers are which make true various propositions.[51] There are important differences between the theories and theorists associated with these terms, but for the purposes of this chapter it will suffice to bundle them together under the heading of "grounding" and pass over these disputes.

Note that grounding is supposed to be transitive. The fact that Smith has a multiplicity of spouses is grounded in some further facts about particular marriage-events, distinct persons, and social practices in virtue of which there are such things as spouses. The fact that Smith is a polygamist is thus grounded in these further facts. The facts about the social practices which constitute marriage, in turn, are grounded in some more basic set of facts, and so on. It is plausible that at some point the grounding relation hits "bedrock", certain facts which are more basic than any of the others, which are not further grounded in anything, and which taken together ground all of the other facts. We'll call the thesis that this is so "fundamentalism", and will use the term "fundamental" to describe any fact which is not grounded by any further fact.

The fundamental vs. non-fundamental distinction is helpful for sifting out these issues from the realist vs. nominalist distinction. All parties in a debate might accept that there are truths about ordinary-level facts (facts about tables and chairs

[50] Lewis (1983)

[51] Armstrong, *Truth and Truthmakers*

and so on), and all might accept that ordinary-level facts are non-fundamental (they are grounded in facts about material composition, social practices, etc.). Nonetheless, one party might be a realist about non-fundamental facts, and the other party a nominalist about them. The nominalist may hold that only the fundamental facts are "real" or truly "exist", in the proper sense of "real" or "exists", and will provide anti-realist truth conditions for our speech about non-fundamentals. The realist will maintain that chairs are real without holding that there is anything fundamental about chairs. The concept of fundamentality or grounding allows us to safely bracket this debate and focus instead on determining what the fundamental facts must be: that is, what facts we need at a minimum to ground all of the others.

## 2.3 Grounding, Reduction, and Supervenience

**2.3.1 Reduction.** So, what exactly is grounding? It might be a primitive notion (I think it is). But let's entertain for a moment two attempts to analyze grounding in other terms which have had serious advocates over the years.

One relation it might seem natural to identify with grounding is reduction. Both grounding and reduction are "conservative" rather than eliminativist: that is, we can maintain realism about the higher-level in virtue of its being grounded in or reducible to some real lower-level domain. Both grounding and reduction are supposed to involve no added ontological costs for the grounded or reduced properties or entities, since they are "nothing over and above" what they are grounded in or reduced to. One might hold that there is a reduction of grounding to reduction.[52] Call this the reduction-grounding hypothesis:

[52] Skiles (2013)

(RGH) *x* grounds *y* iff *y* is reducible to *x*.

However, a non-reductive physicalist might wish to reject (RGH), since it suggests that if mental properties are not reducible to physical properties, then these mental properties are not grounded in the physical and come at added ontological cost.

Of course, "reduction" is ambiguous. We might distinguish a kind of *explanatory/theoretical* reduction from an *ontological* reduction: the first involves a translation of some of *our* sentences about the reduced domain (at least, those involving theoretical or causal roles) into the vocabulary of the reduction base, whereas the second involves a translation within the "language of ontology" of the nature of one thing into the nature of another. The non-reductivist might maintain that there is no possible explanatory or theoretical reduction of one domain to another – that the two represent independent domains which cannot be analyzed in terms of each other – and yet at the same time hold that the nature of the one domain contains nothing not already present in the nature of the other. (One might also hold the reverse)[53]. Clearly, for the non-reductive physicalist (RGH) must be

[53] That is, one might hold that all of the relevant sentences about the *causal* roles played by the reduced vocabulary can be translated into the vocabulary of the reduction base, but nonetheless hold that the one is ontologically irreducible into the other. See Chapter 6 for an account on which this would be true of biological functions.

interpreted in terms of ontological reduction rather than mere explanatory or theoretical reduction.

But what is an ontological reduction, if it doesn't require explanatory or theoretical reduction? It seems as though the notion of ontological reduction is every bit as primitive as the notion of grounding: *x* is ontologically reducible to *y* just means that *x* is grounded in *y*. This makes (RGH) uninformative.

**2.3.2 Entailment or Supervenience.** One alternative relation it might seem natural to identify with grounding is entailment. For instance, Armstrong holds that truthmakers must necessitate the truths they make true. Call this the entailment-grounding hypothesis:

(EGH) *x* grounds *y* iff *x* entails *y*

Along similar lines, it may seem natural to identify grounding with supervenience, provided that supervenience is formulated in a way which involves metaphysical necessitation and entails (EH). Call this the supervenience-grounding hypothesis:

(SGH) *x* grounds *y* iff *y* supervenes on *x* with metaphysical necessity.

The advantage of identifying grounding with supervenience or entailment, rather than reduction, particularly for non-reductive physicalists, should be obvious. One gets to preserve realism about the higher-order domain, without the requirement of explanatory/theoretic reduction from one domain to the other. Frank Jackson's *From Metaphysics to Ethics* (1998) provides an example of the application

of (SGH). Although Jackson does not use the term "grounding", this is clearly the role which he means supervenience to play. The facts about ethics, for example, are supposed to supervene upon the physical facts. This means that the facts about ethics are real, but come at no added ontological cost. The same, naturally, is held by supervenience physicalists about the mental: the mental supervenes upon the physical, and thus is no addition to being.

However, grounding and supervenience are different forms of explanation, meant to answer to different sorts of questions. Grounding answers the question, "why is *x* the sort of thing which it is?" or "in virtue of what is *x* the sort of thing that it is?" Supervenience is meant to answer the question, "why does *x* exist?" or "what can change whether or how *x* exists?". The "because" of grounding is different from the "because" of supervenience.

This difference is reflected in three major problems with understanding grounding in terms of supervenience: the problem of asymmetry, the problem of non-grounding necessities, and the problem of mere covariance.

**2.3.3 Problem 1: Asymmetry.** Grounding is asymmetric. The fact that a ball has the power to roll holds in virtue of the fact that the ball is a sphere, but the fact that the ball is a sphere does not hold in virtue of its having the power to roll. In this respect, grounding is unlike necessitation: the fact that Smith is a polygamist *necessitates* that he has more than one spouse, and vice-versa, but it is having a multiplicity of spouses which grounds being a polygamist and not vice-versa.

The same problem applies to supervenience. Consider the example of the property of being a shape with three angles and the property of being a shape with three sides. It is plausible that these are distinct properties; certainly they are

distinct concepts.[54] Being a shape with three angles supervenes upon being a shape with three sides, and being a shape with three sides supervenes upon being a shape with three angles. Nonetheless, it's not clear that grounding explanations go one way or the other in this case – or that both properties aren't grounded in some simpler property, that of being a triangle. So, supervenience can't be grounding.

To take another example, consider the following properties of gasses: pressure supervenes upon temperature and volume, temperature supervenes upon pressure and volume, and volume supervenes upon temperature and pressure. Even though the distribution of any one of these properties will supervene upon the distribution of the other two, it is not clear which property could be said to be grounded in the other two. If anything, all three are grounded in molecular kinetic energy.

Alternatively, consider that one might hold that the property of being a true proposition supervenes upon the property of being known by a necessarily omniscient God and vice-versa, yet hold that God knows the propositions *because* they are true, not that they are true *in virtue of* their being known by God.[55]

Clearly, (SGH) is too permissive. One could try to block this objection by defining grounding in terms of asymmetric supervenience:

---

[54] See Carnap's *Meaning and Necessity*

[55] That is, if one permits that God's making *p* the case can be distinguished from God's knowing that *p* is the case.

(SGH*) *x* grounds *y* iff *y* supervenes on *x* with metaphysical necessity, but *x* does not supervene upon *y* with metaphysical necessity.

However, (SGH*) is now too restrictive. Suppose advanced philosophers in some ideal future discover the true analysis of the nature of knowledge – something like JTB+. Clearly, the property of knowing *p* to be the case should be grounded in the various JTB+ properties. It is not as though knowing *p* is something over and above JTB+. However, the knowledge properties will supervene on the conjunction of the JTB+ properties *and vice-versa*. So, under (SGH*), knowledge would *not* be grounded in the JTB+ properties.

**2.3.4 Problem 2: Non-Grounding Necessities.** A fact can necessitate another fact without grounding it. Two examples have been given by Kit Fine (2001, 2009). First, that the singleton set of Socrates exists necessitates that Socrates exists and vice-versa, but the existence of the singleton set of Socrates is grounded in the existence of Socrates and not vice-versa. Socrates does not exist *in virtue of* his singleton set's existence. Similarly, it is a necessary truth that Socrates exists if he is not numerically identical to the Eiffel tower, but the existence of Socrates is not grounded in his non-identity to the Eiffel tower. There seem to be many such cases of "accidental necessity" along these lines, properties which do not form part of the *ground* of a thing even though they are necessary and/or sufficient conditions for it.[56] Fabrice Correia offers this example:

[56] We might also say it is a case of a necessary property which is not an *essential* property. The traditional notion of essence, should one choose to accept it, lines up

> Define a cat-singleton as a singleton whose member is a cat. Then the property of being a cat supervenes … on the property of being a member of a cat-singleton, while it is very implausible to say that facts of whether or not something is a cat are ontologically derivative upon facts of whether or not something belongs to a cat-singleton.[57]

Because of this, an emergentist could plausibly hold the position that emergence is a case of an *a posteriori* necessity, while noting that not all cases of *a posteriori* necessities are cases in which the grounding relation holds. E. J. Lowe (1998) offers the example of an individual and the individual's life. Suppose we grant that, given that Socrates exists in a world, Socrates's life also necessarily exists in that world. "Socrates's life" is plausibly a rigid designator: it designates the same life in all possible worlds regardless of qualitative differences between the worlds. In a world in which it was Euthyphro who was put on trial for corrupting the youth and died by drinking hemlock, and Socrates was a popular sophist who subverted the course of justice, "Socrates's life" designates the life of the sophist and not the life of the man put on trial. However, this seems compatible with holding that Socrates

---

much more clearly with ontological dependence than it does with mere modal dependence. Of course, one need not accept essentialism to accept grounding. Grounding is a broader notion than the notion of essence: the *x*'s might be grounded in the *y*'s without *y* being essential to *x*, though the reverse cannot hold.

[57] Correia (2008), 1029

does not live the life he lives *in virtue of* being Socrates: it is not as though it is essential to Socrates that he live his life. The properties of Socrates's life may supervene upon the properties of Socrates without following in some way from the nature of Socrates. The *posteriori* physicalist needs to make an argument for why the grounding relation holds between the physical and the mental, above and beyond asserting that it is a case of some *posteriori* necessity resulting from a rigid designator.

**2.3.5 Problem 3: Mere Covariation.** The history of supervenience is often forgotten. Consider that our notion of supervenience is descended from the moral theories of R. M. Hare and G. E. Moore, who held that the moral properties supervened upon the descriptive properties as a way of promoting *non-naturalism* about moral properties.[58] Because supervenience only purports to involve the co-variation of two properties – no change in one without a change in the other – it did not imply any ontological dependence of the one upon the other. (By analogy, consider how two variables might be correlated over time without the one being causally dependent on the other.) It was in the work of Donald Davidson[59] that supervenience found its conversion from a non-naturalist (or dualist) to a physicalist (or naturalist) thesis. However, as Horgan[60] argues, supervenience never broke free

[58] Hare (1952) 80-81; Moore (1922), 261. Moore does not use the term 'supervenience', but clearly endorses a supervenience thesis.

[59] Davidson (1970)

[60] Horgan (1993), 563: "Supervenience is ontological if it is an objective relation between lower-level properties and facts and genuine, objective, higher-level

of its history. It has remained a relation of mere covariance between properties: a kind of "synchronic correlation" without further explanation.

An emergentist dualist can just as well accept supervenience as a physicalist, if the emergentist is willing to hold that the supervenience is mere covariance without further explanation. Thus, supervenience alone is unable to serve as a definition of physicalism. What is needed is "superdupervenience": supervenience with some sort of physicalistically-acceptable explanation for the supervenience, an explanation of why the is properties of the subvening base are *definitive of* the supervening property.[61]

In other words, what is needed is an *ontological explanation* for the supervenience. Supervenience must be explained in terms of the grounding of one property in another: one needs supervenience *plus* grounding for superdupervenience. So, supervenience (or entailment) can't itself serve as the grounding relation, and (EGH) and (SGH) are false.

**2.3.6 Problem 3: Wilson's Rebuttal.** Jessica Wilson (1999) argues that an explanation in terms of what is definitive of the supervening property in the subvening base is neither necessary nor sufficient for the sort of superdupervenience the physicalist needs. Explanations of the sort Horgan requires for superdupervenience, she acknowledges, are not likely to be forthcoming for either phenomenal states or for the domains of normativity and intentionality. Instead, for

---

properties and facts ... supervenience for a given mode of discourse is robustly explainable if it is explainable as ontological."

[61] ibid., 579

Wilson, what the physicalist needs for superdupervenience is for every *causal power* of a mental property to be numerically identical with some causal power of a physicalistically acceptable base property.[62] Call this the causal-power-grounding hypothesis:

> (CPGH) *x* grounds *y* iff *y* supervenes on *x* with metaphysical necessity *and* each of the causal powers associated with *y* is numerically identical with a causal power *x*.

Motivating (CPGH) is a form of Alexander's dictum, on which properties are to be individuated by their causal powers. Wilson reflects on the causal exclusion arguments of Jaegwon Kim, on which widespread overdetermination appears to be a problem for any view on which mental properties have causal powers which are numerically distinct from the causal powers of their underlying physical base. On Wilson's view, the emergentist's reply to Kim's argument is to deny the causal closure of the physical, whereas the appropriate physicalist response is to hold to the numerical identity of causal powers.[63] Thus, (CPGH) captures a way in which a physicalist can hold that mental properties are grounded in (and thus not distinct from) the physical which emergentists can't accept.

---

[62] Wilson (1999), 41

[63] ibid., 42: "each individual causal power associated with a supervenient property is numerically identical with a causal power associated with its base property"

However, the sort of emergence Wilson has in mind – the sort which is committed to the denial of the causal closure of the physical – is much stronger than the sort of emergence which I have been discussing in this chapter. As I argue in Chapter 3, both emergentists and non-reductive physicalists can accept a robust form of downward causation while denying *neither* causal closure *nor* the numerical distinctness of higher-level causal powers from physical causal powers, provided that (i) there are no novel or emergent causal forces at the higher level, and (ii) such "downward causation" does not have a structure on which effects are to be deduced as conclusions from a set of causes and laws as premises. Because emergentism can accept causal closure, causal closure cannot be the basis for distinguishing emergentism from physicalism or for establishing the physicalistic acceptability of a theory.[64]

Of course, emergentists and physicalists who accept Alexander's dictum might still disagree about whether the causal powers possessed by mental properties are numerically identical with causal powers of various physical base properties, or whether there is merely a persistent covariation between the causal powers of each. But how would such a question be resolved? It would be resolved by whether or not there is an ontological explanation of the mental properties in terms of the physical base properties. Suppose no such explanation existed: we simply have two sets of numerically distinct properties, *P* and *M*, with *P* having one set of causal powers (*PC*) and *M* having a set of causal powers (*MC*). In the absence of an explanation of why *M* is nothing over and above *P*, given the numerical distinctness of *P* and *M*

[64] See also Horgan (2002) for his similar reply to Wilson.

(which the non-reductive physicalist accepts), what motivation would there be for holding that (*MC*) is a subset of (*PC*)? The non-reductive physicalist is left rather like someone insisting that there are two numerically distinct synchronized swimmers, but that their kicks and splashes are *not* numerically distinct. If the powers aren't distinct, why think the properties are – and if the properties are distinct, why think the powers aren't?

So, (CPGH) can't serve as a definition of grounding for the physicalist, because resolving questions about the identity of causal powers depends upon resolving questions of grounding or ontological explanation for the properties which have those causal powers.

**2.3.7 Other Differences.** There are other distinctions between the grounding relation and supervenience. There is such a thing as partial grounding, but not such a thing as partial supervenience.[65] Supervenience is reflexive, but grounding is not.[66] Supervenience is generally understood as a relation between properties in general, whereas grounding is generally held to be a relation between particular entities or

[65] See the discussion in 2.6 below

[66] Every set of properties supervenes upon itself, but nothing grounds itself (except perhaps the fundamental). I see this as presenting no problem, however, insofar as it would be easy enough to modify (SGH) to require that a thing not supervene upon itself, or to modify the notion of grounding so that everything is trivially grounded in itself.

local facts.[67] These are worth noting, but do not seem to prevent us from identifying the two in the way that the problems listed above do.

It is also worth noting that the preceding arguments show that supervenience is not sufficient for grounding, but none of them show that supervenience is not necessary for grounding. One might still hold a weaker, one-directional version of the supervenience-grounding hypothesis:

(WSGH): *x* grounds *y* only if *y* supervenes on *x* with metaphysical necessity.

However, since (WSGH) is compatible with an emergentist's accepting supervenience with metaphysical necessity while denying the grounding of emergent properties in physical properties, the non-reductive physicalist needs something further which can play the role of grounding or ontological dependence.

[67] For instance, the chemical properties supervene upon the physical properties, but it's the fact that there is water in the glass which is grounded in the $H_2O$'s physical structure. However, this should not be regarded as a problem in the way the problems above are regarded as a problem. We can speak of grounding between sets of properties: a moral naturalist holds that the moral facts are grounded in the natural facts. We can also speak of local supervenience as opposed to global supervenience: the properties instantiated by the water in the glass locally supervene upon the properties instantiated by the $H_2O$ molecules.

## 2.4 Why Non-Reductive Physicalists Need Grounding

Let's step back a bit. We are assuming there are truths about our phenomenal states. Given that it's not *prima facie* obvious that phenomenal states are physical states, the physicalist needs to give an explanation for how the truths about the phenomenal are *really* truths about the physical. One suitable explanation is type-identity: every phenomenal property (or type) is identical to some physical property (or type). This can be understood as a kind of explanatory reduction. However, a dominant strain of contemporary physicalism does not claim that this sort of reduction is possible. It may or may not accept the identity of phenomenal event tokens with physical event tokens, but it does not accept the numerical identity of properties. So, the non-reductive physicalist has to give some further explanation for why phenomenal properties can exist and be numerically distinct from physical properties and yet not pose a worry for the physicalist's doctrine that everything which exists is physical.

The non-reductive physicalist could simply assert that phenomenal properties come at "no additional ontological cost" above and beyond the physical properties, even though they are not identical to physical properties. This assertion might include an interesting account of why we should not expect a reductive explanation of the phenomenal in physical terms – perhaps that we fail to directly grasp the true nature of phenomenal properties (namely, their physical nature) through our phenomenal concepts, because phenomenal concepts are so *weird* and first-person-y. However, this doesn't lift the original explanatory burden: how is it that physical properties are *not* distinct from phenomenal properties for the purposes of ontology, given that they are numerically distinct?

At this point, we noted that the non-reductive physicalist can turn to the notion of a special kind of "ontological explanation", or *grounding*. Either grounding admits of further analysis, or it doesn't. Grounding cannot be analyzed in terms of either supervenience or asymmetric supervenience. Grounding might be analyzed in terms of an explanatory or theoretical reduction, but if we do so then non-reductive physicalism is false – so, that won't do. Perhaps grounding can analyzed in terms of a primitive notion of *ontological* reduction, but this isn't especially informative. So, in lieu of some further analysis of grounding, the best strategy at this point for the non-reductive physicalist is to accept a kind of primitivism about the concepts of grounding and fundamentality.

I must emphasize that it is the non-reductive[68] physicalist who needs a primitive notion of grounding, not the emergentist. A physicalist who holds that mental properties are not identical to physical properties and yet are not truly distinct from them needs an ontological explanation of mental properties in terms of physical properties. Supervenience does not provide that explanation, and in the absence of other candidates the physicalist must hold that grounding cannot be further analyzed. It is not as though the emergentist invokes this novel form of explanation – "grounding" – in order to deny it. The emergentist simply denies both:

(D1) mental properties are identical to physical properties; and

[68] The reductive physicalist, recall, could define grounding in terms of epistemic or theoretical reduction.

(D2) there is some ontological explanation of mental properties in terms of physical properties, other than identity, such that mental properties are no addition to ontology.

The reductive physicalist can accept (D1); it is the non-reductive physicalist who bears the burden of denying (D1) while accepting (D2). A skeptic about grounding, who holds that (D2) has no further content beyond (D1), should hold that if (D1) is false then (D2) is false. Thus, a skeptic about grounding must hold that dualism is the immediate consequence of denying (D1), and that it is non-reductive physicalism which is incoherent. To formalize the argument:

(P1) If physicalism is true, then either (D1) or (D2) is true

(P2) If non-reductive physicalism is true, then (D1) is false

(P3) Ontological explanations are either primitive or further analyzable.

(P4) If ontological explanations are further analyzable, then the analysis is in terms of supervenience or entailment, or in terms of explanatory/theoretical reduction, or in terms of something else.

(P5) If non-reductive physicalism is true, then ontological explanations are not analyzable in terms of explanatory/theoretical reduction.

(P6) Ontological explanations are not analyzable in terms of supervenience or entailment.

(P7) There is nothing else, besides supervenience, entailment, or explanatory/theoretical reduction, in terms of which ontological explanations can be analyzed.

(C) If non-reductive physicalism is true, then there is some *primitive* ontological explanation of mental properties in terms of physical properties, other than identity, such that mental properties are no addition to ontology.

Premises (P1) – (P5) are supposed to be uncontroversial. Premise (P6) appeals to the rejection of (SGH), (SGH*) and (EGH) in section 2.3. Premise (P7) is an open question, and perhaps some further account of ontological explanation or grounding can be offered to give us a reason not to believe (P7) is the case.

I do not think this conclusion should trouble the non-reductive physicalist that badly. I see no reason to be a skeptic about primitive grounding or fundamentality relations[69], and it seems to be a fruitful way to understand the non-reductive physicalist's claim. I regard this class of ontological explanation as just as legitimate a species of explanation as causal explanations, and existential (or modal) explanations.

[69] I use the plural, since one might be a pluralist and hold that there are many distinct grounding relations.

What would an ontological explanation for consciousness in physical terms look like? Consider representationalist theories of consciousness, like those advocated by Tye (1995), Lycan (1998), and Dretske (1995). On these accounts, what it is to be in a conscious state is grounded in the nature of what it is to be in a representational state. Being a representational state is supposed to explain what consciousness *is*, on these accounts, not merely necessitate that it exists. An account like Tye's would be an example of a non-reductive physicalist account of consciousness on which the mental is grounded in the physical representational state, contrary to emergentism.

For our purposes, the significance of the argument is that the emergentist is entitled to make use of this primitive notion of grounding in explaining how emergentism is distinct from non-reductive physicalism.

## 2.5 Distinctness as a Denial of Grounding

Let's return to the ontological emergentist's original distinctness claim. We needed a sense of "distinctness" which the emergentist could affirm but the non-reductive physicalist had to deny, even while the emergentist might join physicalists in denying strong modal distinctness. We now have one:

> (Distinctness) If a phenomenon *e* is ontologically emergent from basal conditions *b,* then *e* is not wholly grounded in *b*.

For instance, Smith's being in pain *p* may supervene upon Smith's overall physical state, but the truth of "Smith is in pain" is not grounded in the facts about Smith's physical state. There is a causal sense of "because" in which Smith is in pain

because he is in the physical state he is in – e.g., because he was stung by a jellyfish while swimming off the coast of Cairns. There is also an existential or modal sense of "because" in which Smith is in pain because he is in the physical state he is in – e.g., if one were to change the chemistry of Smith's brain, Smith would experience pleasure rather than pain. However, there is also an *ontological* sense of "because", and the emergentist holds it is not true in this sense that Smith is in pain because he is in the physical state he is in. It is not *in virtue of* being in such and such a physical state that he is in pain, because what it is to be in a particular physical state and what it is to be in pain have two different natures.

The special attraction of grounding for emergentists goes back to issues about reference. There is something particularly bizarre and gerrymandered about reinterpreting the language of "thoughts" and "desires" and "beliefs" in terms of a complex of physical states. While, in principle, one might successfully map every occurrence of "__ believes the sky is blue" onto a particular disjunction of brain states in different brains, it seems as though *why* we'd have the resulting collection of brain states would be explained by their being the base for "___ believes the sky is blue", *not* that they'd form some sort of natural pattern which suggests the content of "____ believes the sky is blue".[70] If reference fixes to properties which are more fundamental or "natural", and the reference of mental terms seems not to be fixed by

[70] Contrast this with artifact kinds, like "table". While "table" might have a complex physical realization, when the social and mental facts are included the reference of table becomes quite natural.

anything else, perhaps this is a reason to think that mental properties are fundamental cuts in the structure of the world.

Understanding emergentism in this way is similar to a recent proposal by Elizabeth Barnes.[71] On Barnes's account, an emergent phenomenon is one which is dependent upon and sustained by its basal conditions, but which is nonetheless fundamental. She contrasts this with the "levels ontology" typically associated with emergence, on which the hierarchical structure of levels determines the facts about fundamentality: the simple stuff at the bottom is the most fundamental, and the higher you go in the hierarchy, the more complex and the less fundamental things get. In place of this, Barnes's version of emergence is one on which the "higher level" phenomena can nonetheless be fundamental: minds remain the most salient candidate for emergence, but other examples she considers without committing to

[71] Barnes (2012). A brief summary of her project, taken from the abstract: "I argue for a new way of characterizing ontological emergence. I appeal to recent discussions in metaontology of fundamentality and dependence, and show how emergence can be simply and straightforwardly characterized using these notions. I then argue that many of the standard problems for emergence don't apply to this account: given a clearly specified metaontological background, emergence becomes much easier to explicate. If my arguments are successful, they show both a helpful way of thinking emergence and the potential utility of discussions in metaontology when applied to first-order metaphysics."

include persons[72], living beings[73], property tropes[74], certain quantum phenomena[75], and other cases[76].

My account is similar to Barnes's in appealing to a notion of ontological explanation to explain how emergent phenomena are distinct from their basal conditions. Particularly, if a property is not grounded at all in its basal conditions or anything else – as seems plausible for consciousness – then the property qualifies as fundamental, and it is emergent on both my account and Barnes's account. However, my suggestion differs in a few notable ways. First, Barnes's account supposes a much weaker connection between emergent properties and their basal conditions – emergents are *sustained* by their basal conditions – whereas the account I offer in this section is intended to be consistent even with the metaphysical necessitation of

---

[72] That is, if one adopts the views of Merricks (2003)

[73] That is, if one adopts the views of Van Inwagen (1990)

[74] That is, if one holds that property tropes are fundamental, but nonetheless depend upon other property tropes to exist: e.g., mass tropes are fundamental, but a thing must have size and shape to have mass – 'point masses' are a useful theoretical construct but not real.

[75] That is, on certain interpretations of phenomena like quantum entanglement and the generation of a field from constituent particles.

[76] She notes a novel use for emergence in the case of someone with a gunky ontology, on which everything has proper parts descending downward, infinitely smaller, whereby a gunk-theorist could nonetheless hold that some intermediate level is fundamental, and thus emergent from its parts.

emergents by their basal conditions. Second, Barnes's account eschews talk of hierarchical levels, but I see no reason to reject talk of "levels", provided that we recognize, as she does, that the facts about fundamentality and grounding need not line up with the facts about supervenience represented by the various levels.[77] Third, on my account, a phenomenon might count as emergent without being fundamental, provided that it is grounded in some other emergent phenomenon. For instance, suppose that persons and conscious experiences have the same physical basal conditions, and that conscious experiences are fundamental. Suppose further that facts about persons aren't fundamental, but that these facts are wholly grounded in conscious experiences like memories, and not grounded in their physical basal conditions. On my account, persons would qualify as emergent, but not on Barnes's account.

Finally, and most significantly, my account does not require that an entity be fundamental in order to qualify as emergent, but only that it not be *wholly* grounded in its basal conditions. My account is consistent with a notion of *partial* grounding, and thus with weaker and stronger kinds of emergence.

## 2.6 Partial Grounding and Real Definitions

I have encouraged regarding the notion of grounding as conceptually and metaphysically primitive, not admitting of further analysis. However, there may be

[77] As I understand it, part of Barnes's motivation in avoiding "levels" is a desire to avoid Kim's overdetermination problem. However, I think we should be able to avoid Kim's overdetermination problem without losing the "levels ontology", so long as the emergentist does not hold to emergence of novel causal forces. See Chapter 3.

certain illuminating and useful ways of explaining and thinking about the notion of grounding without giving an analysis of it; particularly, these may help make clear how grounding could be "whole" or "partial".

One useful notion is that of a thing's real nature, or essence. Another useful metaphor involves the idea of trans-world counterparts. A final useful metaphor involves the idea of a "real definition." Put together, these may illustrate the sense in which an entity can be partially or wholly grounded in another.

**2.6.1 Essences.** For someone who believes in essences, if *A* is part of the essence of *B*, then the fact that *B* exists is partly grounded in the fact that *A* exists. Essential to Socrates is the property of being a human. So, being Socrates is partly grounded in being a human. Of course, being Socrates is grounded in other things too. We might assume that if we took all of the essential properties of Socrates, then we'd have the whole ground of Socrates, but this need not be the case. We shouldn't conflate the notion of grounding and the notion of being an element in something's essence. Kit Fine (2010) argues that the two should be kept sharply separate.[78] For

[78] "I think it should be recognized that there are two fundamentally different types of explanation. One is of identity, or of what something is; and the other is of truth, or of how things are. It is natural to want to reduce them to a common denominator - to see explanations of identity as a special kind of explanation of truth or to see explanations of truth as a special kind of explanation of identity or to see them in some other way as instances of a single form of explanation. But this strikes me as a mistake. And it seems to me that [it is an error to attempt] to assimilate or unify the concepts of essence and ground. The two concepts work together in holding

example, the fact that someone is a philosopher is grounded in the fact that Socrates is a philosopher – but it is surely not an essential part of someone's being a philosopher that Socrates is a philosopher. The notion of grounding is thus *broader* than that of essence – being an essential property may be one of many kinds of grounding.

An emergentist can distinguish a *stronger* type of emergence on which the facts about the emergent phenomena are *wholly* grounded in the facts about basal conditions (e.g., *no* part of a thing's essence or other grounds includes its basal conditions), and *weaker* types of emergence on which the facts about the emergent phenomena are *only partly* grounded in the facts about their basal conditions (e.g., a proper part of a thing's essence includes its basal conditions). For example, perhaps it is essential to my thought of water that the thought have a certain causal relation to water in space and time, but the thought is also essentially conscious. Then, the thought is weakly but not strongly emergent.

I also do not think someone has to accept classical essentialism full-force in order to accept that grounding could work in a similar part/whole way. We might use the term "nature" to remain neutral between essentialism and non-essentialist accounts of the grounding or constitutive conditions of various phenomena. Emergent properties would then be those whose natures included something over and above the natures of their basal conditions.

---

up the edifice of metaphysics; and it is only by keeping them separate that we can properly appreciate what they are and what they are capable of doing together."

**2.6.2 Counterparts.** A platypus is a monotreme. A monotreme is a mammal which lays eggs. The fact that something is a platypus is partly grounded both in its being a mammal and in its laying eggs. There is no possible world in which there is a platypus which is a fish, or in which a platypus has live births, because *whatever it is* that is a fish or has live births in that world, it isn't a platypus. However, we might legitimately ask in such a world which of two species is the closest counterpart[79] of the platypus in our world: platypus*, a species which is in all other respects like our platypus, but has live births, or platypus+, a species which is in all other aspects like ours, but which feeds its young by regurgitation or some means other than providing milk. Alternatively, we might ask whether a world in which the platypus was replaced with the platypus* or the platypus+ is closer to our world.

What would this tell us? It could give us a sense of how much weight each property bore vis-à-vis the property of being a platypus. Perhaps they are equally weighted, and there is no sensible answer to the question. But we can imagine a case in which two grounds are not equally weighted: the fact that someone is a philosopher is grounded both in Aristotle's being a philosopher and in what it is to be a philosopher, but in some sense the world where Alexander the Great was a philosopher is closer than the world in which being a philosopher involves performing magic tricks. Thus, that someone is a philosopher is grounded in the

---

[79] Note that I do not think that one has to accept David Lewis's account of trans-world "identity" in terms of one's closest counterpart at another possible world in order to find something interesting in it about the relative weight of various elements within the natures of things.

essence of being a philosopher to a greater extent than it is grounded in Aristotle's being a philosopher.

**2.6.3 Real Definitions.** Earlier, I noted that someone might hold (WSGH), the view that grounded facts supervene on their grounds. However, if there are partial grounding facts, then we should grant that a fact can *partly* ground another fact without necessitating it.

For instance, the fact that Jeremy is a bachelor is partly grounded in the fact that Jeremy is male, and partly grounded in the fact that Jeremy is part of a society which practices marriage, but neither fact alone necessitates that he is a bachelor. The property of being a brilliant philosopher is not a mere conjunction of two properties – a person can be brilliant and a philosopher, yet do mediocre philosophy. Nonetheless, the property of being a brilliant philosopher is partly grounded in being brilliant, partly grounded in being a philosopher, and partly grounded in a certain relation obtaining between brilliance and philosophy.

It may be helpful to think of facts as structured entities, representable parts of the world consisting of pairings of objects and properties. It may also be helpful to think of properties as having "real definitions": various other properties which serve as criteria for whether a thing has the target property or not. It may then be helpful to think of one fact as grounding another when the objects and properties in one fact are elements in the "real definition" of the objects and properties in the other. For instance, the fact that Jeremy is a bachelor involves the pairing of object named by "Jeremy" with the property of being a bachelor, and being a bachelor is really defined as "being male and never being married and being eligible for marriage and being part of a society which practices marriage."

We might accept that, in principle, for every property there is some *canonical definition*: that is, a definition of the property in terms of the fundamental elements which ground it. If being the Civil War is grounded in an extensive conjunction and/or disjunction of fundamental physical, social, and psychological facts, this conjunction or disjunction is the canonical definition of the civil war.

How far one wishes to go with the notion of a "real definition" is up for grabs. The term "definition" here involves a certain degree of hyperbole. The terms in our *language* do not generally have strict definitions. Neither are we likely to find strict necessary and sufficient conditions for most of our *concepts*, especially the most useful and important ones. Speaking of a "real definition" of what it is to be an *F* in terms of $G_1$, $G_2$, $G_3$… does not commit one to there being an appropriate definition of the concept for *F* or the word for *F*, especially not one in terms of $G_1$, $G_2$, $G_3$…. There may be nothing "definition-like" about the set of properties which ground another property except for their being structured together in a some complex way or another. The only commitment to there being "real definitions" for properties here is the commitment entailed by fundamentalism: for every non-fundamental property, there is some number of fundamental properties which ground it.

Supposing we adopt the metaphor of "real definition", we could say that the strongest form of emergence is one in which *no part* of the definition of the emergent entity includes its basal conditions, and that weaker forms of emergence are those in which a proper part of the definition of the emergent entity includes its basal conditions, but something else (the "novelty") is also included. Combining this with the notion of relative weight within the nature of a thing (from 2.6.2), we could say that a *weaker* kind of emergence is one in which the relative weight of the basal

conditions in the real definition is greater, and a *stronger* kind of emergence is one in which the relative weight of the basal conditions is lesser.

What's the advantage of admitting stronger and weaker kinds of emergence? It allows us to distinguish the strongest case of ontological emergence (that is, consciousness) from various weaker forms of emergence which one might very well still wish to regard as ontologically significant.[80] While the physicalist can acknowledge a distinction between our theories of phenomenal consciousness, and our theories about psychology, economics, sociology, biology, and ultimately physics, the physicalist is committed to these phenomena all having the same nature – being grounded in the same sorts of facts about microphysics. Yet the relationship between physics and macroeconomics seems very different from the relationship between physics and phenomenology, and the issue is not that one is a "bigger jump" than the other on the hierarchy (both involve a big jump, and if anything macroeconomics is "larger scale" than phenomenal experience). The issue is that, unlike in the case of economics, neither the supervenience of the phenomenal on the physical nor anything characteristic about the nature of physical properties seems to give an *ontological explanation* of what it is to be a phenomenal property. This is not an issue if one maintains that phenomenal properties are identical to physical ones (we might just be really confused). However, if one rejects the identity of these properties as the non-reductivist does, and isn't persuaded by existing ontological explanations of consciousness (like Tye's Representationalism), then emergentism seems like a reasonable position to take.

[80] See Chapter 5

## 2.7 Where'd the Mathematical Archangel Go?

One might wonder where C. D. Broad's notion of "ideal deducibility" has gone to. After all, nothing in the claim that emergent properties are not wholly grounded in physical properties rules out that emergent properties could be deduced *a priori* from physical properties by a mathematical archangel. I think this is a legitimate position for an emergentist to hold: even if a mathematical archangel could deduce the emergent properties *a priori*, they might still be emergent.

However, in Chapter 4 I *offer* a new and more complex way in which "ideal reasoner" tests of this sort can be taken to offer a guide to whether there is a physicalistically acceptable ontological explanation for some higher-level phenomenon, and thus whether a phenomenon qualifies as emergent. I suggest that the question of whether or not there is an ontological explanation of the *A*s in terms of the *B*s becomes a question of whether the *A*s could be *a priori* deduced by an ideal reasoner *by means of analysis alone*. There is much to say about what precisely the relevant sort of analysis would consist in, so I will hold off on this question until later.

## 2.8 Preliminary Conclusions for the Emergentist

So, we've uncovered one strategy for the emergentist about phenomenal consciousness to take in order to develop a coherent account. The emergentist can affirm the sufficiency of the physical for the phenomenal alongside the physicalist, while maintaining – contrary to the physicalist – that phenomenal facts are not grounded in physical facts. For the emergentist, the phenomenal facts are distinct from the physical facts in a way which the physicalist cannot grant. Because grounding is not entailed by supervenience, the emergentist is free to interpret the

sufficiency of the physical for the phenomenal as consisting in supervenience with metaphysical necessity: a position Horgan calls "Moorean Minimal Emergentism".[81] Because it is the non-reductive physicalist, rather than the emergentist, who has to introduce a distinctive kind of ontological explanation into the dialectic, the emergentist does not have to be concerned with skepticism about ontological explanation, "grounding", "essences" or the like: if there are no such explanations, then it is the non-reductive physicalist who must either give up non-reductivism (accepting type-identity) or give up physicalism (becoming an emergentist).

In Part B, I will pursue an alternative strategy. Instead of clarifying the distinctness claim in order to sharpen the line between emergentism and non-reductive physicalism, I will attempt to clarify the sufficiency claim in a way to sharpen the line between emergentism and traditional dualism.

## PART B

## §3. Emergence and Sufficiency

### 3.1 Is Emergence Necessary?

One way to understand the emergentist's claim that phenomenal states "arise from" physical states, or that physical states are "sufficient for" phenomenal states, is in terms of the supervenience of phenomenal states on physical states with *metaphysical necessity*. In Part A, I argued that this is consistent with the claim that phenomenal states are not grounded in physical states. However, there are three *prima facie* reasons why an emergentist might wish to resist supervenience with metaphysical necessity.

[81] Horgan (2010)

First, it seems to violate Hume's dictum, that there are "no necessary connections between distinct essences." Whether or not Hume's dictum is a valid metaphysical principle and what the relevant notion of *distinctness* is supposed to be (let alone whether Hume even held the dictum) can be debated.[82] But suppose that someone does hold it, and interprets it as meaning that there are no necessary connections between properties or entities where neither is identical to or grounded in the other. This rules out the version of emergentism described above.

Second, an emergentist might be strongly motivated by conceivability arguments for the possibility of Zombie worlds or Invert worlds, where the physical facts remain the same but the phenomenal facts are different, even while wishing to resist the whole of the traditional dualist's conclusions from these arguments. Emergentism with metaphysical necessity has to bear the same burden as physicalism when it comes to resisting these arguments, but it has even fewer tools at its disposal to do so. (For example, the physicalist might attempt a dual-concept strategy; the emergentist cannot.) It's coherent to accept that my experiencing colors as I do and not in an inverted way is necessitated by my brain state but not grounded in it. But why hold it? If it's not grounded, why think it's necessitated when it *seems* as if it's not?

Third, it seems on the face of things to conflict with the emergentist's traditional claim that emergents are not ideally deducible from their physical basal conditions, thus tending to collapse emergentism into *a posteriori* physicalism.

[82] See Stoljar (forthcoming)

Suppose an emergentist wishes to reject supervenience with metaphysical necessity. One alternative is to hold that mental properties supervene upon the physical with a kind of *nomic* necessity, in virtue of certain contingent trans-ordinal laws. However, many traditional dualists hold this same position: physical states are sufficient for mental states in the sense of nomically necessitating them.[83] It seems as though emergentism risks collapsing into traditional dualism.

I have three goals in Part B. First, I wish to clarify the metaphysical position I have been calling "traditional dualism" and precisely what aspects of it contemporary emergentists may find objectionable. Second, I wish to further illuminate both the concept of "sufficiency" and its connection to supervenience, in order to produce a notion of *manipulative* supervenience which seems a bit richer than the present canonical forms of supervenience and more in line with the emergentist view of the world. Finally, I wish to show how this concept of manipulative supervenience can fulfill the emergentist's sufficiency claim in a way which is intermediate between the positions of the traditional dualist and the non-reductive physicalist.

### 3.2 What is Traditional Dualism?

In the early works of C. D. Broad and Samuel Alexander, emergentism in the mind was set up in contrast to substance dualism, much as emergentism in biology was set up in contrast to substance vitalism.[84] Similarly, the sort of emergentism

---

[83] For example, see Chalmers (1996)

[84] A common point of misunderstanding is to think that the British emergentists were vitalists in the sense of positing vital substances or entelechies. Their view was

which I have been defending is meant to be a dualist position which is in some sense weaker than what I have been calling "traditional dualism".

The typical way to divide up dualist positions has been in terms of what there is a duality *of*, and whether there is a duality of properties, of property instances, of events, of entities, of objects, or so on. Another approach has been to divide dualist views based on their approach to the causal powers of the non-physical, and whether they are interactionist, epiphenomenalist, involving predetermined harmonies, or so on. I doubt either approach is particularly helpful in discussions of what emergence is. It seems as if one might be an emergentist about all sorts of these "grammatical" categories of phenomena and still hold a position distinct both from physicalism and from traditional dualism. It also seems to me that causation is a messy enough topic in its own right to require separate treatment. Rather, dualists might be more interestingly distinguished in terms of what degree of dependence they accept of the mental on the physical.

All dualists hold that there are at least two distinct types of *something or other* in the world: that which has a nature which is wholly physical, and that which has a nature which is not wholly physical. For all dualists about the mind, mental phenomena are *ontologically independent* from physical phenomena. (This "independence" was identified in Part A with a lack of grounding). The traditional dualist holds that this ontological independence indicates a failure of a *different kind* of dependence, which we might call *existential dependence*, or the dependence

---

intentionally proposed *against* and in contrast to substance vitalism. See McLaughlin (1992).

of one thing upon another for its existence.[85] The traditional category of "substance" might be expressed as something on which other things depend for their existence, but which does not itself depend upon anything for its existence.[86] We might say that the traditional dualist holds both:

> Grounding Dualism (GD): mental phenomena are not wholly grounded in physical phenomena.
>
> Partial Existential Independence from the Physical (PEIP): mental phenomena do not wholly depend for their existence on physical substances.

The emergentist accepts (GD), but rejects the traditional dualist's claim (PEIP) that mental phenomena don't wholly depend for their existence on physical phenomena. The emergentist holds that mental phenomena *do* wholly depend for their existence on physical phenomena. Traditional dualism can be further categorized into three degrees of strength, based on whether or not the traditional dualist accepts these further claims:

---

[85] Correia (2008)

[86] Corkum (2012)

Partial Existential Dependence on the Mental (PEDM): mental phenomena at least partly depend for their existence on mental substances.[87]

Total Existential Independence from the Physical (TEIP): mental phenomena do not even partly depend for their existence on physical substances, but wholly depend for their existence on mental substances.

*Minimal Traditional Dualism* is the conjunction of (GD) and (PEIP). A *Substance Dualist* further holds (PEDM), the claim that mental phenomena partly depend for their existence on a mental substance *in addition to* some physical substance. A *Cartesian Substance Dualist* holds the strongest form of traditional dualism, by adding the claim (TEIP). Thus, the Cartesian holds that mental phenomena depend *wholly* on the mental substance and can "float free" of the physical substance.[88] The emergentist rejects (PEDM) and (TEIP) for the same

[87] I am using the term "substances" to indicate something which is itself totally existentially independent from anything else, without meaning to connote by "mental substances" any ghostly translucent-green ectoplasms or the like.

[88] Lowe (2006) distinguishes his non-Cartesian substance dualism from Cartesian substance dualism in the following way: a Cartesian holds that persons (or minds) are substances which have no physical properties, whereas a non-Cartesian substance dualist accepts that persons (or minds) are substances which may have some physical properties without having only physical properties. This seems

reasons that the emergentist rejects (PEIP): the emergentist holds that mental phenomena wholly depend upon the physical for their existence and nothing further – the mind is *ontologically independent* from but wholly *existentially dependent* upon the physical.

There are two questions which need to be clarified further if the notion of existential dependence is to serve to distinguish emergentism from traditional dualism. The first involves defining what the relevant notion of existential *dependence* is. The second is determining the sense in which this dependence could be "*total*" or "*whole*", as opposed to merely "partial".

## 3.3 From Existential to Manipulative Dependence

**3.3.1 The Necessary Condition Account.** We might at first interpret "existential dependence" as indicating a logically necessary condition – there are *no* worlds in which the dependent state occurs without the state it depends upon.

> (Necessary Condition Account) if *y* depends for its existence on *x*, then necessarily if *y* exists, then *x* exists.

However, this is far too strong to be the relevant sense of existential dependence. Both non-reductive physicalists and emergentists alike acknowledge the *weak* modal distinctness – or "multiple realizability" – of the phenomenal and the physical: that is, that it is possible that the same phenomenal properties should

---

consistent with my account here. Lowe's motivation for substance dualism involves reflection on the problem of personal identity.

occur, yet be realized by different physical properties.[89] Suppose we want to say that my being in pain depends upon my being in the brain state that I am in. On the Necessary Condition Account, it follows that I could not be in pain without being in this exact brain state. However, surely it is logically possible that I could be in pain and yet be in a slightly different brain state – or, for that matter, that my pain states could have been realized in a brain with a very different structure.

**3.3.2 The Generic Account.** Perhaps the problem is that the Necessary Condition Account is too specific. Perhaps all we must hold is that there must be *some* physical state which realizes the relevant pain state. Call this the generic account:

> (Generic Account) if *y* depends for its existence on *F*, then necessarily if *y* exists, then some *x* exists such that *Fx*.

The Generic Account forms the first conjunct of Kim's definition of Strong Local Supervenience.[90] However, the Generic account leads to two problems. First,

---

[89] One might suggest that *x* could be given a disjunctive account to account for multiple realizability: necessarily, if *y* exists, then either $x_1$ exists or $x_2$ or . . . However, this won't capture the sense in which *y* is supposed to depend for its existence on the *particular* realizer which in fact occurs, and not merely a disjunctive state of affairs.

[90] Kim (1993), "Varieties of Supervenience": *F* supervenes upon *G* iff *necessarily for every f in F there is some g in G* such that necessarily every *x* that has *f* has *g*.

many physicalists as well as emergentists would not accept that this sort of existential dependence holds between phenomenal and physical states. Jackson (1998), for example, holds that physicalism is a contingent truth, and so allows that physicalists may accept that there is some possible world in which my being in pain is realized by an epiphenomenal ectoplasm. Likewise, while emergentists might hold that emergent properties must by their very nature have some basal conditions or other, it's not clear why we must hold that the basal conditions are necessarily physical in every way the world could have been. Perhaps "emergence from ectoplasm" seems conceivable.

Second, the Generic account is *too* general. We want to say that emergent properties depend upon their *specific* basal conditions, not merely there being some similar basal conditions or other. There is a sense in which my pain depends upon this specific brain state, not just on being in *a* brain state. While my pain could have been realized by another brain state, there is something about this specific brain state which is relevant to my pain.

**3.3.3 The Counterfactual Account.** One might then be tempted to identify existential dependence with a kind of counterfactual dependence:

> (Counterfactual Account) if *y* depends for its existence on *x*, then in the nearby worlds where *x* does not exist, *y* does not exist.

However, Elizabeth Barnes notes that my own existence is counterfactually dependent on the prior existence of my parents: were they never to have existed, I

wouldn't have existed either.[91] Nonetheless, my own existence would continue even if the existence of my parents ceased. Mere counterfactual dependency is far too weak, insofar as it permits a kind of diachronic dependency.

**3.3.4 The Sustaining Account.** As an alternative, Barnes offers an account in terms of one thing "sustaining" the existence of another, a relation which is necessarily synchronic:

> (Sustaining Account) if *x* depends for its existence on *y*, then for every moment *t* at which *y* exists, in the nearby worlds in which *x* does not exist at *t*, *y* does not exist at *t*.[92]

Fortunately, my parents do not sustain my existence in this sense. On the other hand, perhaps a complex object does depend upon its parts in this sense – for each time at which a statue exists, were it not for the parts which compose it, the object would not exist. The Sustaining account captures an intuitive way in which wholes can depend upon their parts. It is tempting to think that this is the relationship between the mind and the body: were it not for my body at *t*, my mind wouldn't be there at *t*.

---

[91] (forthcoming). Not that Barnes refers to "existential dependence" as "ontological dependence", and what I call "ontological dependence" she calls "metaphysical dependence".

[92] Related to this is the notion of "permanent existential necessitation" in Correia (2008), 1016.

However, similar concerns which applied to the Necessary Condition account also apply to the sustaining account. Perhaps there's no nearby world in which my pain isn't realized by some brain state or other – the world in which my pain occurs in a silicon chip is very far away. However, it still isn't clear why there wouldn't be a *nearby* world in which my pain occurs at *t* but is realized by a different brain state at *t*. Even in the case of a statue, it is plausible that the world in which the same statue is composed of one fewer molecule of iron is closer than the world in which the statue ceases to exist because of the removal of one molecule of iron.

**3.3.5 The Generic Sustaining Account.** We might consider instead a more Generic version of the sustaining account:

> (Generic Sustaining Account) if *y* depends for its existence on *F*, then for every moment *t* at which *y* exists, in the nearby worlds in which no *x* exists such that *Fx* at *t*, *y* does not exist at *t*.

This is much closer to what emergentists wish to say about the dependence of the phenomenal on the physical. It is a distinctively synchronic form of dependence that doesn't rule out multiple realizability. However, this account is still subject to the same objections as the earlier generic account: it leaves out the sense in which it's this *specific* brain state happening at *t* that my pain depends upon at *t*. My

experience of pain is not just sustained by the existence of some brain state in general, but the existence of particular brain state.[93]

**3.3.6 The Manipulation Account.** The trouble with all of these suggestions seems to be that the vocabulary of possible worlds is too coarse-grained for our purposes: we are forced to talk of worlds at which a proposition about an entity's existence is or isn't true, a very "on" or "off" question, instead of talking about gradations of different values which a variable might hold. Borrowing a page from recent work in causal explanations[94], we might attempt an account in terms of possible *manipulations* of one variable for another.

> (Manipulation Account) if *y* depends for its existence on *x*, then for every moment *t* at which *y* exists, there are some nearby worlds in which the value of *x* is altered at *t*, and the value of every other variable not on the path between *x* and *y* is kept fixed at *t*, and the value of *y* alters at *t* because of the change in *x*.

I will explore further the various elements in this account (including the relevant notions of being "on the path" and "because of") shortly. For now, it's worth noting that on the Manipulationist account, *y* exists *because* of *x* insofar as *how y* exists can be changed by *how x* exists, not merely that *whether x* exists can be

---

[93] While the emergentist denies a token-identity theory, the emergentist ought to accept a token-*dependence* theory.

[94] Woodward (2005)

changed by *whether y* exists. What is interesting about pains and brain states is not merely that one can eliminate pains by eliminating brains, but that one can *alter* pains by *altering* brains.

Like the sustaining account, this manipulative dependence is supposed to be synchronic rather than diachronic. Unlike the Necessary Condition account and the Sustaining account, the Manipulation account doesn't rule out multiple realizability as a nearby possibility. At the same time, unlike the Generic account and the Generic sustaining account, the Manipulation account is consistent with holding that it is this *particular* brain state on which my experience of pain depends, insofar as changes in this brain state lead to changes in my experience of pain. Given these advantages, I will identify existential dependence with synchronic *manipulative dependence.*

**3.3.7 "On the Path" and "Because".** Part of the definition I offered invokes the notion of keeping the values of everything variable not "on the path" between *x* and *y* fixed. Keeping these values fixed is important to avoiding a kind of pre-emption problem: for instance, if emergent variable *e* (say, a phenomenal state) manipulatively depends upon two independent basal condition variables, $b_1$ and $b_2$, a manipulation of $b_1$ for *e* might be "overdetermined" by a simultaneous alteration in the value of $b_2$. By keeping the value of $b_2$ fixed, we prevent this problem. At the same time, those variables which are along the path between $b_1$ and *e* (e.g., sub-personal psychological states) should be expected to change "along the way" when $b_1$ and *e* are altered, insofar as they are "along the path".

I am taking the notion of a "directed path" from Woodward (2005).[95] To express it more formally: a change in one variable may alter another variable *directly*, unmediated by any other variable. Call this a "direct manipulation". We understand "is a manipulation of" to be the ancestral of the "is a direct manipulation of" relation. If *b* is a manipulation of *e* but not a direct manipulation of *e*, then for any other variable *p* which mediates this manipulation (so that *b* is a manipulation of *p*, and *p* is a manipulation of *e*, and it is thereby that *b* is a manipulation of *e*) the variable *p* is said to be "on the path" between *b* and *e*. (Further, anything which entails *b* or is entailed by *b* should count as being on the path between *b* and *e*.)[96]

I've also relied upon the term "because" in the definition I've offered of existential dependence, thereby appealing to a notion of dependence or explanation in attempting to explain dependence. This rightly implies that the definition I've offered is not an analysis or reduction of existential dependence. This synchronic, upward "because" is supposed to be a conceptual and metaphysical primitive, much as the "in virtue of" in the earlier account of ontological dependence is primitive, and the "because" in accounts of diachronic causation is primitive. While this doesn't amount to a reductive definition of existential dependence in terms of synchronic

---

[95] pgs. 38-61

[96] Consider that *b* entails *b* or *c*, but clearly *b* or *c* should not be counted as a distinct means of manipulating *e*; consider also that if *a* entails *b*, then *a* will naturally count as a means of manipulating *e* and *b* will be "along the path", but if we take *b* as our starting point rather than *a*, then *a* ought not to be included as a distinct manipulation base.

manipulative dependence, it does seem informative enough to move on to clarifying how this sense of dependence can be used to divide up the dualist camps.

### 3.4 Total and Partial Manipulative Dependence

In what sense can synchronic manipulative dependence be total or partial? Here is a simple enough explanation: *partial* dependence is the claim that the dependent can be synchronically manipulated by means of an intervention upon the base, and *total* dependence is the further claim that the dependent cannot be manipulated by means of intervention upon *anything else* besides the base – at least, not anything else which isn't already on the path between the dependent and the base.

On this interpretation, the differences between emergentism and traditional dualism can be understood in the following way. First, let's look at the three traditional dualist positions:

> (i) The *minimal traditional dualist* holds that phenomenal states can be synchronically manipulated by interventions on something fundamental, other than physical states.[97]
>
> (ii) The *Substance dualist* holds (i), and further that phenomenal states can be synchronically manipulated by interventions on an independent mental substance.[98]

[97] The minimal traditional dualist may still hold that having brain states is nomically necessary for having mental states, as Descartes did.

(iii) The *Cartesian*[99] *substance dualist* holds (i) – (ii), and further that phenomenal states can not be synchronically manipulated by interventions on physical states[100].

On this interpretation, then, the emergentist must deny (i) – (iii), holding that:

---

[98] So long as the mind is "hooked up" to the brain, these manipulations of consciousness will cause changes in the brain, but were the mind "unhooked", the substance dualist holds it could continue to be manipulated by changes to the conscious substance.

[99] Obviously, an interactionist holds that phenomenal states can be *diachronically* and *causally* manipulated by interventions on physical states. However, the interactionist's insistence that the relation between brain and mind must be specifically *causal* seems to rule out that the mind depends for its existence on the brain, and thus can be synchronically manipulated in the sense of "manipulation" under discussion here. Note that I don't mean to assert that this is the view of the historical Descartes; it may be a bit of a caricature.

[100] Given interactionist assumptions, phenomenal states may of course be *diachronically* (that is, causally) manipulated by changes in physical states on a Cartesian view.

(iv) phenomenal states can be synchronically manipulated by interventions on physical states.

(v) phenomenal states can not be synchronically manipulated by interventions on anything fundamental other than physical states, including mental substances.

The conjunction of (iv) and (v) is supposed to capture the emergentist's assertion that phenomenal states are totally dependent upon their underlying physical states for their existence – given sufficient changes to the underlying physical states, the phenomenal states will cease to occur. Notably, this sounds very similar to an intuitive and off-hand definition of supervenience: *nothing makes a phenomenal difference without making a physical difference.*

Consider how a guitar can be tuned to different keys or chord structures. Two guitars, despite being of difference sizes and shapes and having strings of different thicknesses, can be tuned to the same key. Still, the tunings of a guitar are closely connected to the physical properties of the various strings – their composition, length, tightness, temperature, and so on. There is no way to change the tuning of the guitar without changing the physical properties of the guitar strings. Nonetheless, the tunings of the guitar are not identical to physical properties of the guitar strings, since the same tuning can be realized by physically very different strings (or something other than strings). Guitar tunings could be said to *supervene* on the properties of guitar strings. There can be no difference in the tuning properties of two guitars for any reason other than a difference in the properties of

their strings (including relational properties, such as the humidity and temperature of their environment). We might express the intuitive notion of supervenience as follows, where $A$ and $B$ are sets of properties:

> *Intuitive Supervenience*: $A$ supervenes on $B$ if and only if nothing can make a difference in the distribution of $A$ except by making a difference in the distribution of $B$.

Understandably, supervenience has long been considered a useful way of describing the relationship between the mind and the physical world.[101] Emergentists have historically accepted this intuitive notion of supervenience, and supervenience has been offered in various ways as a definition of the minimal commitments of physicalism. Traditional dualists necessarily deny supervenience.[102] If we can interpret "makes a difference" in terms of possible manipulations, then the intuitive notion of supervenience is nicely captured by the claim of total manipulative dependence: $A$ supervenes upon $B$ if there is nothing else but $B$ (and that which is on the path between $B$ and $A$) by which $A$ can be synchronically manipulated.

---

[101] For instance, Socrates considers (and rejects) this same sort of "tuning" analogy in the *Phaedo* – an analogy on which the soul is a tuning of the body.

[102] Excepting a kind of supervenience with nomic necessity.

## 3.5 Existing Varieties of Supervenience

**3.5.1 Global Supervenience.** What is the relationship between this intuitive notion of supervenience, and the more formal notions of supervenience developed in contemporary philosophy over the last several decades? There are three distinctions in Supervenience relationships which are well established: Local versus Global Supervenience, Strong versus Weak Supervenience, and Supervenience with Metaphysical versus Nomic (or Nomological) Necessity.

Global Supervenience is the following claim: "*A Supervenes on B if and only if every two possible worlds which are identical in the distribution of B properties are identical in the distribution of A properties*." Imagine the life of Francis of Assisi. Assume that in this world, St. Francis lived a good life. It does not seem as though there could be another world, exactly like our own with respect to the events of the life of St. Francis, the virtues he possessed, and the context they occurred in, yet in which St. Francis was an evil person rather than a virtuous person. So, it seems as though evaluative properties like the goodness of a life globally supervene on descriptive properties like virtue, action, and so forth.

Frank Jackson (1998) developed a variation on global supervenience, or *minimal* global supervenience:

> *Minimal Global Supervenience*. A supervenes on B iff any world which is a minimal B-duplicate of the actual world is an A-duplicate simpliciter.

Jackson's variation is an attempt to formulate physicalism as a thesis which is true in the actual world, without needing to be true of all possible worlds. Suppose

we believe that there are no non-physical substances in the actual world, but there is some possible world exactly like our own physically which also contains a bit of epiphenomenal ectoplasm off in another corner of the galaxy. If this is correct – if physicalism is not a necessary truth in all worlds – then it would not be the case that phenomenal properties globally supervene on physical properties. However, consider a world which is a *minimal* physical duplicate of our own world – a world which is an exact copy of our own physically, and in which nothing additional is added (there is no epiphenomenal ectoplasm). If it's true that in every such world, the phenomenal properties are distributed in exactly the same way as in our world, then phenomenal properties minimally globally supervene on physical properties.

**3.5.2 Weak and Strong Local Supervenience.** Global Supervenience relationships, which quantify over properties and pairs of possible worlds, differ from Individual Supervenience relationships, which quantify over objects and properties. Individual Supervenience theses were divided into *weak* and *strong* supervenience relationships by Jaegwon Kim's early work on Supervenience as follows:

> *Weak Supervenience.* $A$ weakly supervenes on $B$ iff, necessarily, for every object which has property $F$ in $A$, then there is some property $G$ in $B$ which the object has, such that every object in this world which has $G$ also has $F$.

> *Strong Supervenience.* $A$ strongly supervenes on $B$ iff, necessarily, for every object which has property $F$ in $A$, then there is some property $G$ in $B$ which the object has, such that necessarily every object which has $G$ also has $F$.

Suppose that Socrates has the property of being a virtuous person, and the property of being virtuous *strongly* supervenes on the particular virtues that one has. Then there is some collection of particular virtues which Socrates has – say, the virtues of having courage, wisdom, and honesty – such that everyone in any world who has those virtues is a virtuous person. Of course, St. Francis had very different virtues than Socrates, but he was still a virtuous person. But – assuming the properties of being a virtuous person strongly supervene on the properties of having certain virtues – there could not possibly be a person who had the same virtues as Socrates, or the same virtues as St. Francis, and yet failed to be a virtuous person.

Weak supervenience differs from strong supervenience in that the second "necessarily" is missing from the definition – it need not be that in every world, every object which has the *G* property has the *F* property; it need only hold that in our world, every object which has the *G* property has the *F* property.[103]

---

[103] For example, suppose that someone holds artistic beauty is world-relative (as opposed to culturally or individually relative). Such a person might hold that the beauty of a painting weakly supervenes on its physical properties (the distribution of the paint on the canvas) without holding that it strongly supervenes. Such a person could say: "necessarily, every beautiful image has a certain physical constitution and arrangement, and there is no way that another image in this world with the same physical constitution and arrangement could fail to be equally beautiful. However, there is some possible world in which another image, exactly like that one in this world, would fail to be beautiful."

The distinction is relevant to formulations of the supervenience of phenomenal properties on physical properties. Jaegwon Kim observed that many would be non-reductive physicalists (including Davidson) expressed the supervenience of the phenomenal on the physical as merely weak supervenience: they allowed that in another world a brain could be arranged in the same way as my own, but fail to produce the same phenomenal properties that mine has in this world. According to Kim, this is insufficient for serious physicalism – the physicalist should hold that the phenomenal strongly supervenes on the physical, and any world in which a brain is arranged like my own is a world in which the same phenomenal properties are instantiated.

**3.5.3 Metaphysical or Nomological Supervenience.** The final important distinction applies to the second modal operator, "necessarily", in the definition of strong local supervenience. As described above, this "necessarily" is taken to indicate *metaphysical necessity*: that is, it is true that every property which has $G$ also has $F$ across all logically possible worlds. With regard to modality, strong supervenience with metaphysical necessity is akin to Jackson's minimal global supervenience. For simplicity, it's conventional use the phrase "Metaphysical Supervenience" to indicate either minimal global supervenience or strong local supervenience with metaphysical necessity.

However, there is a weaker form of strong supervenience, where "necessarily" is interpreted as *nomological* (or "nomic") necessity, and which quantifies over all possible worlds in which the laws of nature are identical to our own. For a dualist, these laws may include contingent lawful relations between mental and physical properties, synchronic "trans-ordinal" laws linking the lower level domain to the

higher level domain. This variant, which might be called "nomic supervenience", is consistent with emergentism, but not with physicalism.

Which form of necessity should the emergentist accept for the relationship between the mental and the physical? An emergentist might accept nomic supervenience, as many property dualists like Chalmers (1996) do. However, in order to avoid collapse into traditional dualism, an emergentist must hold that the contingent trans-ordinal laws are not an "addition" to the world, a further means by which phenomenal states can be manipulated besides physical states. There are two options for an emergentist who wishes to hold nomic supervenience. First, it could be that the trans-ordinal laws are not fundamental – that is, the laws are mere "Humean" regularities which hold in virtue of other features of the world. Second, it could be that the trans-ordinal laws are fundamental, but that they are not to count as "variables" in our definition whose values can be intervened upon – trans-ordinal laws aren't the sorts of things which can be manipulated.

If an emergentist does not take one of these two options, then physical states would only be "sufficient" for emergence given the addition of the trans-ordinal law to the world. This suggests that total existential dependence is false, for it seems there is some other means (besides the physical) for manipulating mental states: namely, changing the contingent trans-ordinal laws. Were this so, emergentism would collapse into traditional dualism.

Should the emergentist accept metaphysical supervenience instead? This threatens to create a problem for the subclass of emergentists who wish to deny that emergent phenomena are ideally deducible from their basal conditions. After all,

why wouldn't all of the necessary truths be part of the ideal reasoner's deduction base?

To evade this objection, in Section 4, I will try to develop a version of supervenience intermediate in strength between metaphysical supervenience and nomic supervenience – *manipulative supervenience*, or supervenience with manipulative necessity. Although I recognize that coherent versions of metaphysical supervenience and nomic supervenience can be held by emergentists, I think this version of supervenience best captures the sort of intuitive supervenience to which emergentists have historically been committed. Manipulative supervenience captures the notion of "total existential dependence", in a way which makes clear the differences between emergentism and traditional dualism and incorporates the scientific study of the mind through manipulations of the brain. Central to manipulative supervenience is the notion that there is *nothing further* on which emergent phenomena depend for their existence above and beyond their physical base: the physical basal conditions truly are sufficient for emergence.

## §4 Manipulative Supervenience

### 4.1 An Illustration

Suppose that the set of possible worlds is like a scattering of towns in the Swiss Alps. Between some towns there are railway lines, so that one can travel by train from one town to another, even if the town is very far away. But there are not railway lines between all of the towns, even towns which are close. Sometimes there is no way to travel from one town to another.

In the town of Actual, where we live, the philosophers debate about another town. It is rumored that there is a town which is a perfect physical duplicate of

Actual, but where there is no phenomenal consciousness. They call it "Zombieland." The dualists believe the rumors that Zombieland exists, and the physicalists do not.[104]

There are also a couple of perplexing emergentists in the town. Some say that Zombieland does not exist (but not for any explicable reason, mind you). Some say Zombieland does exist, but that it is in a distant canton with different laws of nature. Some admit Zombieland might even be a close neighbor of Actual, with the same laws, on the opposite side of the mountain. When they speculate, they are inconsistent. But when they stop speculating, they all insist together:

"There are no trains to Zombieland."

Whether Zombieland exists or not, careful investigation has discovered that there are no trains from Actual which go there. In fact, there is no town in all of Switzerland in which there is consciousness which offers train service to a duplicate town in which there is no consciousness, or vice-versa. It is not a matter of distance, similarity, or canton law. It is a matter of inaccessibility.

### 4.2 Manipulation Worlds

Obviously, one cannot take any form of transportation from the actual world to another. What I mean by there being a "train" from world *w* to world *w** is this: world *w** counts as a successful *manipulation* of world *w*. As discussed in the previous section, a manipulation is *through* some variable *for* some other variable.

---

[104] Similar things are said about the town of "Invertica", which is supposed to be a physical duplicate of Actual where the phenomenal consciousness is inverted or different.

To be slightly more formal, when I call $w^*$ a successful manipulation of $w$ through $b$ for $e$, I mean that there is a change in the value of $e$ between $w$ and $w^*$ because of a change in the value of $b$.

Again, the traditional account of "manipulation" is taken from Woodward (2005), and it describes diachronic causal relationships. On Woodward's account and my own, a "manipulation" or "intervention" does not require that there be any kind of *agency* (human or otherwise) which is or could possibly be in a position to intervene upon or manipulate the "intervened upon" or "manipulated" variable.

I have been adapting this account specifically for synchronic, non-causal relationships. Emergentists hold that emergent properties are neither identical with nor reducible to their basal conditions, but they are manipulable by their basal conditions. Suppose we define successful manipulation in this synchronic sense, for emergent $e$ and basal conditions $b$, as a world $w^*$ with the following conditions:

(i) *Intervention*: the value of $b$ at $t$ in $w^*$ differs from the value of $b$ at $t$ in $w$

(ii) *Constancy*: every variable that isn't *on the path* between $b$ and $e$ is held constant at $t$ between $w$ and $w^*$.

(iii) *Success*: the variable $e$ has a different value in $w^*$ at $t$ than $w$ at $t$ because[105] of (i). [106]

[105] Again, the "because" relation is primitive in the sense that it is not reducible to a kind of counterfactual dependence or a kind of entailment. It is a synchronic

A variable *x* is *on the path* between *b* and *e* if the manipulation of *e* through *b* consists of a manipulation of *x* through *b* and a synchronic manipulation of *e* through *x*.[107] For example, neural states in the brain are "along the path" between lower-level fundamental physical states and higher-level phenomenal consciousness. If there is a manipulation through *b* for *e*, then I will say that *e manipulatively*

---

relation, analogous to but not identical to efficient causation, the relation emergentists call *e*'s "arising from" *b*. It is a dependence relation in the sense that it *e*'s dependence on *b explains why e* exists.

[106] For Woodward (2005), these are the conditions on causal manipulations of *c* for *e*:

(i) *Intervention*: the value of *c* differs in *w** from *w*

(ii) *Constancy*: every variable that isn't *on the path* between *c* and *e* is constant between *w* and *w**.

(iii) *Success*: the variable *c* has a different value in *w** than *w* because of (i).

[107] I would suggest we take "arises directly from" to be a primitive relation, and then "arises from" to be the ancestral of that relation. So *x* is "on the path" between *e* and *b* in a synchronic sense means the following: *e* arises directly from something which arises . . . which arises directly from *x*, and *x* arises directly from something which arises . . . which arises directly from *b*.

*depends* upon *b*. The emergentist holds that the existence of phenomenal states[108] manipulatively depends upon the existence of some set of basal conditions.[109]

For example, suppose *infatuation* is an emergent property of Gary's brain. There is a possible world in which Gary has mild brain damage, and he is much less infatuated there because of it. That world is a successful manipulation of our world for Gary's infatuation through Gary's brain damage.[110]

There is also a possible world in which Gary has mild brain damage, but he is an equally infatuated person, because he has been exposed to more romantic comedies. That world is an unsuccessful manipulation through Gary's brain damage and Gary's cinematic experiences for Gary's infatuation. However, that world does not count as a manipulation *through Gary's brain damage* for Gary's infatuation, because it fails the constancy condition: Gary's cinematic experiences should have

---

[108] One might instead refer to the *instantiation* of phenomenal properties.

[109] Though a token conscious state need not manipulatively depend on its token physical basis, but only on *there being* such a state to serve as the needed basal condition. Note also that manipulative dependence here is a claim about the existential-dependence of one thing on another, not its identity-dependence. Were my thought of aluminum to arise from a bundle of silicon chips rather than a bundle of neurons, the nature of my thought would be the same, but the nature of my thought might change if "aluminum" were something else in another world.

[110] This account is adapted and simplified from Woodward (2005).

been kept *constant*, since they are not on the path between Gary's brain damage and his infatuation.[111]

The successful manipulation of Gary's infatuation through Gary's brain damage will also probably lower his heart rate. But this *doesn't* violate the constancy condition, Gary's heart rate counts as being on the path[112] between the two.

There is a possible world in which Gary is more infatuated and, coincidentally, Al Gore won the U. S. Presidential Election in 2000. It does not count as a successful manipulation of this world for Gary's infatuation through Al Gore's election, provided that the infatuation difference is not *because of* the electoral difference.

Suppose Gary could just so happen to be more infatuated than he is, but everything else in the world would remain the same. Any such world would *not* count as a manipulation of our world for his infatuation, because it would fail the intervention condition.

To return to our metaphor, we might say that there is a "train" between the actual world and the possible world in which Gary's brain state changes his infatuation level. Likewise, there is a "train" between the actual world, in which I am experiencing a slight pain in my arm, and the possible world in which I am experiencing excruciating pain because different fibers in my brain are firing. There

---

[111] Though it may be that Gary's brain damage is on the path between his cinematic experiences and his infatuation, of course.

[112] See Woodward (2005), 38-61

is a synchronic manipulation of this world for pain through the fibers in my brain, on which my experience of pain depends.

### 4.3 Manipulative Exclusivity

Recall the notion of *total* existential dependence described earlier, as a way of separating the emergentist from the traditional dualist. The traditional dualist accepts that there is *something else* besides my brain states through which my phenomenal states can be manipulated. The emergentist does not accept this.

To return to our story, suppose there are some poets in Actual who claim that there is a train between Actual and Voodooburg. In Voodooburg, my brain states are the same, but I am in a great deal of pain, because in that world a miniature doll in my likeness is being stuck with pins while a shaman thinks nasty thoughts. Other poets claim there is a night train to the spooky village of Zombiegrad. In Zombiegrad, a psycho-cosmic vortex in Sedona, Arizona is closed, and so consciousness never emerged from brain states in the course of natural history.

The emergentists proclaim: these poets are liars! My phenomenal consciousness depends upon my brain states *exclusively*. It is in some sense *determined* by my brain states. My brain states *suffice* for emergence – nothing else is needed. One can not manipulate phenomenal consciousness *in other ways* while keeping brain states fixed, nor can one *prevent* or *interfere with* phenomenal consciousness emerging once the basal conditions are in place.[113]

Take all of the manipulations of the actual world $w_a$ for instantiations of an emergent property *e*. The emergentist could make the following exclusive claim

[113] I am interpreting here the views of Alexander (1920), 3-30

about the dependence of emergent properties on their basal conditions: every successful manipulation of $w_a$ for $e$ is through $e$'s basal conditions $b$.

This rules out Voodooburg: there is no manipulation of $w_a$ for $e$ that isn't through $b$. It also rules out Zombiegrad, where something interferes with consciousness emerging as it does from its basal conditions. If a purported psycho-cosmic vortex were closed, but everything else (brain states included ) were kept fixed, then this attempted manipulation of $e$ that isn't through $b$ would not succeed, given the exclusivity principle.

We can generalize this principle to describe a kind of necessity:

> (*Manipulative Necessity*) *x manipulatively necessitates y* is true in world $w$ iff if there is no successful manipulation $w^*$ of $w$ for $y$ through any $z$, where $z$ is not identical to $x$ and not on the path between $x$ and $y$. [114]

---

[114] The □ operator is then defined as follows: $V(\Box p, w) = T$ iff $(\forall w^*)(Rww^* \rightarrow (V(p, w^*) = T))$. On this definition, accessibility relation $Rww^*$, "$w^*$ is a manipulation of $w$", is defined relative to variables $e$ and $b$ such that: $w^*$ is accessible from $w$ iff $w^* = w \vee [V(b, w) \neq V(b, w^*) \ \& \ D(e, b) \ \& \ (\forall v)(V(v, w) = V(v, w^*) \vee P(b, v, e))]$. The dependence function $D(x,y)$ is defined as: $D(x,y)$ iff $\sim y \Box\!\rightarrow \sim x$ where a primitive relation of dependence ($x$ "arises from" $y$) gives the semantics for the counterfactual □→ operator. The Path relation $P(x,y,z)$ is defined as: P(x,y,z) iff { D(z,x) & [D(y,x) & D(z,y) & D(D(z,x),(D(x,y)&D(y,z)))]} – in other words, $y$ is on the path between $x$ and $z$ when $z$ depends on $x$ and $y$ depends on $x$ and $z$ depends on $y$ and *that it is the case* that $z$ depends on $x$ depends on its being the case that $y$

Notice that manipulative necessitation is relativized to a world.[115] In a given world *w*, it is true in *w* that *x* manipulatively necessitates *y* when there is no manipulation of *w* in which something other than *x* could be intervened upon to change *y* which is not on the path between *x* and *y*. So, it may be that $\Box(x \rightarrow y)$ is true in *w* but not in some other world *w**. Manipulative necessitation defines an accessibility relation between worlds in terms of manipulations of a world, and then defines the necessity operator in terms those worlds accessible from a given world.

The emergentist claims that basal conditions manipulatively necessitate emergent phenomena in the actual world. This is why the emergentists deny the existence of Voodooburg and Zombiegrad: the poets *define* them as places with trains to Actual, but there could never be a train from Actual to such a place.

### 4.4 Total Manipulative Dependence

On my account, emergentists hold that there is no metaphysically possible world in which consciousness arises from brain states, and yet can be manipulated by means of something other than brain states. Consciousness might possibly emerge from silicon states or complex social states rather than brain states, and in

---

depends on *x* and *z* depends on *y*. For the notion of an accessibility relation and its role in various modal logics, see Hughes & Cresswell (1996).

[115] In this regard, it is like nomological necessitation, where the $\Box$ operator is similarly defined by $V(\Box p, w) = T$ iff $(\forall w^*)(Rww^* \rightarrow (V(p, w^*) = T))$ and the accessibility relation $Rww^*$ is defined as $w^*$ is accessible from $w$ iff {laws of $w$}$\subseteq${laws of $w^*$}.

that case it could only be manipulated by whatsoever it emerged from. There are no trains from any town where consciousness emerges from some base or other to any other town where consciousness has been changed through a change in something other than that same base.

The manipulative necessitation of consciousness by a physical world like ours does not rule out the existence of Zombieland. However, we can now see clearly why there could never be a train to Zombieland. The problem is not that Zombieland is too far away, or has different laws *per se*. Rather, the problem is that if Zombieland exists, then is not a manipulation of Actual. It fails the intervention condition. It is not a possible world in which consciousness has been manipulated by means of an intervention upon *something else*, keeping everything *else* constant. Rather, it is a world in which consciousness has been intervened upon directly (by some Leibnizian *Deus ex Machina*) and *everything* else has been kept the same. A train to Zombietown would be a manipulation of Actual for consciousness *through nothing*.

Here is what I have argued for. The existence of Zombieland is consistent with the exclusive dependence of consciousness on brain states. It is also consistent with brain states "fixing" conscious states, and being "enough" or "sufficient" for consciousness in the sense in which nothing further is needed.[116] It is consistent with there being no possible intervention to prevent consciousness from arising once brain states are in place, and thus consistent with brain states making certain that consciousness emerges just as it does and making the emergence of consciousness

[116] But it is not consistent with the deductive sense of "sufficiency"; that is, brain states are not logical or metaphysically sufficient conditions for consciousness.

empirically predictable.[117] All of these things are consistent with the metaphysical and even nomological possibility of "zombies".

### 4.5 Manipulative Supervenience

Let us return to our attempt to define ontological emergence[118] in terms of the *supervenience* of emergent properties upon their basal conditions.[119] Roughly speaking, one set of properties *supervenes* upon another whenever there can be no difference in the supervening property without a difference in its subvening base.[120] As discussed, supervenience is typically defined in terms of some sort of

---

[117] Again, this is the epistemic sense of a "guarantee", not the logical sense in the premises of a valid argument guarantee their conclusion.

[118] Throughout this paper, I have had in mind *ontological* emergentism, a contemporary revival in the philosophy of mind of the views presented by the British Emergentists. There is also what is called *epistemic* emergence or "weak emergence", which is not what I mean to discuss here.

[119] Van Cleve (1990) adopts the view that strong nomological supervenience expresses both the "dependence" and "determination" aspects of the Emergentist claim. Kim (2012), 85-104 argues that Emergentists were historically committed to supervenience in some form. McLaughlin (2008) holds that Emergentism should be defined in terms of nomological supervenience.

[120] Kim (1993), 53-91 in "Concepts of Supervenience" and "'Strong' and 'Global' supervenience revisited" gradually moves from this intuitive but imprecise concept of supervenience, taken from R. M. Hare's *Language of Morals*, to a more precise concept of Supervenience, which he calls "Strong Supervenience."

*necessitation* of supervenient properties on their subvening base. Recall the following definition of strong supervenience:

> *Strong Supervenience*. *A* strongly supervenes on *B* if and only if necessarily, if any *x* has some property *F* in *A*, then there is at least one property *G* in *B* such that *x* has *G*, and *necessarily* everything that has *G* has *F*.[121]

We asked: in what sense of *necessarily* need everything that has G have F?

One option we discussed was *metaphysical* necessity, or strong metaphysical supervenience. Although there is no conceptual or explanatory connection between the basal conditions and emergent phenomena, on this interpretation there is a kind of fundamental metaphysical principle such that in no metaphysically possible world do basal conditions occur without the emergent phenomena.[122]

Another option was *nomic* necessity, or strong nomic supervenience.[123] Here, nature contains explanatorily primitive and metaphysically contingent "trans-ordinal laws" such that there is no world *with the same laws as the actual world* in which the basal conditions occur without the same emergent phenomena.

---

[121] This definition is taken from Horgan (1993).

[122] Horgan (2009) suggests this 'Moorean Minimal Emergentism' as a coherent position on emergence, though he does not himself endorse it.

[123] Van Cleve (1990) and McLaughlin (2008) adopt this definition when interpreting the emergentist claim.

(Of course, some emergentists opt to deny supervenience altogether.[124])

However, here I offer a new option for the emergentist: strong supervenience with *manipulative necessity*, or manipulative supervenience. For everything with a phenomenal property, the emergentist can say, there is necessarily some physical basal property which it has, such that it is manipulatively necessary that everything with that physical basal property *G* has that phenomenal property *F*. In every world in which *F* manipulatively depends upon *G*, there is no manipulation-world of that world in which *F* is manipulated through an intervention upon something other than *G*.

In one regard, this is a weaker claim than metaphysical supervenience, insofar as it does not rule out the metaphysical possibility of "zombie" worlds, since such worlds would not count as manipulation worlds for any world in which consciousness emerges.[125] Yet, in another regard, this is a stronger claim than some versions of metaphysical supervenience for four significant reasons:

---

[124] See Humphreys (1997) and O'Connor (2000), although each author does so for different reasons.

[125] To be a manipulation world, the world would have to be one in which some variable was changed. But zombie worlds and invert worlds are perfect duplicates. Someone might argue that a zombie world with one ammonia particle changed would count as a manipulation of the actual world. However, to be a successful manipulation for consciousness, consciousness would also have to manipulatively depend upon the ammonia particle both in the actual world and in the manipulation world. But consciousness does not, according to our best theories, depend on a

First, manipulative supervenience entails the existential dependence of emergent properties on their basal conditions. Jackson's minimal global supervenience does not entail any sort of existential dependence relation. Kim's strong local supervenience entails the "Generic Account" of existential dependence, but – as mentioned – this account is problematic. On the other hand, manipulative supervenience meshes with an intuitive account on which existential dependence is understood as manipulative dependence.

Second, manipulative supervenience is not subject to the "blockers" problem raised by Hawthorne (2002)[126], unlike Jackson's minimal global supervenience. Suppose that there are non-physical substances in our world which can be "red" or "green". When they are "green" in a world, or altogether absent, nothing blocks the emergence of the mental from the physical. When they are "red" in a world, however, the emergence of the mental from the physical is blocked. Fortunately for us, these non-physical substances are green in the actual world, and so emergence happens. It is true that every minimal physical duplicate of our world is a duplicate with respect to the mental properties, since a minimal physical duplicate won't have the non-physical substances at all, and so nothing will block emergence. However, it is strange to think that this is a picture of the world on which either physicalism or emergentism is true: it is clearly a substance dualist account of our world.

---

particular ammonia particle in the actual world. Likewise, notice that it does not rule out "invert" worlds, since an inverted world does not count as a manipulation of its non-inverted physical duplicate or vice-versa.

[126] See also Leuenberger (2008)

Fortunately, manipulative supervenience rules this out explicitly – there can be no world in which the mental properties in our world are manipulated (let alone eliminated or blocked) by means of an intervention upon a non-physical substance.

Third, manipulative supervenience is not subject to the "lone ammonium molecule" objection raised by Kim (1993). It is consistent with some forms of metaphysical supervenience that there is another possible world, *w**, which is an exact physical duplicate of our world *save* that one ammonium molecule has been added to the rings of Saturn, and as a result the phenomenal properties in that world are different. However, this seems seriously at odds with both physicalism and emergentism. Fortunately, it is inconsistent with manipulative supervenience. Why? Because manipulative supervenience rules out that there is a world in which emergent properties are manipulated by means of interventions upon something other than their basal conditions, and the ammonium molecule is presumably not part of any conscious property's basal conditions.

Finally, manipulative supervenience does not face the problem of necessary beings.[127] It follows from metaphysical supervenience that necessary beings (mathematical abstracta, God) supervene upon the physical world. This is a very strange result for the sort of work supervenience is supposed to play in capturing the notion of total existential dependence. However, necessary beings most certainly do not manipulatively supervene upon the physical world, insofar as there is no manipulation of necessary beings through interventions upon physical properties or entities.

[127] See Jackson (1998)

Manipulative supervenience is in most respects stronger than nomological supervenience, insofar as it rules out worlds in which the physical remains the same, and yet the mental differs *because of* something else – e.g., a change in the laws operating in that world. In this regard, it is able to capture a way in which emergentism is distinct from traditional dualism: the traditional dualist can at most accept nomological supervenience, whereas the emergentist can accept manipulative supervenience.

### 4.6 Does the Distinction Collapse?

I have argued that what I am calling "manipulative necessity" is a distinct and strictly weaker sense of necessity from metaphysical necessity. However, one might object that the distinction depends on admitting certain background conditions into manipulative necessity: considerations about contingent laws of nature, dependence relationships, or facts about what is or isn't manipulable in a given world. Given these background conditions, manipulative necessity may be seen as distinct from metaphysical necessity, but this isn't a fair comparison. But without any such background conditions, the objection goes, manipulative necessity "come what may" simply is metaphysical necessity.

It is certainly true that metaphysical necessity entails manipulative necessity. If *b* metaphysically necessitates *a*, then there is no possible world in which *b* occurs without *a*. It follows that, given a world *w* in which both *a* and *b* occur, there is no manipulation world *w** in which the value of *a* changes by means of a change in some third variable *c*, while the value of *b* remains the same.

However, it is false that manipulative necessity entails metaphysical necessity, even if we remove from consideration contingent facts about dependence

relations in the actual world. Consider the following three models of worlds and the propositions which are true at them. On the M-model, *b* manipulatively necessitates *a* but does not metaphysically necessitate *a*. On the N-model, *b* both manipulatively and metaphysically necessitates *a*. On the O-Model, *b* neither metaphysically nor manipulatively necessitates *a*. The model makes no reference to contingent dependence relationships.

***M-Model***

*w: a, b,* $p_1 \ldots p_n$

$w_1$: *¬a, ¬b,* $p_1 \ldots p_n$

$w_2$ *a, b,* $\neg p_1, p_2 \ldots p_n$

$w_3$ *¬a, ¬b,* $\neg p_1, p_2 \ldots p_n$

$w_4$ *a, ¬b,* $p_1 \ldots p_n$

$w_5$ *a, ¬b,* $p_1 \ldots p_n$

*and there is no* $w_m$ *in which ¬a & b & ¬p, for some p in* $p_1 \ldots p_n$ *such that ¬p is not tautological given ¬a*

***N-Model***

*w: a, b,* $p_1 \ldots p_n$

$w_1$: *¬a, ¬b,* $p_1 \ldots p_n$

$w_2$ *a, b,* $\neg p_1, p_2 \ldots p_n$

$w_3$ *¬a, ¬b,* $\neg p_1, p_2 \ldots p_n$

$w_4$ *a, ¬b,* $p_1 \ldots p_n$

*and there is no* $w_m$ *in which b & ¬a*

*O-Model*

$w$: $a$, $b$, $p_1 \ldots p_n$

$w_1$: $\neg a$, $\neg b$, $p_1 \ldots p_n$

$w_2$ $a$, $b$, $\neg p_1$, $p_2 \ldots p_n$

$w_3$ $\neg a$, $\neg b$, $\neg p_1$, $p_2 \ldots p_n$

$w_4$ $a$, $\neg b$, $p_1 \ldots p_n$

$w_5$ $a$, $\neg b$, $p_1 \ldots p_n$

$w_6$ $\neg a$, $b$, $\neg p_1$, $p_2 \ldots p_n$

Note that on these models, we are omitting to display all of the tautological[128] (e.g., truth-functional and quantificational) consequences of $a$, $b$, $\neg a$, and $\neg b$; for instance, clearly $\neg\neg a$ is true in the world at which $a$ is true. So, $p_1 \ldots p_n$ should be taken to represent all of the propositions which are not tautologically independent of $a$ and $b$, in the sense of not being truth-functional or quantificational consequences of $a$, $b$, $\neg a$, or $\neg b$.[129] Note also that this model does not require that the number of

[128] One might be tempted to say "logical consequences", but one worries this would be question begging.

[129] It might also be necessary to exclude from the model any propositions which are *analytic* truths given $a$ or $b$ or their negations, if these are to count as distinct propositions. This wouldn't beg any questions, since the point of the model is to show a way in which $a$ can fail to occur without $b$ failing to occur, without requiring that something else $p$ fail to occur beyond the obvious cases. Of course, on the M-model $p_1$

possible worlds $m$ and the number of propositions $n$ be finite, though it does require that every proposition $p_1 \ldots p_n$ defined at $w$ also be defined at every world.[130]

This model offers a counterexample to the claim that manipulative necessity entails metaphysical necessity. On the M-model, $a$ manipulatively necessitates $b$, insofar as there is no proposition $p$ which is true at $w$ and isn't a tautological consequence of $b$ or $a$. Contrast this with the O-Model, on which $p_1$ is false and $a$ is also false at $w_6$. At the same time, on the M-model, $a$ does not metaphysically necessitate $b$, insofar as there is a world $w_5$ at which $a$ occurs without $b$. Contrast this with the N-model, where there is no such world and metaphysical necessity holds. Thus, while we developed the concept of manipulative necessity relying on the idea of some special synchronic dependence relationship ("manipulative dependence"), the formal properties which make manipulative necessity modally distinct from metaphysical necessity remain even when we give no special consideration to "manipulative dependence".

---

is in fact logically independent of $a$, whereas on the N-model it is not. If someone objects that the distinction between the M-model and the N-model would then only hold on the assumption of the existence of synthetic necessities, our reply is that if one rejects synthetic necessity, then this is a reason which should count *for* manipulative necessity and *against* metaphysical necessity as a way of capturing close modal relationships between non-analytic pairs of propositions.

[130] If some proposition is undefined at $w$, we can omit it from the model, since it has no bearing on either manipulative or metaphysical necessitation of $a$ by $b$.

One response to the counterexample which I have offered here goes something along the following lines: there must be some *further reason* why $a$ failed to occur in $w_5$ even though it actually occurred in $w$. Since $b$ can't be the reason, something else must be – there must be a reason why we are in $w$ and not $w_5$. This response appeals to a particularly strong version of the Principle of Sufficient Reason:

> (PSR+) If actually $q$, then there is no possible world $w^*$ at which $\neg q$ unless some further proposition $p$ has a different value at $w^*$ than its actual value, and $p$ explains why actually $q$.

Even advocates of some form of the PSR are unlikely to accept PSR+.[131] For one thing, PSR+ is inconsistent with the existence of objectively chancy phenomena. More troublingly, PSR+ is also prone to problems of infinite regress: whatever $p$ in the actual world which explains why actually $q$ in turn needs some explanation which rules out $\neg p$, a proposition true in the actual world which will in turn need an explanation which rules out its negation, and so on and so forth. Perhaps one resolves the regress by holding that the actual world is necessarily the case, and no other world is possible – and if that is so, then the distinction between manipulative and metaphysical necessity becomes a moot point.

---

[131] In its more traditional form, the PSR only requires that there be some explanation for every actual fact, without requiring that there be an explanation which rules out every other possibility.

If, without appealing to some principle like PSR+, one cannot object to the coherence of the M-model or to its distinction from the N-model, then one must accept that manipulative necessity is a distinct and weaker modal relation than metaphysical necessity. In this case, it is coherent for the emergentist to hold that basal conditions are sufficient for emergent phenomena (that is, they manipulatively necessitate them) without being sufficient conditions of them (that is, metaphysically necessitating them).

## 4.7 Replying to the Incoherence Argument

Why might this be an interesting option for the emergentist? I have already suggested one motive, which is that defining emergence in terms of manipulative supervenience permits the emergentist to remain neutral on controversial questions like whether philosophical zombies or inverted worlds are possible. Rather than divide over these questions, it may be useful for emergentists to share a common, neutral, empirically justifiable, non-speculative notion of supervenience.

But there are other motives. For one thing, Jaegwon Kim's well-known causal exclusion arguments rely upon metaphysical or nomological supervenience. Emergentists who do not have a reply to these arguments risk being convicted of epiphenomenalism. However, I suspect that overdetermination arguments pose no threat for manipulative supervenience.[132]

[132] Consider that overdetermination is possible with metaphysical or nomological necessitation: A and B can both independently on their own necessitate C in the relevant sense. But manipulative necessitation rules out overdetermination. It

Perhaps less well known is the incoherence argument against emergentism Kim (2010) offers.[133] Suppose that conscious states supervene with metaphysical necessity on brain states. Let B be the brain state which serves as the basal condition for conscious state C, such that B metaphysically necessitates C.[134] The emergentist is traditionally committed to the claim that C is *not deducible* on the basis of B, *even by an ideal reasoner*.[135] But surely an ideal reasoner would reason according to nothing less powerful than a derivation system which is both sound and complete with respect to the set of necessary truths – all and only the necessary truths would be axioms of the system.[136] Yet that would mean B→C would be an axiom for the ideal reasoner, and C would be ideally deducible from B – making emergence with metaphysical supervenience incoherent. Further, according to Kim,

---

cannot be that both A manipulatively necessitates C and B manipulatively necessitates C, unless A is on the path between B and C or vice-versa.

[133] The argument appears in Chapter 4 of Kim (2010): " 'Supervenient and Yet Not Deducible': Is there a Coherent concept of Ontological Emergence?", 85-104

[134] Of course, for the ontological emergentist (as opposed to the non-reductive physicalist) this metaphysical supervenience must be held to be primitive and inexplicable in other terms, perhaps like many hold the axioms of set theory are metaphysically necessary and not subject to further analysis. See Horgan (2009)

[135] It may well be that C can't be deduced from B by any actual being or any potential mechanism an actual being might create. See C. D. Broad (1925) and his discussion of the "Mathematical Archangel".

[136] Kim (2010), 96-104

accepting nomological supervenience instead won't get us out of the problem.[137] If we accept that the supervenience base must be identical to the deduction base, including the relevant trans-ordinal laws, then nomological supervenience entails ideal deducibility. [138]

Manipulative supervenience is compatible with the metaphysical or even nomological possibility of B without C, and I do not see a clear way in which an ideal reasoner could construct a derivation of C from B using the elements provided in the definition of manipulative supervenience. In this regard, manipulative supervenience may offer the emergentist a way to resolve the tension between the claim that basal conditions are *sufficient* for emergence and the claim that emergents are not ideally deducible from them.

---

[137] ibid.

[138] I suspect that there is another way out of this argument that does not require adopting manipulative supervenience. Nonetheless, the argument does reveal a deep tension between supervenience and non-deducibility in the emergentist's view. My suggestion in defense of metaphysical supervenience would be to understand the emergentist as denying not *a priori* deducibility, but deducibility *by means of analysis alone*. Provided that B→C is a synthetic necessity, C may be necessitated by B but not deducible in the intended sense. Alternatively, a critic of Kim's argument might spot something suspicious about Kim's move from *a priori* deducibility to the discussion of the properties of derivation systems, although I am unable myself to pinpoint any error here.

Here are two claims which are often taken to be equivalent in the literature. The first is this: "the phenomenal facts are fixed by the physical facts"[139], or "no change is possible in the phenomenal facts without a change in the physical facts", or, often intended metaphorically, "once God had made the physical facts, there was nothing further he had to add to get the phenomenal facts."[140] The second is this: "Given the physical facts, it is not possible for the phenomenal facts to have been different", or "the physical properties metaphysically necessitate the phenomenal properties", or "any world which is a minimal physical duplicate of our world is a duplicate simpliciter."[141] I have offered an explanation for why the two claims are not equivalent to each other. The second involves an aspect of manipulative supervenience (at least, the condition of manipulative exclusivity), whereas the second is explicitly metaphysical supervenience. The second entails the first, but not vice-versa, and an interesting form of emergentism can be developed by affirming the first while denying the second.

## §5. Conclusions: Weighing the Options

This chapter has developed several coherent versions of emergence which can be distinguished both from physicalism and from traditional dualism. These

---

[139] Kallestrup (2006)

[140] Chalmers (1996), 41. Of course, Chalmers would say that God does have to add something further even on my account – namely, the emergent mental facts themselves. My point is only that it is not as though he adds something else (a mental substance) in order to get the emergent mental facts.

[141] Jackson (1998)

distinctions were made by asking four questions: (i) do the higher-level facts *ontologically* depend upon the physical – that is, are they grounded in the physical facts? (ii) with what sort of necessity do the higher-level properties supervene upon the physical? (iii) do the higher-level phenomena *existentially* depend upon the physical – that is, can they be manipulated synchronically by means of the physical? and (iv) do the higher-level phenomena existentially depend upon anything else? Consider the options presented in Table 1.

Table 1

*Varieties of Emergence*

| | Ontological Dependence on physical | Supervenience on physical | Existential Dependence on physical | Exist. Dep. on non-physical |
|---|---|---|---|---|
| Physicalism | Whole | Metaphysical | Whole | None |
| Weak Necessary Emergence | Partial | Metaphysical | Whole | None |
| Weak Contingent Emergence | Partial | Manipulative | Whole | None |
| Strong Necessary Emergence | None | Metaphysical | Whole | None |
| Strong Contingent Emergence | None | Manipulative / Nomic | Whole | None |
| Traditional Dualism | None | Nomic/None | Partial | None |
| Substance Dualism | None | Nomic/None | Partial | Partial |
| Cartesian Dualism | None | None | None | Whole |

*Weak necessary emergence* is the view that the relevant sort of emergent properties are totally existentially dependent upon the physical and necessitated by it, and that it is *part* of the essential nature of emergent properties that they are physical, but that there is something further and fundamental to their nature. In this regard, it seems equivalent to the "Dualistic Materialism" of Joseph Almog (2009), when he holds that it is *by their very nature* that conscious states depend upon physical states, yet they are distinct in nature. This view can make only limited use of the usual conceivability or knowledge arguments for dualism, and is perhaps the closest one can get to being a physicalist without actually qualifying as one. It has the advantage that nearly every argument for physicalism against dualism counts as an argument for it rather than against it.

*Strong necessary emergence* is the view that emergent phenomena are ontologically independent from their physical base even while being metaphysically necessitated by them. Unlike forms of weak emergence and contingent emergence, strong necessary emergence has the burden of explaining the apparent violation of Hume's dictum – how there can be metaphysically necessary relations between ontologically independent properties.

*Strong contingent emergence* is the view that emergent phenomena are totally dependent upon their physical base for their existence, without being necessitated by them. This total dependence can be expressed in terms of manipulative supervenience of emergents on their physical base, or it can be expressed in terms of nomic supervenience provided that the trans-ordinal laws count as "nothing further" and are not a means by which a manipulation can occur (i.e., if nomic supervenience is so defined as to entail manipulative supervenience).

The challenge to this view involves asking *why* we should be in a world in which emergence occurs as opposed to a world in which it doesn't, given that the two worlds are physically identical and nothing else can interfere with emergence. Someone who accepts manipulative supervenience must reject that there is an answer to such a question. However, this may be hard to accept.

*Weak contingent emergence* is the view that emergent phenomena of the relevant sort are partly but not fully grounded in the physical, while remaining totally existentially dependent upon it – and that this partial lack of grounding in the physical means that it is possible that zombies or inverts exist, although no such world is a manipulation of our own. This view shares the costs of strong contingent emergence but is somewhat less forcibly dualistic.

I believe that it is perfectly legitimate to refer to the view I've called minimal traditional dualism as a fifth brand of emergentism, when the minimal traditional dualist accepts nomic supervenience and partial existential dependence on the physical. Many emergentists do hold this sort of view about consciousness, interpreting the sufficiency claim as involving nomic supervenience and holding in virtue of emergent "trans-ordinal" laws. Because my purpose in this chapter has been to distinguish emergentism from traditional dualism, I have focused on ways in which emergence might hold with stronger forms of necessity than nomic necessity. By doing so, it isn't my intention to rule out emergence with nomic necessity, by definition.

It may of course be that different forms of ontological emergence apply to different phenomena in the world – in fact, I believe it is so. When attempting to place phenomenal experiences in the world, it turns out the emergentist has a

variety of coherent options. In Chapter 3, I will put aside many of these distinctions and presume a kind of metaphysically necessary emergence. However, in Chapter 4, I will return to these distinctions and attempt to relate them to the traditional question of whether emergent facts would be deducible *a priori* by an ideal reasoner or "mathematical archangel". In Chapter 5, I will focus specifically on the advantages of weak necessary emergence in describing a variety of phenomena outside of the philosophy of mind.

Chapter 3

# EMERGENCE AND CAUSATION

## §1. The Problem of Mental Causation

The sandwich is half-finished when I hear my phone ring. I am in the middle of spreading mustard on the bread – my knife stalls in mid-air. My brother is calling. I hate to delay eating my sandwich, but a long time has passed since I last talked to him. I decide to answer the phone. I push a button and say, "Hello."

The movement of my finger to push the button and the vibration of my vocal chords in such a way so as to make the sounds for "Hello" – these are physical events. They happen *because* I decide to answer the phone. I decide to answer the phone because I hear the phone ring and see my brother's name on the Caller-ID, and because my desire to talk to my brother outweighs my desire to eat a sandwich. These are conscious mental events. Conscious mental events cause physical events.

The alternative is *epiphenomenalism*: the claim that, contrary to appearances, our thoughts and reasons and decisions play no part in making anything happen, but float freely above all the physical phenomena and their causal patterns. Although physical events cause mental events, mental events do not cause physical ones. Like other views that entail widespread massive illusions, we should prefer to avoid epiphenomenalism if we can.

At the same time, we should prefer not to explain mental causation as an instance of *telekinesis*. Telekinesis would be said to occur if some physical event were caused by mental events alone, with no physical causes – or if a physical event had physical causes, but these causes were somehow *missing something* without the addition of a mental cause. The sciences generally eschew telekinesis, for good

reasons. So, widespread telekinesis seems like a rather implausible theory for how people go about answering their phones.

A very simple explanation which avoids both epiphenomenalism and telekinesis is the type-identity theory. On the type-identity theory, conscious mental properties (or types of mental events) are identical to certain physical properties (e.g., types of events in the brain). My belief that my brother is calling and my desire to speak to him are identical to two distinctive kinds of events in the brain – a belief-type event and a desire type event. Together these cause a third event in my brain – my decision – which goes on to cause more events in my brain which transmit the signal to my extremities to push the button and say "Hello." Conscious mental events[142] cause physical events because consciousness is by nature physical.

An emergentist holds that mental events are not identical in nature to physical events. Emergentists do not accept any form of identity theory. In contrast to historical emergentism, the form of emergentism I have been defending does deny that are novel fundamental causal *forces* which arise at the mental level, and accepts the causal closure of the physical as normally understood. By accepting causal closure, it leaves no room for telekinesis. Emergentism also accepts a kind of

[142] For the remainder of this chapter, I will use "mental *x*" as shorthand for "conscious mental *x*." I accept that there are unconscious mental events, properties, kinds, and states, but what is most of interest in my discussion of emergentism is phenomenal consciousness. Further, for the remainder of this chapter, I will use the language of mental and physical "events", with the understanding that it could be freely translated into claims about mental and physical properties, states, etc.

strong supervenience of the mental on the physical – necessarily, for every mental event, there is some physical event which necessitates the existence of that mental event. So, to avoid epiphenomenalism, the emergentist must give an explanation for how mental events can cause physical events while remaining distinct from them – the problem of "downward causation".[143]

It can be argued that the non-reductive physicalist is in the same boat as the emergentist on this issue. Unlike the emergentist, a non-reductive physicalist accepts that mental events are by nature physical events, and may accept that particular mental event *tokens* are identical to particular physical event *tokens*. However, like the emergentist, the non-reductive physicalist denies the type-identity theory, and holds to a kind of strong supervenience instead. I suspect the non-reductive physicalist will also need to give an account of how downward causation is possible.

In this chapter, I will attempt to tackle a well-known and compelling argument that against both emergentism and non-reductive physicalism. The causal exclusion arguments of Jaegwon Kim have for over two decades thrown a wrench into emergentist and non-reductive physicalist accounts of mental causation, by arguing that accepting supervenience and causal closure, while denying type-identity, leads either to epiphenomenalism or to an implausible form of widespread overdetermination.[144]

---

[143] Notably, this problem applies to *any* emergent property with causal powers, not just to conscious mental states.

[144] See Kim (1993), (2005), (2010), (2011), (2011b), and elsewhere.

I believe the response I offer in this chapter has the advantage of offering a very simple explanation of the error in Kim's argument, in a way which doesn't require serious revisions to our ordinary pre-theoretical understanding of causation or causal closure. On my account, there is an ambiguity between the ordinary use of the word "sufficient", meaning that something is complete and nothing more is needed, and a technical sense of "sufficient" used in formal logic, meaning that a deduction of a particular conclusion from something would be valid. The ordinary sense of "sufficient" is the one which appears in discussions of downward causation and intuitive prohibitions on overdetermination.[145] However, Kim reinterprets both as instances of the logician's sense of "sufficient" – a tricky move which philosophers, so immersed in the specialized language of logic, are perhaps the least likely of all people to notice.

Many rebuttals to Kim's arguments have been offered before mine, of course. By adding my response to other recent responses to Kim's argument, I hope to show that there are now many plausible routes by which an emergentist or non-reductive physicalist can give an account of mental-to-physical "downward causation" which meet Kim's objection, and that concerns about overdetermination should no longer stand in the way of either of these theories.

I will begin in section 2 by providing some important background on the topic of causation. Following this, in section 3, I offer a summary of the major stages of Kim's argument. In section 4, I will describe some recent responses to Kim. Then, in

[145] As well as, arguably, definitions of causal closure.

Sections 5-7, I offer my argument for the distinction between the ordinary and technical senses of "sufficiency", and develop this into a reply to Kim.

## §2 Background: Causation

### 2.1 Force

At the risk of over-simplifying a complex issue, it may be helpful to clarify a few important ideas in the broader discussion of emergence and causation.

First, the notion of *force* is used in a technical way in these discussions, as opposed to the more ordinary notion of an exertion of strength, power, or control. Here, "force" is meant to refer deferentially to whatever kind of thing it is which the physicists are referring to when they talk about "fundamental forces". At the moment, the standard model in physics has four fundamental forces: electro-magnetism, gravity, the strong force, and the weak force. These are fundamental insofar as their force cannot be explained in terms of any other force. The term *causal force*, as I use it here, means anything which is either a fundamental force, or else a higher-level force whose force is entirely derived from fundamental forces. When a S.W.A.T. team uses "brute force" to open a door, they need not use an electro-magnet. Their action involves causal force insofar as it "borrows" all of its force from various fundamental forces occurring in the same spacio-temporal region.

Historically, emergentists like J. S. Mill believed that there were emergent fundamental forces: that at higher levels of organization in chemistry, biology, and psychology, new forces arose, both comparable to the forces of physics and over and

above them.[146] While this was a legitimate position to hold given the state of knowledge at the time, this is no longer a credible view: we now have an explanation of how high-level causal forces in chemistry and biology entirely derive their force from the fundamental forces of physics. If anything, the push of current physics is towards consolidating or unifying the fundamental forces further, rather than expanding them. So, all sides in this discussion – reductivist, non-reductivist, and emergentist – should be in agreement as to what the causal forces in the world are and are not.

## 2.2 Accounts of Causation: Deductive, Counterfactual

Two general classes of accounts of what causation is are especially relevant to understanding Kim's argument. Obviously, causation is a rich topic in philosophy with a long history. Many accounts of causation do not fit either of these classes I will give.[147] However, understanding these highly influential types of account is helpful for understanding Kim's argument.

**2.2.1 Deductive Accounts.** Deductive (or "nomological") accounts model causation along the lines of a formal deduction. Such an approach can be found in

---

[146] One might speak of "emergent causal forces" in the sense that a causal force might be exercised by an emergent phenomenon, so long as the causal force is entirely derived from the fundamental physical forces. However, this locution is too easy to confuse with historical claims about emergent fundamental forces, so I will omit it.

[147] For one such example, see the account of Salmon, (1971).

Aristotle's *Posterior Analytics*.[148] In the deduction, the effect serves as the conclusion of an argument in which the cause, a series of background conditions[149], and some general law serve as premises. For example, suppose I knock over a table on which a cup of coffee is sitting:

A1. (Background Conditions). The cup of coffee is on the table, and nothing else is between the cup and the floor.

A2. (General Law) Ceteris paribus, if there is nothing between an object and the floor, the object will fall to the floor.

C. The table is removed from underneath the cup of coffee. (Cause)

E. The cup of coffee falls to the floor. (Effect)

Laws play a prominent role in deductive accounts of causation. These laws should not be understood too strictly – they might hold only given *ceteris paribus* conditions, or they might be probabilistic. Laws can apply at all levels – from folk psychology or sociology down to fundamental physics – and it is debatable whether the high-level laws can be reduced to the low-level laws.[150] One might be a robust

---

[148] cf. Book I, Ch. 6-22; Aristotle held that that causes were natural laws which served as middle premises in valid syllogisms.

[149] Background conditions tend to be implicit in ordinary explanations. In most situations, an explicit list of all the relevant background conditions could never be given.

[150] The classic reductionistic approach is found in E. Nagel (1961)

realist about these laws, maintaining that they are part of the natural fabric of the actual world which is being uncovered by scientific investigation. One might instead give a deflationary account of these laws, understanding them merely to be true generalizations of observed regularities without observed exceptions. [151]

**2.2.2 Counterfactual Accounts.** Counterfactual accounts model causation as involving some element of *difference-making*:[152] a cause makes some difference in what happens in the event it affects. This may take the form of a strict counterfactual account, like that of David Lewis (1973): *C* causes *E* iff in the nearest world where *C* is not the case, *E* would not have occurred. Which world qualifies as *nearest* may be determined by maintaining as many of the background conditions and general laws in the actual world as possible (cf. *A1* and *A2* from the deductive model), or it may be determined by some contextual or normative ranking of similarity relations between worlds.

However, a counterfactual account may be more complex, as on the interventionist account of Woodward (2005). For the interventionist, *C* causes *E* when C provides a means of manipulating E – that is to say, when changes in the

---

[151] Hempel, C. and P. Oppenheim., (1948)

[152] It's misleading to say that counterfactual accounts are those which make claims about other possible worlds. Deductive accounts make claims about other possible worlds too – namely, which ones are ruled out. But deductive accounts do not require that causes make a difference: nothing rules out overdetermination. Further, one might develop a counterfactual account which makes no reference to possible worlds.

values of the variables in *C* will lead to changes in the values of the variables of *E*. Woodward's account has some advantages over Lewis's account, including moving away from a binary system to a much richer set of values, avoiding causal pre-emption problems[153], resolving more clearly the issue of which alternative scenarios are relevant, and providing a clear non-deflationary account of how counterfactual causes can be generative and productive.[154] In this chapter I will focus on Lewis's account because it is likely more familiar to my audience, I should note that I prefer Woodward's account, since many objections to counterfactual accounts of downward causation can be answered by the manipulationist but not the Lewisian.

Finally, accounts of causation on which causes involve some kind of "pushing" or "pulling" – or a transmission of information or energy – also fall into the category of accounts where causes must make some difference or other. Even though these accounts are not typically expressed in counterfactual terms, they are in agreement that effects would not happen without the "push" of their causes, and that causation need not be modeled as involving any sort of deduction.

---

[153] Problems which, notably, lead Lewis (2000) to modify his prior view.

[154] Again, debates between deflationary vs. non-deflationary accounts of causation are not essential to any of these views. Lewis (1973) gives a counterfactual account of causation which is deflationary, and Woodward (2005) gives an interventionist counterfactual account which is not deflationary; however, there are non-deflationary strict counterfactual accounts of causation, and perhaps someone will develop a deflationary interventionist counterfactual account someday.

## 2.3 Related Causal Concepts.

**2.3.1 Cause and "Cause".** There are a number of other causal terms which it will be helpful to distinguish. The term "cause" itself can be used both to indicate the *real* causes (according to some account of causation), and to indicate a sub-class[155] of the real causes – the causes it is permissible in ordinary discourse to use the word "cause" for. For instance, on some counterfactual accounts of causation, earth's narrowly escaping a large asteroid in the 12th century counts as a cause of why Napoleon lost the battle of Waterloo – for, had an asteroid struck earth in the 12th century, life would have been wiped out, and there would have been no Napoleon or battle of Waterloo. However, it is not permissible for historians to discuss how "missing the asteroid caused Napoleon's defeat", because this falsely implicates that missing the Asteroid was especially significant or relevant to Napoleon's defeat. Similarly, on some regularity accounts, a man's taking birth control pills qualifies as the cause of his not getting pregnant (since men who take birth control pills regularly fail to get pregnant). However, this doesn't make it permissible to advertise birth control pills to naïve young men.

**2.3.2 Causal Responsibility**. Whether or not it is permissible to call one thing a cause of some effect depends upon the degree of causal *responsibility* the

---

[155] Perhaps on some deflationary accounts of causation, the real causes are actually smaller than the set of things which it is permissible to call "causes", and the ordinary sense of "cause" is more an extension or metaphor of the real sense. Because it is difficult to reconcile this view with any sort of non-reductive approach to mental causation, I'm omitting it here.

cause has for the effect. Degrees of causal responsibility are likely determined by a mixture of probabilistic, normative, and contextual factors, although may be no precise formula for it. One cause may be more responsible than another cause for an effect: for instance, the rainstorm may be more responsible than the broken water main for the flood, though both contributed to it.

It is conventional in ordinary English to give some causes the honorific title *the* cause, e.g.: "Archduke Ferdinand's assassination was the cause of World War I." Strictly speaking, there is rarely such a thing as *the* cause of an event (surely, the First World War had too many causes to name). What someone means by "*c* is the cause of *e*" is that the *c* has a greater degree of causal responsibility than any other cause.

**2.3.3 Causal Power and Generative Causes**. A causal *power* is a property that some phenomenon (object, property, etc.) has, by which it is able to generate, produce, influence, or prevent changes in other phenomena. Boiling water has the power to dissolve sugar. A causal power may remain merely potential; when exercised, it produces an instance of causation.

However, not every instance of causation needs count as an exercise of causal powers. For instance, one might accept that there is such a thing as causation by absence, yet deny that absences have causal powers, insofar as they do not produce or generate anything. For emergentists, it is important that mental-to-mental and mental-to-physical causation both be considered generative or productive causes: that is, that they be real exercises of causal powers. Non-reductive physicalists need not assign the same importance to mental properties having distinct causal powers:

it may be enough that they represent real causal dependencies and figure in true causal explanations.

**2.3.4 Dependence, Overdetermination, Explanation.** A causal *dependence* relation holds between an effect and some temporally prior[156] cause without which the effect would not have occurred as it did. Causal dependence is often expressed as counterfactual dependence. The movement of a car causally depends on the gasoline in the tank: were there no gas in the tank, the car wouldn't move. Not all counterfactual dependencies need be considered causal dependencies.[157]

Not every instance of causation is a case of causal dependency. In a case of *overdetermination*, for example, there is causation without causal dependency. Suppose a man is sentenced to be executed by a firing squad of three soldiers, each of whom is an expert marksman aiming at the heart, such that any one shot is certain to kill the man. Each shot is a cause of the man's death, but for no shot is it true that, were it not fired, he would not have died. This case of causal *pre-emption*

---

[156] Causal dependence thus should be distinguished from the synchronic, "upward" dependence of emergent phenomena upon their subvening bases, without which they would not exist as they did. For clarity, I believe it is better to regard this "metaphysical dependence" or "existential dependence" relation as non-causal, although Kim and other authors from time to time discuss it as though it were a kind of causation.

[157] For instance, back-tracking counterfactuals are generally regarded as non-causal. See Lewis (1973)

raises issues for counterfactual accounts of causation.[158] How precisely to define overdetermination is debatable, of course. It isn't simply a matter of having two causes: there must be two *sufficient* causes to be an overdetermination case.

A causal *explanation* is an explanation of why some phenomenon happened in terms of what caused it at a prior time.[159] On deductive accounts of causation, a causal explanation has the form of derivation of effects from a set of premises, including the cause, the relevant background conditions, and the general law. On counterfactual accounts of causation, causal explanations correspond to causal dependence relations. For instance, the gas explains the movement of the car insofar as the movement of the car causally depends upon it – but *something more is needed* to explain the movement of the car, because the movement of the car causally

---

[158] Lewis (2000) revises his previous account to address the issue of pre-emption; Woodward (2005) sees an advantage of his account in that it escapes pre-emption problems.

[159] Diachronic causal explanations are probably not the only types of explanations: there may also be a synchronic metaphysical explanations of higher-level phenomena in terms of lower-level phenomena. In the interest of full disclosure, I also believe we ought to accept synchronic *ontological* explanations of a thing in terms of its nature (see Chapter 4), and diachronic *teleological* explanations of present states in terms of some non-arbitrary possible future state (see Chapter 6), given a bit of demystification of each, thus completing the Aristotelian quartet. However, nothing I say in this chapter depends upon accepting either of these latter two forms of explanation.

depends upon a whole host of other phenomena, like the running of the engine and the inflation of the tires. In either case, one thing might cause another without providing a useful causal explanation for it.

Causal explanations occur at distinctive levels. On one level, asphyxiation is the cause of a suffocated person's death; on another level, the deaths of particular cells are; on another level, the cessation of certain chemical processes involving $O_2$ movement are; on another level, getting involved with the mafia might be the cause of death. It is open for debate whether these "levels" are mere conveniences, or part of the metaphysical structure of the world. If levels are part of the metaphysical structure of the world, then levels of causal explanation will correspond to distinctive laws operating at that level, and/or to distinctive causal dependence relations at that level. The emergentist, naturally, takes all talk of "levels" very seriously, and the reductivist does not.

With these distinctions in mind, I will turn now to Kim's arguments against emergentism and non-reductive physicalism based on the objection from widespread overdetermination.

## §3 Kim's Argument

### 3.1 Overview

Jaegwon Kim's well-known causal exclusion arguments have many variations, all of which argue that middle positions in the philosophy of mind (e.g., non-reductive physicalism and emergentist dualism) are committed to there being physical events which have *both* a sufficient mental cause *and* a sufficient physical cause. According to Kim, this would amount to an implausible kind of widespread overdetermination, leaving as the only alternatives (i) epiphenomenalism, or (ii) a

reductivist physical-mental type-identity theory.[160] Two principles play an important role in Kim's arguments:

> *(CCPD) Causal Closure of the Physical Domain*. If a physical event has a cause (occurring) at time *t*, it has a sufficient physical cause at *t.*
>
> *(EP) Exclusion Principle*. No event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination.[161]

The definitions of both principles invoke the notion of a *sufficient cause*. As mentioned, I believe the phrase "sufficient cause" admits of two interpretations: one on which a cause is "sufficient" in the technical sense of being a logically "sufficient condition" of its effect, and one on which a cause is "sufficient" in the ordinary language sense of being "enough", such that nothing else is needed to bring the effect about. However, Kim interprets "sufficient cause" in his principles CCPD and EP as involving sufficiency in the technical sense, and he holds the same for mental causation of physical events.

Kim's objections are typically divided into two sub-arguments. First, his *supervenience argument* makes the case that the non-reductivist who accepts both supervenience and the distinctness of mental and physical properties is committed

---

[160] Kim (1993), 350-357.

[161] Kim (2011), 215-217

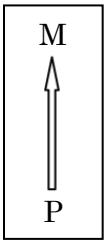
to downward "mental-to-physical" causation. Second, his *exclusion argument* makes the case that mental-to-physical downward causation presents an implausible case of widespread overdetermination. Obviously, it is the exclusion argument which troubles the emergentist – emergentists already accept the conclusion of the supervenience argument. However, understanding the supervenience argument is helpful to seeing exactly what form of downward causation is entailed by non-reductivist accounts of the mind.

### 3.2 Kim's Supervenience Argument

At a minimum, both non-reductive physicalists and emergentists agree that mental properties supervene upon physical properties but are not identical to physical properties. Let M be some mental state of mine – say, my belief that I am a human and all humans are mortal. It follows that there is some underlying physical state (i.e., brain state) P, such that P necessitates M. (See Figure 1).

**Figure 1**

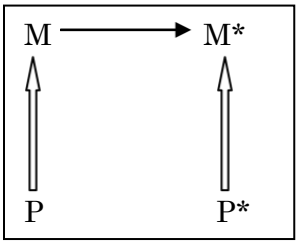
M Supervenes Upon P



Kim cites the following principle, which he calls "Alexander's Dictum": *to be real is to have causal powers*.[162] He concludes from this that for mental properties to be real, distinct, and not reducible to physical properties, mental properties must have their own real, distinct, and non-reducible causal powers. *M* must have some causal power which is not a causal power of *P*. (See Figure 2) This could be the power to bring about some physical effect, but Kim doesn't want to assume downward causation off the bat. Instead, we can assume that mental properties simply have the power to bring about other mental events on their own level – which sounds innocent enough. Let *M** be my mental act of concluding that I am mortal, an event which is caused by my belief *M* that I am human and all humans are mortal. There will be some physical event *P** such that *M** supervenes on *P**.

**Figure 2**

M* is overdetermined by M and P*




All of this sounds uncontroversial enough – non-reductive physicalists and emergentists both hold that there are causal patterns and relations at the mental

[162] Kim (1993), 348; citing the emergentist Samuel Alexander, *Space, Time, and Deity*, vol. 2, p. 8

level of some sort. However, the supervenience argument soon becomes more serious.

Kim argues that M must cause M* *by causing* M*'s supervenience base, P*. Since M must cause P*, anyone who accepts both supervenience and the distinctness of M and M* is committed to downward causation. How does Kim argue for this?

Why is the instance of *M** present? *Ex hypothesi*, it is there because an instance of *M* caused it; that's why it's there. But there is another answer: it's there because *P** physically realizes *M** and *P** is instantiated on this occasion.[163]

Kim sees *M* and *P** as competing claims for why *M** has occurred. Why did I conclude that I am mortal? Was it because I believed that I was human and all humans are mortal? Or was it because I was in a certain brain state? It's not that *M* and *P** caused my conclusion jointly, each contributing a piece of the puzzle: *P** necessitated *M** on its own. To simplify the argument from here, it is helpful to assume one must adopt either a deductive or a counterfactual account of mental causation.[164]

***Option A.*** On a deductive account of causation, an argument with *M* and various background conditions and laws as a premise must offer a valid deduction of *M**. It follows that "the given instance of *M* was a *sufficient condition* for that

---

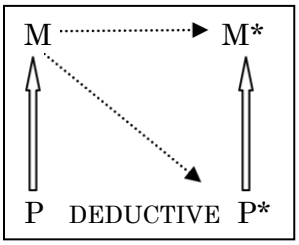
[163] Kim (1993), 352. Emphasis original.

[164] I am simplifying the argument for ease of exposition. Kim is well aware that there are accounts of mental causation which fall into neither category, of course, and has argued that his objection applies to them equally.

instance of *M**'.[165] My beliefs that I am human and that all humans are mortal must, given certain laws of thought, necessitate that I conclude that I am mortal. However, since my brain state *P** also necessitates my conclusion *M**, *P** is also a sufficient condition of *M**. (See Figure 3).

**Figure 3**

M causes P* (Deductive Model)




Kim understands the fact that both *P** and *M* are sufficient conditions of *M** to entail that both are sufficient causes of *M**. According to (EP), no event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination. But Kim does not believe this is a genuine case of overdetermination:

> Nor is it plausible to suppose that the occurrence of *M** on this occasion was somehow *overdetermined* in that it has two distinct and independent origins in *M* and *P**. For this, too, conflicts with the assumption that *M** is a property that requires a physical realization base in order to be instantiated,

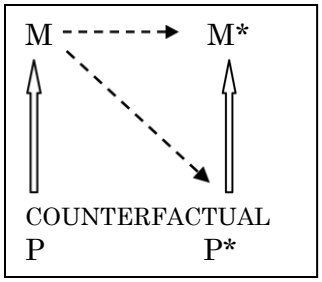
[165] Kim (1993) 352, Emphasis added.

> and that this instance of *M** is there because it is realized by *P**. . . I believe the only coherent story we can tell here is to suppose that the *M*-instance caused *M** to be instantiated *by causing P*, M*'s physical realization base, to be instantiated*. [166]

If one accepts Kim's claim that this cannot be a case of overdetermination, then it follows that M's causing M* and P*'s causing M* cannot be distinct causes of M*, but that instead M must cause M* by causing P*.

**Figure 4**

M causes P* (Counterfactual Model)




***Option B.*** On a counterfactual account (See Figure 4), for *M* to cause *M**, the following counterfactual must be true: were *M* not to have occurred, then *M** would not have occurred. Remember, *P** necessitates *M**. Were *M* not the case but *P** still the case, then M* still would have occurred because of *P**. In other words, were I still in the brain state corresponding to my conclusion, I still would have concluded that I was mortal, even if I had not believed beforehand that I was human and all

[166] Kim (1993), 352. Emphasis original.

humans were mortal. So, the only way for the counterfactual to be true is for *P** itself to counterfactually depend upon *M*: were it not for my believing that I am human and all humans are mortal, I never would have gotten into the brain state corresponding to a conclusion to begin with. We have the downward causation of *P** by *M*.[167]

However, one should note that Kim can only compel an opponent at this stage to accept the *counterfactual* causation of P* by M. He can't compel his opponent to accept that M is a cause of P* on a deductive account of causation by the supervenience argument alone, so long as *Option B* remains open. This will be important to keep in mind as we move into the exclusion argument.

### 3.3 Kim's Exclusion Argument

Kim now cites his principle of causal closure, (CCPD), to argue that *P** must itself have some sufficient *physical* cause – call this *P+*. This move should be uncontroversial enough – the emergentist accepts that *P** has a sufficient physical cause. Kim makes the further choice to identify *P+* with *P*, although he does not further justify this move. It follows that *P** now has two causes: *P* and *M*. Kim wants to know: which of them does the work? Does my believing that I am human and all humans are mortal cause my brain state? Or is it my *brain state* corresponding to these beliefs which really causes the brain state corresponding to my conclusion?

At this stage, Kim assumes that for *M* to be a serious cause of *P**, *M* must be a sufficient cause of *P**. (As noted earlier, this is not guaranteed by the

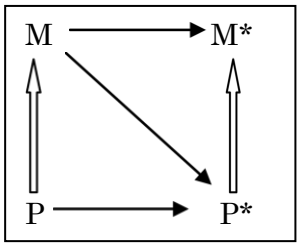
---

[167] Thus, "same level causation can occur only if cross level causation can occur". Kim (1993), 353.

Supervenience argument: *P** need only counterfactually depend upon *M*). Granting that *M* must be a sufficient cause of *P**, and that *P* is also a sufficient cause of *P**, it follows under Kim's definition that *P** has two sufficient causes. According to (EP), if *M* and *P* are distinct causes, then *P** must be overdetermined by *M* and by *P*. This step is crucial to Kim's argument – and it is why his arguments are often identified by their involving "overdetermination". (See Figure 5)

**Figure 5**

P* is Overdetermined by M and P




Kim's trap is now set for his opponent. In the final phase of his argument, he argues that overdetermination of P* by both P and M is implausible, since overdetermination only happens either by an accident or as a result of some sort of intentional planning, but this case is not either of those cases. According to Kim, then we should reject that *M* has any causal power over and above the causal power of *P* in causing *P**. It is really *P* which produces *P**, not *M*. Kim's opponent is left with two options – either to hold to a kind of epiphenomenalism on which *M* is distinct from *P* and yet has no causal powers over and above *P*, or else to follow Alexander's dictum and conclude that *M* is identical to *P*, acceding to the type-identity theory.

## §4 Existing Responses to Kim

### 4.1 Redefining Overdetermination

While Kim's objection is remarkable for its difficulty and staying power within the philosophical community, a number of insightful and persuasive rebuttals to Kim's objection have been offered recently. The first, by Karen Bennett (2003) proposes a more plausible definition of overdetermination, which ultimately gets the emergentist off of the hook.

Kim defines overdetermination as any case in which an event has two distinct sufficient causes. Bennett accepts that this may be a necessary condition of being an overdetermination case, but she rejects it as a definition. She adds further necessary conditions to what it is to be a cause of overdetermination.

(Bennett's Definition) $c_1$ and $c_2$ causally overdetermine e only if:

(SF) $c_1$ is a sufficient cause of e, and $c_2$ is a sufficient cause of e.

(O1) were $c_1$ to occur and not $c_2$ , then e still would occur.

(O2) were $c_2$ to occur and not $c_1$ , then e still would occur.

(NT) (O1) and (O2) are not trivially true.[168]

Notably, these additional conditions tie the definition of overdetermination into a counterfactual account of causation. To clarify, "trivially true" in this case would mean that there is no possible world in which the antecedent conditions obtain. For example, if $c_1$ entails $c_2$, then the antecedent of (O1) is impossible,

[168] Bennett (2003)

making (O1) trivially true. We wouldn't want this to generate overdetermination cases.

Returning to Kim's causal exclusion argument, the following conditions must be true non-trivially for P and M as causes of P* in order to qualify as a case of overdetermination:

> (O1) There is a possible world in which *M* occurs without *P*, and in the nearest world in which *M* occurs without *P*, *P** still occurs.
>
> (O2) there is a possible world in which *P* occurs without *M*, and in the nearest world in which *P* occurs without *M*, *P** still occurs.[169]

Bennett argues that either condition (O2) is false, or else condition (O1) is trivially true. First, she observes that neither *M* nor *P* is realistically a sufficient cause of *P* on its own*. Rather, they are sufficient causes only in light of certain background conditions. The event *P* in my brain at *t* is sufficient to cause the event *P** in my brain at *t** only *ceteris paribus*, in light of background conditions involving my neurons firing in the appropriate way, my not being hit by a bus or suffering an aneurysm between *t* and *t**, a nuclear blast not obliterating earth between *t* and *t**, and so on. Call this conjunction of events *P^A*.

Second, recall how at the beginning of the Causal Exclusion argument, Kim identified the physical cause of *P** in the brain, *P**, with *P,* the supervenience base of

[169] ibid.

*M*. However, this is implausible for a number of reasons. For one thing, an externalist about mental content will hold that mental event *M* – my belief that I am human and all humans are mortal – supervenes upon a much larger set of physical states than simply my brain event *P**, now identified with *P*.[170] Among other things, its supervenience base must include the natural kind *human* and on the natural kind *mortal* and the history by which I obtained the concepts *human* and *mortal*. For another thing, we have no reason to believe that the causal patterns in the brain strictly obey folk psychology: the pattern of neurons which fire to cause brain event *P** is highly unlikely to be identical to the pattern of neurons which fire when I experience the conscious mental state of believing that I am human and all humans are mortal. It is far more likely that M will supervene on some conjunction of *P* and various background conditions: something like *P^A*, where *A* includes a number of facts about the world besides what is going on in my head, such as the fact that the earth is not destroyed by a nuclear blast.

Nonetheless, suppose it is the case that *P on its own*, without any background conditions, necessitates *M*. It follows that condition (O2) is trivially true. Since *P* entails *M*, there is simply no possible world in which *P* occurs but *M* does not occur. So, there is no threat of overdetermination.

Suppose, more realistically, that it is a complex statement containing a number of background conditions on which *M* supervenes, like *P^A* . Now, condition (O2) is false. Were *M* not the case, yet *P* still the case, then *A* would have had to be different in some way (again, if *A* were not different, then *M* would still have

[170] See Kobes (2009), and Burge (2009)

occurred). However, if *A* had been different, then *P* would no longer be sufficient to guarantee *P**. To illustrate why this is the case, suppose instead that P is the event in my brain from which my desire to move my arm M arises. Let P* be the moving of my arm.[171] Suppose one were to replicate *P* in a slice of brain in a vat that is being artificially stimulated. This is a case in which *P* occurs, but *M* does not occur (I can't desire to move my arm, since "my arm" fails to refer to anything). It's also, not coincidentally, a case in which *P** does not occur – the brain in a vat doesn't move anything. So, the nearest world in which *M* is absent but *P* is the case is a world in which *P** does not occur. Thus, there is no overdetermination.

Bennett's response provides a plausible and well-justified escape route for the emergentist from Kim's snare. Most notably, Bennett's account reveals the importance of counterfactual accounts of causation to our understanding of what is objectionable about overdetermination: *it is not simply an event's having two sufficient conditions which makes something a case of real overdetermination, but an event's causally depending upon neither cause.* However, in part because the response is so highly technical, one is left worrying that she has never really diagnosed the central problem with Kim's argument. I will turn next to an approach which aims to identify the problem with Kim's argument in a mistaken intuition.

---

[171] Note that, while the illustration substitutes for *P** a bodily state instead of a brain state, the same applies even if *P** is a brain state: something which it causally depends upon could be missing were *A* changed.

## 4.2 Contextual Parameters

To many people, it certainly *seems* as though there is something wrong with a physical event having both a physical cause and a mental cause. To them, it seems as though the physical cause ought to exclude the mental cause: it cannot be both that my arm moves because I decide to move it, and at the same time that my arm moves because of an underlying neurophysiological process which guarantees that it will happen. My decision to move my arm can only count as exercising causal powers if (i) my decision is a *joint cause* alongside the neurophysiological process, which *adds* novel causal forces to it (i.e., a theory of telekinesis), or else (ii) my decision *overdetermines* the movement of my arm. These are the only options.

While I do not share this intuition, if one does have a strong intuition that mental causes must be either joint or overdetermining to be real, then technical solutions like those given by Bennett (2003) are unlikely to be very convincing – it will seem more probable that something has gone awry in the mechanics of the rebuttal than that the underlying intuition is mistaken. Even if there is nothing wrong with mental causation, it would be helpful to understand why it seems so.

One explanation is that our intuition results from ignoring features of context.[172] Recall that causal explanations occur at different levels, and recall also that counterfactual accounts of causation require alternative possibilities to be ranked in terms of their nearness to the actual world, a ranking which may depend

[172] Horgan, T. (2001). Maslen, C., Horgan, T., and Daly, H. (2009).

upon the context of study.[173] Suppose that one important contextual feature is the *level* at which causal relationships are being studied: which alternative possibilities are nearer the actual world will change depending upon whether one is talking physics, biology, or psychology. Normally, this implicit level-parameter doesn't cause many problems; normally, we are only discussing causation at a particular level, in which case it is safe to assume that overdetermination only occurs in the case of intentional arrangement or coincidence.

When philosophers discuss downward causation, they ignore this contextual parameter. Consider my decision to answer a phone call, and the physical movement of my hand to push a button. On what does the movement of my hand causally depend? If we ask the question at the level of physiology, then the relevant alternative possibilities will be cases in which things went differently physiologically: I wouldn't have pushed the button if my muscles hadn't contracted. These will be different alternative possibilities than those relevant at the level of psychology, on which I wouldn't have pushed the button if I hadn't wanted to, even if my muscles were functioning in the same way. Both of these will be different than the possibilities which are relevant when we ask the question at the level microphysics. Philosophers have the intuition that there is something wrong when they run roughshod over the contextual parameters and treat downward (mental-physical) or upward (supervenient) causes as same-level causes.

---

[173] This context-sensitivity does not entail a deflationary account of causation. One can hold that different real causal relations are picked out depending upon the context of a discussion.

This solution to the exclusion argument has the disadvantage of requiring a denial of many of the underlying assumptions in Kim's argument: Kim might, for example, reply by ruling out contextual counterfactual accounts of causation as in conflict with the more "serious" productive or generative causation that an emergentist wants.[174] However, it offers a clear and plausible explanation for why *we should not trust our intuitions about overdetermination*. Thus, Horgan (2001) explains the intuition in Kim's arguments as a kind of cognitive illusion – one which persists even if one knows the contrary to be the case.

### 4.3 Who's Afraid of Overdetermination Anyway?

A third response to Kim's argument which I wish to highlight is that of Bernstein (2010), who embraces the conclusion of Kim's *Reductio* that there is widespread overdetermination in the case of downward mental causation, but argues that widespread overdetermination in the absence of coincidence isn't problematic after all. Consider that there are a variety of events in the world which occur regularly and might be called "overdetermination": does the rock thrown at the window break it, or do the particles of which the rock is composed break the window? In some sense, both do. But this isn't problematic, because the rock and its parts aren't mereologically distinct. Overdetermination needs to be defined in a way which excludes these "trivial" cases and yet includes what the nonreductivist or emergentist wishes to say about the relationship between the mind and the brain.

---

[174] For instance, it is often observed that counterfactual accounts allow for causation by absence, but absences are not supposed to be productive or generative (or so it is said). See Kim (2010), 251-253

Why should anything be wrong with overdetermination to begin with? It's not generally objected that overdetermination is physically or conceptually impossible. Does overdetermination in mental causation seem too widespread to be a coincidence and demand further explanation? Perhaps, but the emergentist offers a further explanation, in terms of predictable trans-ordinal laws (or manipulability relations). We don't generally complain that law-like regularities in nature are *prima facie* implausible by arguing that regularities have to be either intentional or coincidental.

Bernstein's reply has the advantage of compatibility with the whole of Kim's supervenience and exclusion arguments, without requiring modifications to his definition of overdetermination or his assumptions about causation – she simply embraces the conclusion. At the very least, those who defend Kim's argument need to give a further explanation of *why overdetermination is supposed to be unacceptable*.

I will now turn to develop my own response to the problem of overdetermination. I believe that Bernstein, Horgan, and Bennett have given plausible responses to the problem of overdetermination which are fully compatible with my own. Like Bennett's account, mine recognizes the difference between an event with two sufficient conditions and a genuine case of overdetermination. That said, my own response does not require modifying Kim's definition of overdetermination nor embracing the conclusion he considers a *Reductio*. It does require commitment to a counterfactual account of downward causation, but not to a specifically contextualist account. I also believe it has the advantage of explaining

why a certain form of overdetermination is supposed to be unacceptable and why so many philosophers find the intuition in Kim's case compelling.

## §5 Two Notions of Sufficiency

### 5.1 Two Ways to be Sufficient

I believe that all humans are mortal. I also believe that I am human. From these two premises, I form a new belief – I conclude that I am mortal. There are two relationships between my beliefs and my conclusion which could be described with the word "sufficient". First, there is the sense in which *the truth of my beliefs* is sufficient for the truth of my conclusion: given that I am human and all humans are mortal, I *couldn't fail to be* mortal. Second, there is the sense in which *that I hold these beliefs* is sufficient to bring about my conclusion, even though I could have held the same beliefs and yet failed to come to the conclusion, because *nothing further was needed* for me to come to the conclusion that I was mortal.[175] I'll label these the *deductive* and the *completive* notions of sufficiency, respectively.

I believe that the causal incompatibilist makes the mistake of reasoning within Kim's argument as follows:

Given $A$, $S$ is *sufficient* for $T$ to be the case

$\therefore$ Given $A$, $S$ is a *sufficient condition* of $T$

[175] For example, it does not seem to me that some further *act of will* is needed for me to make the inference. But I could have failed to make the inference when I did nonetheless, even taking as background assumptions that I reflect on my beliefs and that I am not suffering from any obvious defect of rationality.

I believe this is fallacy of equivocation, where the ordinary completive sense of "sufficiency" is confused with the logician's deductive sense of a sufficient condition. In this section, I will describe each notion of sufficiency, and attempt to explain why neither entails the other.

## 5.2 The Deductive Notion

The deductive notion of sufficiency – call it sufficiency$_D$ – is the technical sense of sufficiency familiar from introductory logic, the sense of "sufficient" used in the phrase "sufficient condition". One state of affairs (or property instance, or event) is sufficient$_D$ for another when and only when it *entails* the other.

Let $S$ and $T$ be states of affairs, and $A$ be some complex background[176] state of affairs (the background may include laws of nature, definitions, *ceteris paribus* clauses, etc.). Then:

> (sufficiency$_D$) $S$ is sufficient$_D$ for $T$, given background assumptions $A$, iff it is not possible for $S$ and $A$ to be the case and yet $T$ fail to be the case.

Assuming $A$, that $S$ is the case *guarantees* that $T$ is the case: one could structure a valid argument with $A$ and $S$ as premises and $T$ as a conclusion. There is

[176] Bennett (2003), pp 23-25 notes the difference between a "strong" and a "weak" sense of sufficient cause, depending on whether or not background conditions are permitted. My definition of sufficiency$_D$ is her "weak" interpretation of sufficiency, since $A$ is permitted for the deduction of $T$ from $S$.

*no way it could fail to be* that *T* is true, given that *S* and *A* are true. S and *A* together *logically necessitate T*. If the contents of *A* are limited to laws of nature, then *S* can be said to *nomologically necessitate T*.

### 5.3 The Completive Notion

The *completive* sense of sufficiency – I'll call it sufficiency$_C$ – is the sense familiar from ordinary language, where the word "sufficient" means something like "that's all you need" or "that's enough". For instance, a diet of 2,000 calories a day provides sufficient energy for the average adult human, a credit card is sufficient to put down a hotel reservation for a vacant room, and, for most people, a measles vaccine is sufficient to avoid getting the measles. Nothing about the possibility of a valid deduction is implied by these statements.

We might specify a relation of *minimal sufficiency*, on which one thing is minimally sufficient for another when (i) nothing more than it is needed, *and* (ii) it is still needed. For instance, $1,000,000 is sufficient to purchase a movie ticket in 2013 US dollars. It's also "more than sufficient": one hardly needs $1,000,000 to purchase a movie ticket. So, while $1,000,000 may be sufficient to purchase a movie ticket in the completive sense, it's something more like $12 which is *minimally* sufficient to purchase a movie ticket. One actually needs at least $12 to buy the movie ticket – one doesn't need $1,000,000. A few Gigabytes of RAM is minimally sufficient to run the software on my computer; a billion Terabytes of RAM would be sufficient also,

but not minimally sufficient.[177] My eating a popsicle is sufficient for any necessary truth, but not minimally sufficient.

When I consider the fact that all humans are mortal and the fact that I am human, nothing else needs to happen for me to conclude that I am mortal. My beliefs compel me to come to the conclusion that I am mortal, whether I want to or not. *Nothing further is needed*, given these two beliefs and my considering them, for me to conclude that I am mortal. Further, though these particular beliefs were not necessary for me to come to the conclusion ("all mammals are mortal" and "I am a mammal" might have done it), my conclusion counterfactually depends on my having had the beliefs I had – in those nearby worlds in which I doubt the mortality of all humans, or doubt my own humanity, I wouldn't have come to the same conclusion.

Let $S$ and $T$ be states of affairs, and let $A$ be the same sort of complex background state of affairs mentioned earlier. Let $V$ be the vocabulary in which $S$, $T$, and $A$ are expressed. Let $P$ be any arbitrary true sentence in $V$. Then we'll define minimal completive sufficiency, or sufficiency$_C$ as:

(sufficiency$_C$) $S$ is sufficient$_C$ for $T$, given background assumptions $A$, iff both:

i. (dependence condition) were $S$ not the case, then $T$ would not have been the case in some nearby$_A$ world, *&*

[177] Credit belongs jointly to Jennifer Matey and to David Chalmers for convincing me of this point on separate occasions.

ii. (exclusivity condition) there is no $P$ which is counterfactually$_A$ independent of $S$ and not entailed[178] by T for which is it true that, if $P$ were not the case and yet $S$ were the case, $T$ would not have been the case in some nearby$_A$ world.

Here, the background conditions $A$ play the role in evaluating the counterfactual of sorting other possible worlds into those which are more or less nearby the actual world (with respect to how much they differ relative to $A$).[179] This differs from the role $A$ plays in the definition of sufficiency$_D$, where it serves as a premise in the deduction of $T$.

Notice that sufficiency$_D$ doesn't require $T$ to depend on $S$, but sufficiency$_C$ does. On the other hand, sufficiency$_C$ doesn't rule out the possibility of $S$ and $A$ being the case and yet $T$ failing to be the case, but sufficiency$_D$ does. However, both sufficiency$_C$ and sufficiency$_D$ rule out the possibility that some other $P$ independent

---

[178] Otherwise, it would be possible for all T that P = "T v U", in which case $T$ counterfactually depends on infinitely many instances of $P$.

[179] On a sophisticated method of evaluating counterfactuals, like Woodward (2005)'s interventionism, the counterfactuals which are relevant (e.g. to causation) are restricted in some way – to those involving possible interventions, for example. Here I would take these restrictions to be something built into $A$. Likewise, on various contextualist methods of evaluating counterfactuals, $A$ should be construed as containing a contextual parameter.

of $S$ is a necessary condition of $T$ – sufficiency$_C$ does so explicitly, and sufficiency$_D$ entails this is the case.

## 5.4 A formal example of the distinction

To see how it can be the case that $S$ might be sufficient$_C$ for $T$ but not sufficient$_D$ for $T$, or that $S$ might be sufficient$_D$ for $T$ but not sufficient$_C$ for $T$, consider the following two models (see Table 2), where *a, b,* and *c* are independent sentences in some vocabulary which are true in $w_a$, the actual world, and $w_n$ are the various other possible worlds at which they are assigned truth values. Let's further specify that worlds $w_{1\text{-}6}$ *are nearby* $w_a$, where our set of background assumptions $A$ ranks the nearness of worlds.

Table 2

*Completive and Deductive Models*

| C-Model | D-Model |
|---|---|
| $w_a$: *a, b, c*<br>$w_1$: ¬*a, b, c*<br>$w_2$: *a,* ¬*b,* ¬*c*<br>$w_3$: ¬*a,* ¬*b,* ¬*c*<br>$w_4$: *a, b,* ¬*c*<br>& there is no $w_n$ *nearby* $w_a$ with: ¬*b, c*<br>& there is no $w_n$ *nearby* $w_a$ with: ¬*a, b,* ¬*c* | $w_a$: *a, b, c*<br>$w_1$: ¬*a, b, c*<br>$w_2$: *a,* ¬*b,* ¬*c*<br>$w_3$: ¬*a,* ¬*b,* ¬*c*<br>$w_5$: ¬*a,* ¬*b, c*<br>& there is no $w_n$ with: *b,* ¬*c* |

Notice that *b* is sufficient$_C$ for *c* on the C-model and not on the D-model. The dependence condition holds on the C model but not the D model: on the C-model it is true that, in all nearby worlds in which ¬*b* is the case, ¬*c* is also the case. But the D-model contains $w_5$, in which ¬*b* is the case and ¬*c* is not. (The exclusivity condition

holds on both the C model and the D model: on neither model is there a world in which $\neg a$ is the case and $b$ is the case, yet $\neg c$ is the case.)

Notice also that $b$ is sufficient$_D$ for $c$ on the D-model and not on the C-model. There is no possible world on the D-model on which $b$ is the case but $c$ is not the case. However, there is such a possible world on the C-model: $w_4$. Even though $b$ is sufficient for $c$ in the ordinary sense – $b$ is enough for $c$ and nothing more than $b$ is needed for $c$ – $b$ is not sufficient for $c$ in the technical sense, and it would be invalid to deduce $c$ from $b$.

### 5.5 Intuitive Examples

Intuitive examples where a state of affairs is sufficient$_D$ but not sufficient$_C$ for some other state of affairs are easy to find in existing philosophical literature about causal explanation. That a man takes birth control pills is sufficient$_D$ but not sufficient$_C$ for his not getting pregnant.[180] Each soldier's bullet in the firing squad was sufficient$_D$ for the execution but not sufficient$_C$.[181]

In other words, nomological necessitation is not the same thing as nomological dependence.[182]

---

[180] See Salmon (1971), 34

[181] Since the dependency condition involves a counterfactual, it faces the "preemption" issue: it isn't true for any one soldier's bullet that the execution depended on it. See Lewis (2000). Only the sum of the bullets is sufficient$_C$ for the execution.

[182] Consider how the height $h$ of a pole and the angle $\theta$ of the sun above the horizon, are sufficient$_D$ causes of the length $s$ of a shadow cast by a flagpole, since $s$

Intuitive examples where a state of affairs is sufficient$_C$ but not sufficient$_D$ for some other state of affairs are much more difficult to come by. Practically speaking, we can usually get away with the fallacious inference that because *nothing else is needed* besides a cause to bring about a certain effect, the effect was guaranteed by the cause.[183] Nonetheless, suppose a historian truthfully says: "Once hope of economic improvement had dissipated and social media made organizing possible, the attempted assassination of the opposition leader was sufficient to cause the

---

can be deduced from $h$ and θ given laws about the rectilinear propagation of light. Of course, $h$ and θ are also sufficient$_C$ for $s$ – were they different, $s$ would have been different. But while $s$ and $h$ are also sufficient$_D$ for $\theta$ – because $\theta$ can be deduced from them – they are not sufficient$_C$ for $\theta$. Changing the height of the flagpole and the length of its shadow will not change the angle of the sun on the horizon in nearby world. Example from Woodward (2011)

[183] Note that coming up with a clear example requires considering causal patterns that occur on a particular higher level than fundamental physics – those of a special science like forestry or economics or history – in large part because CCPD rules out any physical event with a cause that doesn't have a sufficient$_D$ cause. This isn't a problem with forestry or history, since closure does not apply to higher domains. This doesn't mean there are no law-like, predictable regularities in higher domains. Just that it isn't plausible that *every* historical event (for example) could be deduced from antecedent conditions in history with the use of these historical laws, even in an ideal and perfect Historical Science of the future.

revolution."[184] The historian seems to be claiming that the assassination was sufficient$_C$ for the revolution – nothing more was needed.[185] But the historian almost certainly does not mean the assassination was sufficient$_D$ for the revolution. The historian can't mean that *it would be impossible* for the assassination to have happened with the same background, and yet for there to have been no revolution. In fact, it might even have been a *nearby* possibility.

## 5.6 Towards a Causal Distinction

Neither notion of sufficiency is intrinsically causal.[186] However, we might develop this distinction into one between a *sufficient$_D$ cause* versus a *sufficient$_C$*

---

[184] This is not intended as reference to an actual historical event. I'm working off the view that things like economic hardship, violated expectations, a government's lack of perceived legitimacy, ability to organize an opposition, or so on, are something like the ingredients of revolution.

[185] I take the historian to mean that the attempted assassination was sufficient$_C$ for revolution because (i) were there no assassination, there wouldn't have been a revolution, and (ii) beyond the assassination, given the other ingredients already in place, nothing else was needed for a revolution. The claim is thus that if these antecedent conditions obtained, and yet revolution failed to come about, then there would be no other sentence in the historian's vocabulary left such that the failure of the revolution could be attributed to the failure to obtain of the state of affairs expressed by the sentence.

[186] For one thing, both notions of sufficiency are as compatible with *synchronic* relationships as they are with *diachronic* relationships.

*cause*. Let's say an event *S* is a sufficient$_D$ cause of some effect *T* given certain background assumptions *A* only if it is not possible for *S* and *A* to be the case but not *T*. [187] An event *S* is a sufficient cause of some effect *T* given background assumptions *A* only if in the nearby worlds with respect to A where *S* is not the case, *T* is not the case, and there is no other sentence *P* of which this can truly be said while keeping *A* fixed (where P is some sentence in the relevant vocabulary[188] which is counterfactually independent of *S,* keeping A fixed[189], and not entailed by *T*) [190]

---

[187] It might be noticed that each sense lines up imperfectly with a different class of accounts of what a sufficient cause is. Deductive accounts of causation (like those which follow after the D-N model of explanation) are those on which all causes must be sufficient$_D$ for their effects given background conditions and laws of nature. Counterfactual accounts of causation are those on which causes must meet the dependence condition (but not the exclusivity condition) of the definition of sufficient$_C$ with respect to those worlds which are nearest given background conditions and laws of nature.

[188] By the "relevant vocabulary", I mean those predicates which express properties within the special science which operates at the level at which the purported cause occurs. Microbiology and Forestry operate on different "levels", and one need only exhaust the vocabulary of Forestry to make true a 'nothing else is needed' claim on the level of forestry, without having to further exhaust the vocabulary of microbiology in order to meet the exclusivity condition.

[189] Notice why it is essential that A be kept fixed for all counterfactual evaluations. Otherwise a distal cause (for which the counterfactual is true) will rule

It should be noted that the distinction between a sufficient$_D$ cause and a sufficient$_C$ cause is not whether they involve laws or background conditions or not – both of them do.[191] Likewise, the difference is not that a sufficient$_D$ cause is *generative* in some way that a sufficient$_C$ cause is not.[192] Nor is the difference that one sense of

---

out every more proximal cause. Let C1 be a cause of E at t1, and C2 be a cause of E at t2. Suppose were it not for C2, E would not be the case; and were it not for C1, E would not be the case. This seems to rule out the exclusivity condition, even if nothing else was needed at t2 for C2 to cause E. Surely C1 does not depend on C2, which happened later in time! However, this is only true *when we use a different set of background assumptions than A*. A is generally taken by Lewis (1973) and others to include all events prior to t2. If *A* is kept fixed, then "were it not for C2, C1 would not be the case" becomes a *backtracking counterfactual*, which is in fact true.

[190] Both of these are "only if" claims, not "if and only if" claims, because I want to leave it open that something further is needed to make the relationships specifically causal.

[191] One might suppose that counterfactual accounts do not rely upon laws of nature or lawful regularities, but they must in fact do so in order to differentiate near from distant worlds. This remains true even if contextual factors are in some cases given priority. (Kim 2011, 211-213) Instead, the distinction is between whether these laws are used as part of a deduction from causes to effects, or whether they are used to map out the nearness of other possible words.

[192] In fact, it seems to me that the ordinary "that's enough" sense of sufficient$_C$ is exactly what we tend to have in mind when we talk about causes bringing about

"sufficient cause" is deflationary and the other realist.[193] The difference is between a cause that rules out all alternative effects, and a cause that exhausts all causes that could change the effect. That is, the difference is between a deductive account of causation, on which a cause must be sufficient$_D$ for its effects, and a counterfactual account of causation, where causation is causal dependence, and the exhaustive set of all causes on which an event depends sufficient$_C$ for its effects.

---

their effects. I seems clear to me that my belief that I am human and that all humans are mortal generates and produces in me the conclusion that I am mortal. Whereas it is a common mistake for introductory logic students to suppose that sufficient conditions bring about that which they are sufficient conditions of, and so it seems like a fallacy to me to suppose that there is something productive or generative going on simply because one can get a deduction going.

[193] Indeed, the distinction between deflationary vs. realist accounts of causation is orthogonal to the distinction that I am making here. Lewis (1973) gives a counterfactual account of causation which is deflationary, but Woodward (2005) gives an interventionist counterfactual account which is not deflationary. Deductive accounts of causation can be deflationary, like that of Hempel (or, famously, Hume), under which laws of nature are only regularities without an observed exception. Deductive accounts can also be non-deflationary – we can take laws of nature to be real laws, uncovered by scientific investigation. The deductive/counterfactual distinction involves the form that causal claims take, as opposed to the ontological status of causal relations.

## §6 An Alternative Response to Kim

### 6.1 Return to the Scene of the Crime

Let's return to Kim's Supervenience Argument. Notice that Kim's opponent need not accept that mental events are sufficient causes of physical events in the technical sense, and should at most accept that mental events are sufficient causes in the ordinary sense. This move offers another way out of Kim's exclusion argument. It allows for a robust sense in which mental events are sufficient causes of physical events, while remaining compatible with the supervenience of the mental on the physical.

Let M and M* be mental events and P and P* be physical events, where P is the supervenience base of M and P* is the supervenience base of M*, and M causes M*. For simplicity, suppose that P also causes P*.[194] This is the sort of picture of mental causation which middle positions on the mind are generally willing to accept.

Kim argues that M must cause M* *by causing* M*'s supervenience base, P* as follows. First, assume one must adopt either a deductive or a counterfactual account

---

[194] Of course, a good deal hangs on this assumption, and I see in Bennett (2003) two reasons for why this assumption isn't plausible. First, if externalism about mental content is true, then the physical base of M and the physical base of M* aren't just in the brain – they include historical relations between the brain and the world and how things actually are in the world. But one of these states is unlikely to be a sufficient cause of the other. Second, the sorts of background conditions *A* which are needed for *P* to be a sufficient cause of *P** are unlikely to be part of the supervenience base of M.

of mental causation.[195] On a deductive account, M causes M* only if M is sufficient$_D$ for M* to happen. But M* supervenes on P* -- which means that P* is also sufficient$_D$ for M*. So, M* has two sufficient causes, and according to the Exclusion Principle, either this is a genuine case of overdetermination or M and P* are not distinct causal chains – M causes M* by causing P*.

On a counterfactual account, on the other hand, M causes M* only if M* counterfactually depends on M. But, once again, M*'s supervenience base P* is sufficient$_D$ for M*. So, were M not the case but P* still the case, M* would still necessarily be the case because of P*. For M to be the cause of M*, then it would also have to be true that *were M not the case, then P* would not be the case*. So, on a counterfactual account, M must cause P*. Either way, Kim forces his opponent to accept the downward causation of P* by M.[196]

However, one should note that the Kim can only compel his opponent at this stage to accept, at most, the *counterfactual* causation of P* by M. He can't compel his opponent to accept that M is a sufficient$_D$ cause of P* by the supervenience argument alone. (See Figure 6)

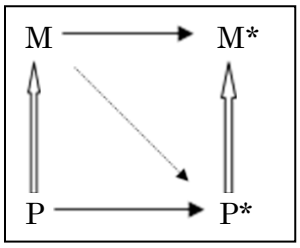
---

[195] I am simplifying the argument for ease of exposition. Kim is well aware that there are accounts of mental causation which fall into neither category, of course, and has argued that his objection applies to them equally.

[196] Thus, "same level causation can occur only if cross level causation can occur". Kim (1993), 353.

**Figure 6**

M need only be sufficient$_C$ for P*




## 6.2 Kim's Exclusion Argument

Kim's Exclusion argument follows from here. Kim notes that P* has two causes – a physical cause (P) and a mental cause (M). Now, according to CCPD, P must be a sufficient$_D$ physical cause of P*. Kim also thinks we must accept that M is a sufficient$_D$ cause of P*. Why?

To vindicate *M* as a full and genuine cause of *P**, we should be able to show that M can bring about P* on its own, without there being a synchronous physical event P that also serves as a sufficient cause of P*.[197]

In other words, a full and genuine cause can't rely on *something else* in order to bring about its effect. M must be *all that is needed* for P*, why it must be *enough* to bring about P – why it can't be that a particular synchronous physical event P is *also* needed for P*.

Of course, this is an argument for why M must be a sufficient$_C$ cause of P*, not a sufficient$_D$ cause. And this seems to me the most plausible claim that can be

[197] Kim (2011), 215. Note: I have adjusted the use of P* and M* from the original for consistency with the model used elsewhere in this paper.

made about "downward causation" from the mental to the physical. Mental causes are *enough* to bring about their physical effects. The physical effect would not have occurred in the absence of the mental cause, so the dependency condition holds. The physical effect doesn't depend on anything else in this way either – had the particular physical base P of M not occurred, but M had still occurred (which, in all nearby possible worlds, happens through some physical base or other), then P* still would have occurred. So, the exclusivity condition holds. M is sufficient$_C$ for P*.

However, it is not plausible to suppose that mental causes *entail* their physical effects or that physical events can be deduced from mental events, even given background conditions and ceteris paribus clauses. It is not true that M *guarantees* the occurrence of P* – in fact, M might occur and P* fail to occur. Nothing could rule out the possibility of M failing to cause P*, though it might be a very *distant* possibility. So, it is not plausible that M is sufficient$_D$ for P*.

If M is merely sufficient$_C$ for P* but not sufficient$_D$ for P*, then the exclusion principle doesn't apply. Only if we were to grant that M is a sufficient$_D$ cause of P*, and that P is also a sufficient$_D$ cause of P*, would it follow under the Exclusion Principle that either P and M aren't distinct, or else that P* is overdetermined by both M and by P.

### 6.3 Reformulate the Exclusion Principle?

One might wonder if Kim really is committed to interpreting the Exclusion Principle as involving sufficient$_D$ causes. He is. The exclusion principle makes reference to overdetermination. For Kim, an event is overdetermined iff it has two

distinct sufficient$_D$ causes.[198] The firing squad case is intended to be an example of this.[199] In genuine overdetermination, two causal chains converge, where each cause, independent of the other, guarantees (necessitates, *determines*) the outcome. This is something which, one might argue, could happen only by chance or in designed scenarios (like the firing squad).

Why not define a different sense of overdetermination, one in which an event has two distinct sufficient$_C$ causes? Call it "overdetermination$_C$". P* is overdetermined$_C$ by M and by P – since P is also a sufficient$_C$ cause of P*.[200] Why not reformulate the exclusion principle in terms of overdetermination$_C$?

But overdetermination$_C$ doesn't describe firing squad scenarios, since both causes are sufficient$_D$ but neither is sufficient$_C$. So the classic "genuine cases of overdetermination" aren't overdetermination$_C$. In fact, the only possible cases of overdetermination$_C$ are cases where one cause counterfactually depends on the other cause. So, whereas Kim's Exclusion Principle is a substantive claim, the sufficiency$_C$

---

[198] Kim defines overdetermination for effect E and causes C and C* as "E would have occurred even if either C or C* had not occurred … the other would have been sufficient to bring it about." Kim (1993), 252. Note on 253, however, Kim identifies this relationship with "completeness" and contrasts it with the D-N model.

[199] Kim (2011) and elsewhere.

[200] Since P is sufficient$_D$ for P*, it meets the exclusivity condition. Since M depends on P (*ex hypothesi* per the emergentist and non-reductive physicalist), and P* depends on M, then by transitivity P* depends on P, meeting the dependence condition.

version of the exclusion principle is analytic simply given the definition of sufficiency$_C$. By definition, if two causes are minimally completively sufficient (sufficient$_C$) for one event, *then the two causes must be dependent on each other*.[201] Hence the following is principle:

> *(EPC) Exclusion Principle$_C$*. No event has two or more *independent* sufficient$_C$ causes.[202]

This is, of course, exactly the relationship which emergentists and non-reductive physicalists have been claiming holds between M and its supervenience base P: M is dependent upon P.

---

[201] Otherwise, it would be false that "there is no P counterfactually independent of S and not entailed by T for which is it true that, if P were not the case and yet S were the case, T would not have been the case in some nearby$_A$ world."

[202] Notice three differences. (i) "independent" (which follows from the definition of sufficiency$_C$), rather than "distinct". Obviously, the emergentist holds that M and P are metaphysically dependent but distinct. (ii) this principle doesn't make reference to time. Thus, an event can't even have a sufficient$_C$ distal cause and a sufficient$_C$ proximate cause with the same background conditions A – whereas an event can have a sufficient$_D$ distal cause and a sufficient$_D$ proximate cause without involving overdetermination. (iii) this principle doesn't add 'unless it is a genuine case of overdetermination', since all cases of overdetermination$_C$ are cases where one cause is dependent on the other.

### 6.4 Summary

In summary, Kim's supervenience argument at most compels the supervenience theorist to accept *counterfactual* downward causation. I have argued that the most plausible accounts of mental-to-physical downward causation involve sufficient$_C$ ("all you need") causation but not sufficient$_D$ ("no way to fail") causation. Neither counterfactual mental-to-physical causation nor sufficient$_C$ mental-to-physical causation are cases to which Kim's Exclusion Principle applies, and they present no threat of overdetermination with the sufficient$_D$ physical-to-physical causes guaranteed by causal closure, leaving a way out of Kim's Exclusion argument.

It is my suspicion that the principle that no event has two or more independent sufficient$_C$ causes is the sort of thing which might be at work leading us to find Kim's Exclusion Principle intuitive. No event can have more than one minimally "complete cause", unless one cause depends in some way on the other. It seems to me that a serious form of emergentism is compatible with this intuitive claim, and with the causal closure of the physical. I will briefly turn to a few objections which might be levied against my reply to Kim.

## §7 Objections to Emergence and Mental Causation

### 7.1 Who Needs Ordinary Language?

Philosophers can start to become weary and exhausted by arguments claiming there are multiple senses of a word – particularly ones which contrast a "folk" use against a technical use. "Who cares what the folk think?" someone might say. On this objection, logicians have discovered the real sense of sufficiency – the notion of a sufficient condition – and it is up to everybody else to catch up.

I do not share this disdain for ordinary language. However, in defense of my particular argument about the two senses of "sufficiency", it is worth pointing out that the "ordinary" notion of sufficiency (the sense of being "enough") is the sense in which the sciences regularly use the term and the only notion of sufficiency which can be justified by the process of experimentation. The question of whether a cause is sufficient$_C$ for its effect is *empirical*; the question of whether it is sufficient$_D$ for its effect is not empirical. The notion of completive sufficiency should not be dismissed.

Consider an experiment in which a researcher is establishing the relationship between the temperature of a liquid and the rate at which it evaporates. The researcher might set the value of the temperature variable at a variety of settings, keeping the values of all other variables fixed, and then measure the corresponding changes on the evaporation-rate variable. A process like this establishes a kind of dependence of the evaporation-rate variable on the temperature variable.

Suppose a researcher wished to compile a set of all of the variables on which evaporation rate depends, and that this researcher had an inexhaustible source of grant money, technological ability, graduate assistants, and time. Perhaps he is a scientific archangel, who can conceivably go through every single variable in the entire cosmos and test it in this manner: for every *x*, keeping other variables fixed, what effect does a change in variable *x* have on evaporation rate? Perhaps he could even tweak around with the fundamental laws of physics. What would the result of this experiment be? The result would be an exhaustive list of everything on which evaporation rate depends: pressure, temperature, density of the substance, and so on. Perhaps he could then say with full confidence, for any given setting of these variables, exactly what the evaporation rate would be, because a setting on this

exhaustive list of variables would be sufficient$_C$ for the given effect. The evaporation rate would depend upon these variables and upon nothing else.

He might express his results as a *law* of evaporation-rate. However, he would have to mean "law" in the regularity-theorist's sense of "law". He could not mean that there is some law of nature such that, if the antecedent conditions of the law hold, some effect must follow with *logical* necessity. Observation of experiments does not give us laws of that sort. So, the researcher cannot mean that the set of causal variables is sufficient$_D$ for the given effect because his process of experimentation has not involved observing all of the logically possible worlds and ruling them out. The only way to establish that *S* is a sufficient condition of *T* is to show either that the definition of what it is to be *S* contains what it is to be *T* (that is, from the perspective of ontology[203], *T* is *analytic* given *S*) – or else to maintain that there is some sort of necessary brute fact which links *T* and *S* which admits of no further explanation.[204] Neither of these are studied by controlled experiment.

---

[203] Of course, in a given language, *T* might be synthetic and a posteriori given *S*, as Saul Kripke famously showed in the case of Water = H2O. However, from the perspective of ontology, what it is to be Water is identical to what it is to be H2O: there is nothing contained in the "real definition" or essence of one which is not contained in the essence of the other (regardless of sociolinguistic facts).

[204] For instance, certain kinds of non-naturalism in meta-ethics hold that the natural facts are sufficient$_D$ for the moral facts without being analytic given them.

## 7.2 Compatibility with Laws?

It should be clear, then, that I am not denying the existence of causal laws. My response given here is compatible with holding a nomological account of physical-to-physical causation, of mental-to-mental causation, and of the physical-to-mental supervenience relationship. All that my response to Kim requires is that one deny that *downward* mental-to-physical causation can be expressed by laws which carry any sort of modal force (that is, beyond mere statements of regularity). When my decision to answer the phone causes my hand to push the button, it is incorrect to say that my decision made it necessary that my hand move to push the button. Regardless of which account one prefers of causation generally, I think it is quite strange to hold that *mental causation* fits the deductive model of causation, or that mental causes are ever sufficient$_D$ for physical effects.

Consider how my believing that I am human and that all humans are mortal causes me to conclude that I am mortal, which causes me to utter the word "shucks". My utterance of "shucks" is a physical effect of a mental cause – my believing a pair of propositions.

Let $p^*$ be a description of the physical and neurological state corresponding to my utterance of "shucks", and let $m$ be my belief that I am mortal. There seem to be many slightly different ways[205] that $p^*$ could have occurred which are equally compatible with $m$: a different neuron could have fired, a molecule could have been missing from my brain. My utterance of "shucks" could have been realized in ways

[205] Two distinct "way things could have occurred" are two distinct physical states, which could be described differently in the language of fundamental physics.

which are different at a fine-grained level. This shouldn't be a worry for the claim that *m* caused *p** – mental causation just doesn't operate at such a fine-grained level. But it is a worry for any claim that *m* was – given some law of nature and set of initial conditions – sufficient$_D$ for *p**. My belief that I am mortal didn't guarantee or entail or necessitate my saying "shucks".

What does it mean to say that my belief in my own mortality caused me to say "shucks"? It seems natural to think it means something along one of these lines: I wouldn't have said "shucks" if I didn't just come to the conclusion that I was mortal; or I wouldn't have said "shucks" in the time and manner that I did if I didn't come to the conclusion that I was mortal in the time and manner that I did; or that the best way to have stopped me from saying "shucks" right then, if you had wanted to, would have been to prevent me from coming to the conclusion that I was mortal.

Maybe one could compile an exhaustive list of everything that could have possibly gone wrong and prevented me from saying "shucks" once I had come to the conclusion that I was mortal. That list together with the fact that I concluded I was mortal would then suffice$_C$ for me to say "shucks", since nothing else would be needed. But there's no reason to be so optimistic as to think that even an ideal reasoner could derive from all of that together that I, as a matter of logical necessity, must say "shucks". This isn't to suggest that something further had to intervene to make sure that I actually said "shucks" upon concluding that I was mortal. Concluding I was mortal was enough to make me say "shucks". It's just to deny that my failing to say "shucks" would be *contradictory to* my concluding that I was mortal given the laws of nature, and some set of background conditions.

Most counterfactual accounts do not require that causes be sufficient for their effects in either sense – causes may be only partially responsible for their effects. That said, the strongest causal claim possible on counterfactual accounts is that of sufficiency$_C$ – that an effect depends (in the relevant counterfactual way) on its cause, and that it doesn't depend on anything else. It is not asserted that causes must be sufficient$_D$ for effects.

## 7.3 Forces, Powers, and Production

Someone might believe that in order to be a productive cause – to really make something happen – a cause has to be deductively sufficient for its effect. Someone who forms this belief falls into the converse of the fallacy mentioned earlier:

Given $A$, $S$ is a *sufficient condition* of $T$

$\therefore$ Given $A$, $S$ is *sufficient* for $T$ to be the case

There is a certain phenomenology associated with drawing a conclusion from a set of premises. Introductory Logic students often have to be taught that this feeling of being *oomphed* or "following from the premises" is not part of validity. As discussed earlier, that a set of premises is a sufficient condition for the conclusion of a valid argument does not mean that the conclusion *depends* upon the premises in any sense. The dependence condition of the ordinary sense of "sufficiency" does not hold.

So, deductive causation is no more productive or generative than counterfactual causation. The notion of a productive cause is most likely tied to the pragmatic notion of a possible manipulation of one variable by another, as discussed

in Woodward (2005).[206] In this sense, since it is possible for someone to manipulate whether I push buttons or not by intervening upon the decisions that I make and my reasons for making them, my reasons and decisions count as productive causes of my moving my hand to push the button. The ability to manipulate states of my hand counts as a causal *power* of my decision, legitimizing the ontological status of my decision for one who follows Alexander's dictum.

Does my decision exercise causal force? Here, I would hold that my decision exercises no causal forces over and above those which occur at the fundamental physical level. I would deny any sort of novel, emergent causal forces, and nothing on my reply to Kim requires that such forces exist.

### 7.4 Deny the Argument at an Earlier Stage?

Someone might notice that my distinction between deductive and completive sufficiency might be used to halt Kim's argument at an earlier stage, blocking the supervenience argument rather than the exclusion argument. Someone might deny instead (a) that supervenience requires P to be sufficient$_D$ for M, (b) that P is a sufficient$_D$ cause of P*. This is correct. I've chosen to focus on the exclusion

[206] I believe my notion of a sufficient$_C$ cause is compatible with Woodward's account of causation – i.e., a sufficient cause of *y* would be a set of all of the variables by which the value of *y* can be changed. Woodward offered a unique response to Kim's supervenience argument at the 2012 meeting of the Philosophy of Science Association; Woodward's response is very different from my own approach, but not incompatible with it either.

argument because an emergentist is committed to downward causation anyway, and because there are independent reasons which Kim can cite for holding (a) and (b).

To see how someone might deny (a), consider the account of *manipulative supervenience* which I offered as one option for the emergentist in Chapter 2. On this account, the supervenience of *M* on *P* does not require that *P* metaphysically necessitate *M*, nor that there be a strict law by which *P* necessitates *M*, but only that it be impossible for *M* to be manipulable by any means except for an intervention upon *P*. in other words, that *P necessarily* be sufficient$_C$ for *M* without being sufficient$_D$ for *M*. This would block Kim's use of the supervenience argument to infer downward causation.

To see how someone might deny (b), consider my reflections about experimentation given earlier, and how some contemporary theories of physics make it consistent to deny that prior states of the world are sufficient$_D$ for future states even while holding that nothing else is responsible for future states being as they are, making them sufficient$_C$ for future states. From this perspective, the strongest version of principle (CCPD) justified by experimental evidence is one on which every physical event has a sufficient$_C$ physical cause, but not necessarily a sufficient$_D$ physical case. Thus, within Kim's argument, CCPD at most offers us that *P* is sufficient$_C$ for *P**, much like *M* is sufficient$_C$ for *P**.

## 7.5 Commitment to Determinism?

While I believe that discussions about the nature of free-will and determinism in the philosophy of action are not going to be settled by debates about mental causation in the philosophy of mind, it is not uncommon to connect the two

on some level. Suppose that someone asks whether Emergentism is a form of determinism, or a form of libertarianism?

I believe the best answer is that the form of Emergentism I have offered strongly suggests (though it does not entail), a unique brand of compatibilism. On this view, one can accept the following claim of the Libertarian:

> For at least some human actions, an agent chooses to do *A* at time *t,* and yet could have done otherwise than *A* at *t*, given the whole state of the universe prior to *t*.

At the same time, one can accept the following claim of the Determinist:

> Every human action is an event, and every event at time *t* has a sufficient$_C$ cause in the whole state of the universe prior to *t*.

This form of compatibilism is neither in the spirit of orthodox determinism (which envisions a world in which prior states of the universe are deductively sufficient for future ones) nor is it in the spirit of orthodox libertarianism (on which something further is needed for events to happen – the force of will – given prior states of the universe). At the same time, this form of compatibilism is consistent with the *literal claims* of both determinists (that every future event has a sufficient cause) and of libertarians (that there is a possibility of doing otherwise). While I highly doubt that this resolves the debate, or that droves of adherents will flock to

this sort of emergentist compatibilism, it does show that the current debate may need to be framed in a different manner.

## §8 Conclusions

There are many plausible replies to Jaegwon Kim's causal exclusion argument against emergentism. On the reply which I have offered, Kim's argument depends on equivocating between the logician's deductive sense of a *sufficient condition* and the ordinary completive sense on which something is sufficient when nothing else beside it is needed. It is the ordinary sense alone in which it is plausible to think that mental events are sufficient to cause physical events. Because the form of overdetermination which Kim finds objectionable requires that an event have two sufficient conditions, no such overdetermination occurs when a physical event has both a completively sufficient mental cause in addition to some deductively sufficient physical cause.

My response to Kim's argument has the advantage of leaving Kim's definition of overdetermination intact, accepting (for the sake of argument, at least) that there is no widespread overdetermination, and not requiring a commitment to a particular account of causation generally – except to require that high-level mental-to-mental causes and downward mental-to-physical causes admit of some counterfactual account. Like Bennett's response to Kim, my response involves observing the differences between deductively sufficient conditions and counterfactual dependence claims, but without requiring as technical a mastery of Kim's argument.

My account gives a simple explanation for why overdetermination is supposed to be unacceptable. Contrary to Kim, it is not overdetermination$_D$ which is unacceptable – that is, the claim that an event has two distinct sufficient$_D$ causes.

Rather, it is overdetermination$_C$ which is unacceptable – that is, the claim that an event has two independent sufficient$_C$ causes. This is unacceptable because it is *analytic* that it is not so: necessarily, if an event depends on one cause alone *and* on some other cause alone, then one of the causes must depend on the other.

I also believe my account offers a simple explanation for why so many philosophers find the intuition in Kim's case compelling. Philosophy often makes use of a certain heuristic device, on which questions remain open so long as alternate possibilities have not been ruled out by contradiction. If nothing else is needed for my pushing the button but my decision to answer the phone, then surely my decision guarantees my pushing the button; it simply has to happen, because if it didn't, then why wouldn't it? Surely there would have to be some reason that my pushing the button failed to happen – and that would mean that the physical button-pushing depended upon something other than my decision, and so my decision wasn't enough after all. So long as not every possibility has been ruled out, a true philosopher will expect an answer to the question of why one is the case and not another. And so, philosophers conclude that a truly sufficient explanation would naturally be a deduction-enabling sort of thing. Intuitions say that a complete explanation rules out all other possibilities. Yet, nothing in the logic of things requires this to be the case.

In the next chapter, I will consider the objection that the lack of *a priori* deducibility of consciousness from physical descriptions fails to justify the emergentist's belief in an ontological gap between the physical and the mental.

Chapter 4

# COSMIC HERMENEUTICS

## PART A

### §1. Introduction

Let P be a true sentence in some language, and Q be a sentence in the same language or some other language. Imagine that there exists an ideal reasoner, like Laplace's demon or C. D. Broad's Mathematical Archangel – a conscious intellect not subject to any of the contingent limitations, biases, or errors of finite minds. Suppose this ideal reasoner knows that P, and is tasked with deducing Q *a priori*. There are two questions we might ask. First, if the ideal reasoner could deduce Q from P *a priori*, what would that say about Q? Second, if the ideal reasoner *couldn't* deduce Q from P *a priori*, what would *that* say about Q?

Suppose that P is the conjunction of all of the true sentences in the language of fundamental ontology, and Q is the conjunction of all of the true sentences in any language. Is Q ideally deducible from P? This special assignment for the ideal reasoner might be called *Cosmic Hermeneutics*.[207] We might ask is whether cosmic hermeneutics is possible, and what its possibility or impossibility would mean for metaphysics.

Suppose that we restrict the sentences in P to true sentences in the language of fundamental physics. Suppose we restrict Q to a conjunction of sentences about some higher-level domain, such as the sentences about the qualitative experiences of a subject. Call this Physical-to-Phenomenal Cosmic Hermeneutics (PPCH). Would Q

[207] The term belongs to Horgan (1983)

be deducible from P by an ideal reasoner? If not, what would it mean for the relationship between the physical domain and the higher-level domain?

Many hold that *PPCH is possible if and only if physicalism is true.* Jackson (1998) holds that physicalism is true because he holds that PPCH is possible. Byrne (1999), alongside Chalmers (1996) and Jackson (1982), holds that physicalism is false because PPCH is not possible. Chalmers (2012) holds that because Cosmic Hermeneutics is not possible for sentences about qualia and for indexical truths, we ought to conclude that qualia and indexicals need to be added to the set of fundamental truths. Then, when we take the resulting set of fundamental truths as our basis – the truths about physics, the truths about conscious experiences, the truths about indexicals, and a *totality* sentence stating that there are no further truths which can't be deduced from this set of fundamental truths – Chalmers holds that Cosmic Hermeneutics *is* possible.

Many, such as Block and Stalnaker (1998), have attacked one half of the biconditional: they argue that physicalism might nonetheless be true, even though PPCH is not possible. For Block and Stalnaker, truths about phenomenal states or other higher-level domains are plausibly cases of *a posteriori* necessities, which are entailed by the truths of fundamental physics, but which are not deducible from them even by an ideal reasoner. In response, Jackson and Chalmers (2001) have defended a robust form of two-dimensional semantics on which certain *a posteriori* necessities are guaranteed to be *a priori* for an ideal reasoner with an adequate understanding of the relevant concepts.

Few have attacked the other direction of the biconditional: the claim that PPCH might be possible, and yet physicalism might be false. However, in this

chapter, I will argue that a form of emergentism is compatible with the possibility of Cosmic Hermeneutics. An emergentist could hold coherently that Cosmic Hermeneutics is possible, and yet that physicalism is false.

My position will appear at first to be at odds with Broad (1923), in which emergentism for a domain is defined in terms of the inability of an ideal reasoner to deduce the facts of that domain from the physical facts. Nonetheless, I will locate the emergentist's interest in the task of Cosmic Hermeneutics not in the question of whether the higher-level truths are ideally *a priori* deducible from the physical truths, but rather in the question of what resources would need to be *a priori* for the ideal reasoner in order for the deduction to occur. In itself, that a higher domain of facts is ideally deducible from certain fundamental facts does not tell us that the higher domain is not also fundamental. However, on my view, if a higher domain of facts is merely derivative from a set of fundamental facts, then the higher domain of facts should be ideally deducible from those fundamental facts given only *analytic a priori* premises. If a higher domain of facts is emergent from a set of facts, then it may be deducible from them, but it will be deducible only given *synthetic* premises – and the strength of the synthetic *a priori* resources minimally needed for the deduction will indicate something about the strength of the kind of emergence involved.

I will begin in Part A with an overview of the topic of Cosmic Hermeneutics and relevant history and background. I will start in section 2 with its origins in Laplace and Broad's ideal reasoners, and connect the topic to debates about realism and fundamentality in the 20th century. I will continue in Section 3 by describing the project of Radical Interpretation in the 1960s-1980s. In section 4, I will follow the

shift of the conversation into conceptual analysis and two-dimensional semantics over the last few decades, ending with a discussion in Section 5 of the most recent work on the topic, Chalmers's *Constructing the World* (2012).

In Part B, I will develop my own account, on which Cosmic Hermeneutics may be possible for ontologically emergent domains, provided that the deduction only makes use of ontologically analytic premises.

## §2. Background

### 2.1 Broad's Mathematical Archangel

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated, and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. (Laplace, 1840) [208]

> If the emergent theory of chemical compounds be true, a mathematical archangel, gifted with the further power of perceiving the microscopic structure of atoms as easily as we can perceive hay-stacks, could no more predict the behavior of silver or of chlorine or the properties of silver-chloride without having observed samples of those substances than we can at present

[208] Laplace (1840)

> . . . If the mechanistic theory be true, the archangel could deduce from his knowledge of the microscopic structure of atoms all these facts but the last. He would know exactly what the microscopic structure of ammonia must be; but he would be totally unable to predict that a substance with this structure must smell as ammonia does when it gets into the human nose. The utmost that he could predict on this subject would be that certain changes would take place in the mucous membrane, the olfactory nerves and so on. But he could not possibly know that these changes would be accompanied by the appearance of a smell in general or of the peculiar smell of ammonia in particular, unless someone told him so or he had smelled it for himself. (Broad, 1923)[209]

Both C. D. Broad's "mathematical archangel" and the Laplace's "sufficiently vast intelligence" attempt to resolve a metaphysical question through an epistemic test involving an ideal reasoner – a mind short of omniscient, but with infinite and infallible inferential powers. Laplace's demon is tasked with deducing the past and future facts from the present facts, and the demon's success is taken to entail the truth of diachronic determinism. In contrast, Broad's mathematical archangel is tasked with deducing the higher-level facts at a time from the lower-level facts at the same time, as a test of synchronic determinism. The archangel must deduce the chemical behavior of ammonia given a description of its atomic structure at that same moment, or what it's like to smell ammonia given the simultaneous events in

[209] Broad (1923), 71

the mucous membrane and olfactory nerves. The mathematical archangel's success would indicate that *mechanism* is true; the archangel's failure would indicate that ontological emergentism is true.

Broad accepts that, given *empirical* observations of the behavior of silver chloride, or of the smell which accompanies ammonia, it is possible to formulate laws which predict the presence of emergent phenomena given a description of their underlying conditions. However, Emergentists like Broad and Lewes[210] and others have tended to define an emergent domain as one which is not deducible *a priori* from its basal conditions, even by an ideal reasoner. The metaphysical claim is to be cashed out in terms of an idealized epistemic test.

It's generally accepted that emergentism was wrong about chemistry. Although it might be impossible in practice, in principle an ideal reasoner given the sub-atomic facts and laws would be able to deduce the events which we describe as the behavior of silver-chloride. The theoretical reduction from chemistry to physics we have now was not yet available when Broad wrote.[211] Broad recognized the possibility he was wrong, noting that there appeared to be no *a priori* barrier to

---

[210] "The distinction here indicated between Components and Constituents, or between Parts and Elements, will be seen hereafter to have its importance. . . . The combinations of the first issue in Resultants, which may be analytically displayed; the combinations of the other issue in Emergents, which cannot be seen in the elements, nor deduced from them." Lewes (1874), 98

[211] See McLaughlin (1992).

mechanism about chemistry or biology in the way in which mechanism seemed to be incompatible with phenomenal experience.[212]

However, emergentism might be still right about phenomenal consciousness. No widely-accepted reductive account for phenomenal consciousness has appeared in the years since Broad wrote, and there seems to be little hope held out for one. Emergence for consciousness remains on the table, as an open matter for debate.

Within this debate, whether or not consciousness qualifies as ontologically emergent continues conventionally to be defined in terms of the possibility of an ideal reasoner deducing the facts about phenomenology from the facts about the brain.[213] An "epistemic gap" between basal conditions and an emergent property is supposed to remain under ideal conditions if and only if an "ontological gap" holds between the two.

But why should we consult with a mathematical archangel on matters of ontology? What is the link between ideal deducibility and the structure of the world? What is the task of ontology – and what, if anything, does it have to do with considerations about what can in principle be known or deduced?

## 2.2 The Task of Ontology: From Reality to Fundamentality

One answer is that that the task of ontology is to tell us what exists. This assumes that we do not tell ontology what exists. That is, it assumes that to exist is not merely to be the sort of thing about which people are disposed to say "yes", when asked "does such-and-such exist?". Rather, ontology is supposed to have a *regulative*

[212] Broad (1923), 72.

[213] O'Connor et. al. (2012)

role in telling us whether it is or isn't *correct* to say that *x* exists. Unless one is a Meinongian, there are a whole host of things which don't exist that people talk about. Perhaps we should say ontology is all about what *really* exists, as opposed to what is merely being represented in language or thought as though it existed.[214]

Yet, this distinction between what really exists and what is merely represented as existing has posed persistent problems over the last two-and-a half millennia.[215] In one direction, the distinction between appearance and reality has been at times drawn in such sharp contrast that one loses all hope of knowing anything about reality as it is and not merely as it appears to be.[216] In another direction, the distinction between the two may be so thoroughly obscured that it becomes trivial that everything is as it seems to us to be.[217]

One hopes that philosophy has learned something from these detours over time and recognizes its freedom to pursue a middle path. Suppose someone approaches us with a radically revisionary ontology, which tells us that what *really* exists is inconceivably different from the sorts of things we ordinarily say exist –

---

[214] Reynolds (2006) identifies the function of "real" as indicating that one is reporting on the world rather than on the content of other representations. This gives a clearer understanding of "real" than the usual opposition of "real" to "mind-dependent" or "evidence-dependent", or the collapsing of "real" onto the question of whether a domain is one in which every sentence is true or false.

[215] e.g., at least since Parmenides's *On Nature*.

[216] Kant (1781)

[217] Ayer (1936)

perhaps that matter and energy are a mere projection of some greater mind.[218] We can recognize that this is the sort of claim which goes beyond anything we could possibly hope to know about, and so exclude it from the sorts of propositions which for all intents and purposes are to be regarded as true or false. We can decide that the words *real* and *world* for our purposes should still pick out the matter and energy in this scenario, not the greater mind, because even though matter and energy would turn out not to be *ultimate* reality, this isn't the kind of scenario the word "real" developed in our language to deal with.[219]

Suppose instead that someone seeks to push us in the other direction, and insists that what is "real" simply *is* our linguistic representation of the world. They insist that there is no added "metaphysical" sense of "really" or "merely" in which something could be *really* real or *merely* a representation. They argue that our sentences are the result of shared experiences and could only possibly be about possible shared experiences – how else would they find their way into our language to begin with?[220] Such a person might suppose that he or she is offering a *deflationary* account of ontology. However, we can respond that there no way in which to express the account being's "deflationary" in the framework it offers. After all, if "reality" just means "appears to be" and "what we talk about in our language", then, well, reality really is as it seems to be to us, and as we speak about it – and, as a matter of fact, we speak about it as the sort of thing which sometimes isn't as it

---

[218] Berkeley (1713)

[219] See Neuber (2012) on Carnap's "empirical realism".

[220] See Carnap (1928)

seems and which our representations can occasionally get wrong. So, what appears at face value to be an anti-metaphysical account turns out to vindicate common sense metaphysics.[221]

Now, suppose that someone gets going too far on this new path, adopting wholesale the naïve realism of ordinary language. Chairs and birthday parties and loans and rabbits exist, and their existence is a "brute fact".[222] Unfortunately, the same applies to the feminine mystique, the white man's burden, the American dream, and the average family of 2.5 children. To resolve this, we ought to stop and recognize that it is also part of our ordinary way of speaking about the world that we defer to experts in the sciences, and permit them a greater burden in telling us what kinds of things there are. Perhaps they even tell us that we are all drastically mistaken (such as when we believe that solid objects contain no empty space). In ordinary language, we respect that chairs and loans and birthday parties should not be regarded as having the same metaphysical status as electrons, because we defer to external sources to help us determine what reality really is.[223]

We might let the sciences handle *all* of the work. We might choose to let the word "exists" include all and only those things which appear in our best scientific theories and are needed in order to explain phenomena. This might include quarks

---

[221] The methodology of Wittgenstein (1953) and Austin (1956) reflects this deflation of the deflationary account.

[222] Strawson (1959), Anscombe (1958)

[223] Putnam (1975), Burge (1979)

and gravity and electromagnetism and, with a deep sigh, mathematics.[224] But chairs and rabbits and birthday parties won't really exist, since they aren't necessary to explain how things appear to us.

Yet this seems odd. It is not as though chairs and loans and rabbits and birthday parties are supposed to be *fakes*. There is *something* there in the world which the words "chair" and "loan" and "rabbit" and "birthday party" are about, which they naturally latch on to. There are truths about them, which are true in virtue of the way the world is. But they aren't something *more* in the world, new things in addition to the physical stuff (the quarks and electrons and gravitational forces) and the psychological stuff (our thoughts about them).[225]

And so, at this stage in the dialectic, philosophy seems to be reaching for a compromise – a way to reconcile the permissive naïve ordinary ontology and the eliminative austere Occamist ontology. What we need are two ways to be "real" or "exist". One can apply to every object and property which we ordinarily make positive existence claims about: horses and apes are real, but unicorns and Big Foot are not. Another can be that of being *fundamentally* real, and it applies only to the very special properties and objects in virtue of which truths are true, the genuinely "brute" realities: quarks or strings might be real in this sense, but chemical compounds or mereological sums of particles – though only one step up from the fundamental level – are not.

---

[224] Quine (1992)

[225] Lewis (1984)

### 2.3 From Fundamentality to Cosmic Hermeneutics

This is how ontology came to have a new task. Not only does it tell us what exists, but now it also tells us which of the things which exist are fundamental or brute, and which of them exist only in a derivative way, as nothing "over and above" the more fundamental things. Assigned this new task, ontology will need some principled way to tell the difference between (i) the things which are (relatively)[226] fundamental, (ii) the things which are real but not fundamental, and (iii) the things which are not real at all.

There are some obvious cases. It is obvious that a pair of potatoes is nothing over and above the two potatoes taken by themselves. So long as you know all of the facts about the potatoes, you know all of the facts about the pair of potatoes. Nothing will surprise you about the pair if you know the potatoes taken alone. It should be just as obvious that a heap of rice is nothing over and above the rice in the heap. You probably can't know everything about the rice in the heap. But surely, if you knew everything about the rice and its spacio-temporal location and the relations in which it stood to everything else, you'd know everything about the heap. You could deduce it all, if you were smart enough and had the time. The heap is reducible ontologically to the rice, and the potato-pair to the potatoes.

What does this mean? Well, on the one hand, if you believe that potatoes and grains of rice are real, then you should also believe that pairs of potatoes and heaps of rice are real. Being committed to the existence of one commits you to the existence of the other, since the existence of one can be deduced from the other. On the other

[226] See Dunnaway (2013) for the motivation for relative fundamentality.

hand, it suggests that pairs and heaps aren't as close to the fundamental level as individual potatoes or grains. If you have the individuals, you get the groupings for free.

What makes the obvious cases of ontological reducibility so "obvious" is our sense that the facts about the higher-level domain would be easily *scrutable* given the facts about the basis they are being reduced to. When you have grains of rice in front of you, it's obvious you also have a heap – there's no mystery as to where the heap of rice came from.[227] Because it's so obvious, it's hard to see why anyone would want to insist that *ontologically* the heap is fundamental. The heap plays no added explanatory role. Recall that, as part of our compromise, we're allowing the demands of scientific explanation alone to tell us what is fundamentally real, in exchange for preserving non-fundamental realism for most of the ordinary things we talk about.

Sometimes there are less obvious cases. It is not so obvious that knowing all the facts about the subatomic particles in a grain of rice would give you the ability to know all of the facts about a grain of rice. Even the most intelligent human being likely couldn't deduce the facts about the grain from the facts about the subatomic particles. Nonetheless, it's uncommon to hold that the grain of rice is something over and above the subatomic particles.

In order to preserve the analogy with the "obvious" cases, we may appeal to an ideal reasoner – given infinite computing power and memory – whom we suppose

---

[227] There might be a conundrum as to *when* the heap comes into existence from the addition of grains of rice or when it goes out of existence from the subtraction of grains of rice, but not a mystery.

in principle could deduce the facts about the grain of rice from the facts about subatomic particles. The subatomic particles are fundamental, but the grain is not, we say, because the grain facts would be *a priori* scrutable for an ideal reasoner given the subatomic particle facts. It's not as though we actually consult with any mathematical archangels when making this claim. Rather, the claim is meant to illustrate the ontological reducibility of the grain to the particles, by helping us imagine a mind for which it would be obvious that when you have the particles you have the grain, much as it is obvious for us that when you have the two individual potatoes you have the pair.

Let's sum up our story so far. Epistemology was troubled by conflicting tendencies to be skeptical about reality and to take everything at face value. In metaphysics, this translated into a tension between austere and permissive accounts of reality. A compromise intended to resolve this tension allows there to be many real things, but some especially real, or fundamental – the minimally necessary set needed for a scientific explanation of the world – and others whose reality is derived from or grounded in those fundamental things. In some obvious cases, whenever one phenomenon is grounded in a more fundamental phenomenon, the facts about it are easily and obviously scrutable from the facts about the fundamental phenomenon. We came to expect the same to apply in non-obvious cases – if one phenomenon is grounded in a more fundamental phenomenon, the facts about it ought to be *a priori* scrutable from the facts about the fundamental phenomenon *by an ideal reasoner*, even if we can't do so ourselves. A failure of ideal deducibility then became an indicator that either something was fundamental or else that it wasn't something we ought to regard as real at all. The ideal reasoner became the means of arbitrating

the real and reducible on the one hand, and the unreal or irreducibly real on the other hand.

Thus, the new task of ontology becomes identified with trying to show how most of the "folk" properties and objects referred to in ordinary language are the sorts of things which an ideal reasoner would be able to deduce given as small a set of concepts and facts as possible. This can be a revisionary project: perhaps some of our ordinary folk theories are bad theories and should be thrown out. However, if we are to be charitable to ourselves, the hope is that most of our sentences will be translatable into the language of the fundamental. This is how the task of ontology in locating the fundamental properties comes to be identified with the task of cosmic hermeneutics.

## 2.4 Forms of Scrutability

Philosophers have thought of Cosmic Hermeneutics in several different ways. There are at least three questions which have to be answered when considering whether or not some truth would be *scrutable* by an ideal reasoner, given the fundamental truths. I use the term "scrutable" as synonymous with "understandable" or "knowable" on some basis, including but not limited to that which is "deducible".[228] Combining the answers to these questions in various ways gives us a variety of types of scrutability.

---

[228] I'm borrowing 'scrutable' from Chalmers (2012), insofar as some people may be uncomfortable with calling some of these forms of scrutability 'deductions', which perhaps suggests that one could construct a finite derivation, that the premises must be relevant to the conclusion, that the premises must logically guarantee the

**2.4.1 Base.** First, what is the content of the scrutability base? The ideal reasoner is given some starting point *B*, perhaps a domain of facts[229]. This domain might be taken to be the "fundamental" facts, though it need not be. For example, *A* might be scrutable given the fundamental facts, but not the fundamental physical facts taken alone; *A* might be scrutable given any basis, including an empty basis (for instance, if *A* is itself *a priori*); *A* might be scrutable from *B* without *B* being fundamental; *A* might also be locally scrutable given some subset of the physical facts, or it might only be globally scrutable given the totality of physical facts, and so on.

**2.4.2 Form.** Second, what form of representation appears in the scrutability base for the ideal reasoner (e.g., the premises of the deduction), and what form of representation is the ideal reasoner supposed to output (e.g., the conclusion of the deduction)? Some options:

*(a) Sentential:* representations in natural or idealized language,

*(b) Propositional:* conceptual representations, sets of worlds[230]

*(c) Ontological:* real properties, truthmakers, natures of things.[231]

---

conclusion rather than merely making it epistemically certain or a matter of warranted belief, and so on. Perhaps there are non-deductive forms of scrutability.

[229] I'm using "fact" here as ambiguous between "true sentence", "true proposition", and "truthmaker", depending upon the answer to 2.4.2

[230] Whether one is a Russelian or a Fregean about propositional content, or takes a possible-worlds view of content, will obviously produce different interpretations of a scrutability claim.

**2.4.3 Resources.** What sorts of rules of inference or logical transitions is the ideal reasoner allowed to apply to the premises in order to reach the conclusion? We might think of this as asking about which cognitive "resources" are to be provided to the ideal reasoner, when the ideal reasoner is attempting to "scrute" *A* from its scrutability base *B*. Below, I list a number of options for what forms of scrutability might hold between output *A* and base *B*. Chalmers (2012) distinguishes varieties of scrutability by the relation between a sentence *S*, scrutability conditions *C*, and a subject *s*. Instead, here I have distinguished a broader set of varieties here in terms of which inferences or cognitive abilities the ideal reasoner is permitted to use to understand one thing on the basis of another.

I have divided the sorts of resources an ideal reasoner might be given into three classes, with each consisting of 3-4 distinct categories within that class. This list of resources is supposed to be progressively less restrictive, with categories further down the list including all resources available higher on the list in addition to some new type of inference pattern or reasoning capacity.

*Class A: Analytic A priori Resources Only (a) – (d)*

*(a) Minimal logical scrutability* holds iff *A* is scrutable from *B* when only classical first-order logic (with identity) is permitted as a rule of inference.[232]

---

[231] Chalmers (2012), 73-91 distinguishes only Sentential and Propositional scrutability. However, an ideal reasoner might be imagined as having access not only to our representations but to the "language of ontology".

[232] Or, if some other logic is the correct one, the rules of that logic alone will be permitted. If pluralism about logic is true, and there is no fact of the matter about

*(b) Definitional scrutability* holds iff $A$ is scrutable from $B$ when only (a) and definitions of words in terms of necessary and sufficient conditions, or translations of synonymous terms between languages, are permitted as rules of inference. Chalmers identifies definitional scrutability with the project of Carnap in the *Aufbau*.[233]

*(c) Mechanical-Analytic scrutability* holds iff $A$ is scrutable from $B$ when only (b) and all analytic *a priori* inferences which involve only the mechanical application of a rule, independent of content or meaning, are permitted.

*(d) Conceptual-Analytic scrutability* holds iff $A$ is scrutable from $B$ when only (c) and intuitions about the application of a concept to various possible scenarios are permitted.[234]

---

whether classical or non-classical logic is correct, then there will be various types of minimal logical scrutability.

[233] Chalmers (2012), 6-12 notes a number of problems with definitional scrutability which have been raised since Carnap by Quine, Wittgenstein, Kripke, and others. While these are a problem for definitional scrutability, they need not be a problem for analytic scrutability more broadly conceived.

[234] Chalmers (2012), 388-392 does not distinguish between (c) or (d) as varieties of analytic scrutability. However, it seems to me that the contemporary philosophical practice of "conceptual analysis", with its heavy reliance on intuitions about concepts, can be meaningfully distinguished from those forms of the analytic *a priori* which does not rely on intuitions about meaning. I take (c) to be the more basic concept of the analytic – i.e., a mechanical procedure; intuitions about concepts are

*Class B: Analytic and Synthetic A priori Resources (e) – (g)*

(e) *Mathematical scrutability* holds iff $A$ is scrutable from $B$ when only (d) and inferences which utilize the axioms of set theory are permitted.

(f) *Simulative scrutability* holds iff $A$ is scrutable from $B$ when the ideal reasoner is permitted only (e) and whichever inferences can be justified by means of conducting a mental simulation of some aspect of the world. For example, charitable reconstructions of a speaker's intentions are permitted as a means of inferring the meanings of their terms. Likewise, imagining what would happen, or what patterns would appear, given a system following certain rules or procedures, counts as justification by simulation.

(g) *A priori scrutability* holds iff $A$ is scrutable from $B$ when all inferences which could be justified *a priori* are permitted, including (f). Thus, a sentence $S$ is inferentially scrutable from conditions $C$ for subject $s$ iff $s$ is in a position to know *a priori* that if $C$, then $s$.

*Class C: A Posteriori and A priori Resources (h) – (j)*

(h) *Inferential scrutability* holds iff $A$ is scrutable from $B$ when all inferences which justify belief in $A$ on the basis of *knowing* $B$ are permitted, both *a priori* and *a posteriori*, even if these do not guarantee the truth of $A$.[235] Thus, a

---

regarded as analytic in virtue of the assumption that they inform us about meanings, and when the meanings of terms are sully specified translations can be performed mechanically.

[235] See Chalmers (2012), 47-52. S is conditionally scrutable from C for s when s is in a position to know given C that S is true.

sentence *S* is inferentially scrutable from conditions *C* for subject *s* iff, if *s* were to come to know *C*, *s* would be in a position to know *S*.[236]

(i) *Conditional scrutability* holds iff *A* is scrutable from *B* when all inferences which justify belief are permitted, both *a priori* and *a posteriori*, on the *assumption* that *B* is the case, even if one is not justified in believing *B*. Thus, a sentence *S* is conditionally scrutable from conditions *C* for subject *s* iff *s* is in a position to know that *if C*, then *s*.[237]

(j) *Entailment scrutability* holds iff *A* is scrutable from *B* when all necessary truths are known by the ideal reasoner and available for use in inferences, including those which can only be justified *a posteriori*.

Which of these varieties of ideal scrutability should we be concerned with, if our interest in Cosmic Hermeneutics is determining the fundamental ontological structure of the world? In the remainder of Part A, I will consider three approaches. On the approach of *Radical Interpretation*, which I will discuss in Section 3, Cosmic Hermeneutics is concerned with sentential simulative scrutability. On the approach of *Conceptual Analysis*, Which I will discuss in Section 4, Cosmic Hermeneutics is concerned with propositional, conceptual-analytic scrutability. On the *Cosmoscope* approach, which I will discuss in section 5, Cosmic Hermeneutics is concerned with sentential, *a priori* scrutability. In Part B, I will advocate for an alternative to all three approaches.

---

[236] ibid., 40

[237] ibid.

## §3 Cosmic Hermeneutics as Radical Interpretation

### 3.1 Radical Translation to Radical Interpretation

Normally, the task of translating from one language into another is aided by prior linguistic knowledge. One does not generally try to translate 19th century German philosophy into English without first knowing ordinary German. A knowledge of related Slavic languages may aid a translation into Macedonian. Quine (1960) invoked the thought experiment of "radical translation" to see what a pure translation unaided by prior linguistic knowledge would look like. In radical translation, the project is to translate sentences[238] from a language one knows nothing about (*jungle*) into one's own language, given no intermediary, no common tongue, no dictionary, but only observations of various behaviors. This was meant to illustrate how limited the information one could obtain about meaning from superficial observations of behavior was, and so was meant to lead to skepticism about any sort of analytic "meaning" beyond what could be obtained empirically.

Donald Davidson modified Quine's thought experiment in a way which avoids the skepticism about meaning. Davidson replaces the notion of translation with that of an *interpretation* – our task is not merely to find synonymous sentences in one language for sentences in another, but to understand what the words in the language *mean*. In ordinary interpretation, we know ourselves to be speaking the same language as others. We can reflect on what we mean by a term and then

[238] The emphasis on holistic sentences as opposed to words is important, and perhaps one reason Quine's view of language is seen as in conflict with contemporary generative linguistics.

presume, in the absence of evidence otherwise, that others mean by their terms what we would mean by them. When others speak in a way which is amiguous between multiple interpretations, we assign an interpretation to their speech based on assumptions about their beliefs, goals, and intentions in the context of our conversation. However, in radical interpretation, we know neither the meanings of the other speaker's utterances (since we do not speak the language ourselves) nor the intentions or beliefs of the speaker in question (which would be necessary to form hypotheses about what they might mean). We must simultaneously form a theory about the meaning of the speaker's utterances in order to interpret their psychology, and a theory about the speaker's psychology in order to interpret their utterances. How would we ever get started?

For Davidson, the starting point is a *principle of charity*:

> This is accomplished by assigning truth conditions to alien sentences that make native speakers right as often as plausibly possible, according, of course, to our own view of what is right. What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. Applied to language, this principle reads: the more sentences we conspire to accept or reject (whether or not through a medium of interpretation), the better we understand the rest, whether or not we agree about them. The methodological advice to interpret in a way that optimizes agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances and other behavior of a creature

> as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything.[239]

David Lewis (1974) made the connection between Davidson's project of Radical Interpretation and the older tradition of Cosmic Hermeneutics from Laplace and Broad. Whereas Davidson's interpreter is supposed to be relatively like a human being, and thus not radically different from the speaker, despite the lack of a shared culture, Lewis conceives of the project of Radical Interpretation as performed by an ideal reasoner like Laplace's demon. Davidson's project is intended to provide a theory of how we might come to interpret natural languages through observations of behavior, and thus to support a coherence theory of truth and knowledge. On the other hand, Lewis's Radical Interpretation is conceived of as giving the ideal reasoner access to some set of fundamental facts and no further facts, such as the facts of the speaker and his or her material parts as a physical system.[240] Lewis is not trying to figure out how *we* perform the real-life task of determining what other people mean or believe, but rather asking "how do *the facts* determine the facts?"[241] How is it that the physical facts determine the psychological facts and the semantic

---

[239] Davidson, "Radical Interpretation", 324

[240] Lewis (1974), 331; although Lewis recognizes the same thought experiment can be performed if taking as fundamental any non-physical "psionic fields, astral bodies, entelechies, or what-not".

[241] ibid., 334

facts? Lewis assigns the ideal reasoner the task of getting to know a person, Karl, as an instance of the sorts of psychological and semantic facts which are often thought to avoid reduction to the physical – his goal is Physical-to-Phenomenal Cosmic Hermeneutics.

## 3.2 Interpretive Principles

The ideal reasoner wishes to know the facts about both Karl as a speaker, in terms of the meanings of the utterances in his language **M**, and Karl as a person, in terms of his attitudes (his beliefs and desires). The ideal reasoner starts knowing nothing about either, but has a complete physical description **P** of Karl and the cosmos in which he lives (which we assume his meanings and his psychology both supervene upon)[242]. We can distinguish Karl's attitudes as expressed in Karl's own language, **Ak**, as well as his attitudes as expressed in our language, **Ao**. The ideal reasoner is thus constrained by a few principles:

**3.2.1 Principle of charity.** On Lewis's modification of Davidson's principle, the interpretation of **Ao** should ascribe to Karl beliefs and desires which are not so drastically removed from our own that, if we were to give Karl our own life histories

[242] Lewis seems to have in mind only "Karl as a physical system", and thus seems to require *local* supervenience. However, Lewis's thought experiment need not conflict with externalism about semantic content, and so instead we should understand the whole physical cosmos as the *global* supervenience base for both Karl's meanings and Karl's beliefs. Local supervenience entails global supervenience, but only local supervenience conflicts with externalism.

and give ourselves the entire life history which Karl has given in **P**, we would end up with **Ao** and Karl would end up with attitudes more or less like our own.

**3.2.2 Rationalization Principle.** Much as the principle of charity constraints **Ao** by Karl's past life history, so the rationalization principle constrains **Ao** by the prediction of Karl's future behavior in **P**. The beliefs and desires ascribed to Karl should provide good reasons for his behavior, generally speaking – he should be presumed to be a rational agent following something much like rational decision theory.

**3.2.3 Principle of Truthfulness.** When **M** assigns a truth condition to the meaning of Karl's language, then **Ao** should ascribe to Karl beliefs and desires which are consistent with a convention of truthfulness. If 'Ionlay' is true when a Lion is present, for example, then Karl should generally believe that a lion is present when he hears others utter 'Ionlay'.

**3.2.4 Principle of Generativity.** The truth conditions assigned by **M** to the meaning of Karl's language should be as uniform and simple as possible. They should be consistent with our ability to generate or construct novel sentences out of old words in rule-governed ways.

**3.2.5 Manifestation Principle.** Karl's attitudes in **Ak** should manifest themselves publically in his dispositions to speech behavior in **P**, except where in **Ao** we prescribe him as having some special reason for secrecy.

**3.2.6 The Triangle Principle.** Given the meaning of Karl's language in **M**, it should be that **Ao** = **Ak**: the beliefs and desires we assign him in our language should translate into the ones which, given the meaning of his language, we assign in his language.

## 3.3 Interpretive Methods

Given these principles, Lewis sees three methods by which the ideal reasoner might successfully interpret Karl's language. Each of these methods might be iterative – the reasoner might go through the steps many times.

**3.3.1 Method One - Davidson's Method.** Use the principle of charity to develop **Ao** from **P.** Use the Manifestation principle to develop **Ak** from **P.** Revise the beliefs in **Ao** and fill in **M** according to the Triangle Principle, so as to balance the degree of match between the beliefs in **Ao** and the beliefs in **Ak** with the demands of the Principle of Generativity on **M**. Revise the desires in **Ao** based on the Rationalization Principle and **P**, and then fill in the desires in **Ak** to match the desires in **Ao** according to the triangulation principle. Repeat until a stable solution is reached.

**3.3.2 Method Two - Lewis's Preferred Method.** Start by determining **Ao** from **P** by means of the Rationalization Principle and the Principle of Charity. Fill in **M** from **Ao** based on the Principle of Truthfulness and the Principle of Generativity**.** Finally, given **Ao** and **M**, fill in **Ak** by means of the Triangle Principle.

**3.3.3 Method Three –Quinean Holistic Non-Method.** Try to fill in **Ao**, **Ak**, and **M** all at once, balancing the various principles as best as possible. Because we have no guarantee of determinate answers to the meanings or attitudes of Karl given the principles assigned above, we can't operate in stages, and the best we can do is use the theory we are building of Karl to continue building our theory of Karl.

**3.3.4 A Burgean Variation.** A variation on Lewis's methods might be obtained by substituting the principle of charity with Principle (B) from Burge's (2010) *Origins of Objectivity*. Charity requires that the interpreter assume the

speaker is every bit as rational as the interpreter would have been if the interpreter had the speaker's life history. However, it is not implausible that humans, by and large, are far less rational than an ideal reasoner given an identical life history and biological composition would be. Instead, Burge's principle requires that "for an individual to have any representational state (such as a belief or perception) as of a subject matter, that state must be associated with some *veridical* representational states that bear referential, indicational, and attributional representational relations to a suitably related subject matter."[243] To simplify somewhat, in order to attribute to a subject the mental state of a *false belief that p*, we must assume that there are nearby[244] mental states, subjects, beliefs, or contents *p'* closely related to *p* for which there exist true beliefs that *p'*. Such a principle might be used in place of the principle of charity to obtain a similar simulation of a subject's beliefs and a similar interpretation of a subject's language, but without the need for optimism about the rationality of human beings.

**3.3.5 Ebbs's Variation.** Ebbs (2009) proposes what might be taken as a fourth method for radical interpretation. For Ebbs, the project of "radical interpretation" is not all that different from our everyday interpretation of what others are saying. Ordinarily, we are entitled to take for granted in the absence of

---

[243] Burge (2010), 68

[244] "Nearby" might be understood metaphysically, as involving the nearness of other worlds, or it might be understood in terms of some other suitable causal or constitutive relation; it should go without saying that "nearby" is not understood here geographically.

other evidence that other speakers of our language are using words with the same meanings as we have when we use them, *for the same reasons* that we are entitled to take for granted that our present selves are using words with the same meanings we gave them in the past. Ebbs introduces the notion of an *intersubjectivity constraint*: our theory of the truth conditions of terms in **M** must be accompanied by an account of why it is epistemically reasonable for one to apply **M** to other speakers' sentences and to one's own sentences as one used them in the past. On Ebbs's account, the only way in which the ideal reasoner could come to know the meaning of Karl's sentences would be *to become a speaker of Karl's language* and begin to participate (at least mentally) in Karl's linguistic community – the Mathematical Archangel couldn't stand apart and interpret from above.

### 3.4 Radical Interpretation as Simulative Scrutability

Notably, all three methods – if we understand them to tell us about how "the facts determine the facts" and not merely how the process of translation works – involve the ideal reasoner attempting a kind of mental *simulation* of Karl's psychology. All three methods make heavy use of the Principle of Charity and the Rationalization Principle, which require the ideal reasoner to consider what he would consider rational, reasonable, and good where he in Karl's shoes and to explain Karl's behavior in terms of reasons for acting. All three methods depend upon the "simulated" **Ao** – the ideal reasoner's theory of Karl's action in the interpreter's own language – in order to derive Karl's **M** and **Ak**, with the dependence being direct in Lewis's Method Two. Ebbs's approach would be the most dramatic example where interpretation requires simulation; the Burge-esque approach would be a more moderate example.

Since it requires the interpreter to simulate in some way the speaker's internal mental states, the Radical Interpretation model is best categorized as involving the *simulative scrutability* of the semantic and psychological facts given the physical facts. Insofar as once accepts Lewis's principles of interpretation as *a priori*, this simulation of Karl's beliefs also qualifies as *a priori* scrutability. However, it certainly does not qualify as either a kind of mechanical-analytic or conceptual-analytic *a priori*. Getting to Karl's beliefs via simulation is very different from getting to Karl's beliefs via analysis.

What's the ontological upshot of Radical Interpretation, then, whichever of these four methods guide it? On the one hand, Radical Interpretation tells us that if Cosmic Hermeneutics from the physical facts to the psychological and semantic facts is possible with simulative scrutability, then the psychological and semantic facts, insofar as they are determinate,[245] are exclusively determined by the physical facts. On the other hand, if Radical Interpretation suggests that Cosmic Hermeneutics is *only* possible by means of simulation, and is not possible by means of analysis, then this provides some motivation for thinking that semantics and psychology are at least weakly emergent.

Notably, Lewis does not expressly discuss the radical interpretation of Karl's *phenomenology* – if Karl could have the same beliefs and desires yet a different phenomenology, then a simulation which determines Karl's beliefs and desires will still underdetermine what it is like for Karl to believe and desire what he does.

---

[245] e.g., Davidson acknowledged a degree of indeterminacy for radical interpretation.

### 3.5 Interpretive Constraints as Meaning Constraints?

I have categorized the project of radical interpretation through interpretive constraints as involving *simulative scrutability*, understanding a simulation to be a procedure which is *a priori* but not *analytic*.[246] However, Horgan (1984) made the case that the "interpretive constraints" in Lewis's project should instead be understood to involve only "meaning constraints" – that is, to involve only conceptual-analytic scrutability. Horgan holds that if Cosmic Hermeneutics were possible through "meaning constraints" alone, then that should be sufficient to demonstrate the truth of physicalism. Thus, Horgan interprets Lewis's radical interpretation as making a case for non-reductive physicalism.[247]

The notion of a "meaning constraint" reflects a looser notion of meaning than the pre-Wittgensteinian, pre-Quinean understanding of meanings as "definitions", but it still connotes a sort of analyticity. A "meaning constraint" is what helps us adjudicate cases in which a familiar word is used by a speaker in such a peculiar way that we determine the word simply doesn't mean in the speaker's mouth what it means in our own. Supervenience is not a meaning constraint. Consider that,

---

[246] The process of simulation is "empirical" in the broad sense – it is the sort of thing one must perform, and observe the results, and not merely reason through – and yet the beliefs one forms on the basis of a simulation are justified *a priori* for the same reason that derivations from conditional proofs are a justified *a priori*: one sees what necessarily follows from a given assumption, and concludes the conditional.

[247] On the assumption that there is no other way to derive conclusions about mental states, except through simulation.

although Moorean non-naturalists in Meta-Ethics accept that ethical truths supervene upon the physical truths, they would deny that the *meanings* of ethical language are in some way dependent upon the meanings of physical terms. On the other hand, Horgan held that Lewis's interpretive principles *are* meaning constraints – particularly, the principles of Charity and Rationalization are supposed to be entailed by our concept of a *person*.[248]

It is not clear to me why either principle must be a constitutive part of our concept of a person, or why our concept of a person must be an constitutive part of either the concept of meaning or the concept of an attitude like a belief or a desire; perhaps one could hold that there are sub-personal beliefs or desires, or that some non-human animals can be assigned rationality or meaningful units of thought without meeting other criteria of persons. However, putting this aside, even if it is analytic from the concept of belief or of meaning that beliefs and meanings are the sorts of things which are to be interpreted by means of Lewis's constraints, the process of interpreting Karl by means of Lewis's constraints is not thereby made analytic or mere application of a concept. Intuiting from introspection that I *would have to hold certain beliefs* in order to rationally exhibit particular behaviors in given circumstances is every bit as much a candidate for the synthetic *a priori* as intuiting that I would *have a reason* to exhibit particular behaviors in given circumstances. Cosmic Hermeneutics still involves a kind of simulation.

I will turn now to a method by which Cosmic Hermeneutics *is* supposed to be possible by means of conceptual analysis alone. Whereas the subject matter for

[248] Horgan (1984), 34

Radical Interpretation took the form of sentences, the subject matter for conceptual analysis involves propositions and the categories of everyday thought.

## §4 Cosmic Hermeneutics by Conceptual Analysis

### 4.1 The Plan

Frank Jackson's *From Metaphysics to Ethics* (1998) sets up the project of "serious metaphysics" as involving a particular method of Cosmic Hermeneutics. This method is concerned with the *propositional conceptual-analytic scrutability* of ordinary truths from a limited, compact base of fundamental ingredients. It occurs in four stages.

**4.1.1 Stage One: Determining Application of Concepts.** In the first stage, the ideal reasoner conducts an analysis of the various concepts which occur in ordinary language and thought. For every concept helped by speakers of English, or Spanish, or American Sign Language – including concepts which have no linguistic expression whatsoever – the ideal reasoner attempts to determine, for all of the possible scenarios and entities in the cosmos, which entities are included under each concept. This process is supposed to be entirely *a priori*.[249]

---

[249] Jackson (1998), 46-52. Notably, this is a much stronger notion of the *a priori* than those inclined to methodological naturalism in philosophy will admit. Devitt (2005) and others have expressed a concern that one cannot be a naturalist or a materialist and accept that Jackson's form of *a priori* conceptual analysis tells us anything significant about reality. However, although I am not a materialist, I do not share this concern. In contemporary philosophy, the *a priori* is understood to be a type of non-empirical *justification* for belief which does not depend upon sensory

For example, suppose that T is a sentence in folk psychology, like "John believes what he sees." The proposition expressed by T contains the concept of *belief* and the concept of *seeing*. The ideal reasoner considers the concepts of *believing* and of *seeing* in terms of various ways the world might be such that the concepts "believes" or "sees" would apply there to various things. The ideal reasoner would identify various descriptions as cases of believing and others not as cases of believing. Notably, the ideal reasoner need not be able to define "belief" in order to do this.[250]

---

experiences like perception, memory, or testimony – claiming that a belief is justified *a priori* is not a claim about the origins, history, or source of the justified beliefs. There is no need for *anamnesis*. A materialistically-acceptable history of how I came to possess various *a priori* beliefs might be told, such as a life history of how I came to learn a concept through social interactions with others who spoke my language, or an evolutionary history involving the death of my ancestors who failed to grasp logical, moral, or mathematical truths. A history of a belief is not the same as a justification for a beliefs.

[250] "Once an essential role for explicit definitions is eschewed, the model of conceptual analysis that emerges is something like the following. When given sufficient information about a hypothetical scenario, subjects are frequently in a position to identify the extension of a given concept, on reflection, under the hypothesis that the scenario in question obtains. Analysis of a concept proceeds at least in part through consideration of a concept's extension within hypothetical scenarios, and noting regularities that emerge. This sort of analysis can reveal that

Most ordinary speakers have some level of conceptual competence like this, not only for highly concrete concepts but for rather abstract ones: given a description of a scenario and asked whether or not it is a case in which "Smith *knows* that Brown is in Barcelona" or "Al's disease is not *arthritis*", most people will find something to say much of the time. There are, of course, borderline cases and unclear cases. Running through *every* possible scenario would require infinitely more conceptual competence than any human has! However, the ideal reasoner can be assumed to have complete competence when it comes to the relevant concepts. So, the ideal reasoner might, given his infinite capacities, run through every single possible world, in order to determine every application of the concept *believed* or *seen* or any other concept in that world.[251]

**4.1.2 Stage Two. Determining the Functional Role.** Given this intuitive knowledge of the application conditions of various concepts, the ideal reasoner will then attempt to relate the various concepts to one another, in terms of the characteristic role they play vis-à-vis the other concepts. This will produce a list of all of the conceptual truths for each concept. For instance, beliefs are the sorts of things which are 'characteristically caused by perceptions', 'combine with desires to generate actions', and 'are a necessary condition of knowledge'.[252] These conceptual truths are *a priori* – one does not have to look at whether there are actually beliefs

---

certain features of the world are highly relevant to determining the extension of a concept, and that other features are irrelevant." Chalmers and Jackson (2001), §3

[251] Jackson (1998), 31-37

[252] Papineau (2009)

in a world and what sorts of things people there believe in order to determine these facts about beliefs. The ideal reasoner develops an exhaustive list of the conceptual truths for each concept. This list is a specification of the functional role the concept plays for those who use it.

Note that the functional roles in these cases need not involve an explicit analysis.[253] They might only include only characteristic roles for the application of each concept, not necessary and sufficient conditions.[254] These functional roles are not *definitions*. [255]

The only concepts for which the ideal reasoner is unable to determine a functional role will be those which can't be given a role in terms of other concepts. These *primitive concepts* will be the "building blocks" out of which all of the others are constructed, the "semantic primes" such that every concept in every human language can be given a functional role in terms of them.[256]

---

[253] Chalmers and Jackson (2001)

[254] One supposition at this stage is that a concept which can be given some functional analysis by the ideal reasoner can be given a complete functional analysis. However, even if we eschew talk of "definitions" or "necessary and sufficient conditions", it still seems possible that the ideal reasoner will only know *a priori* some of what is characteristic of some concept or part of the role that it plays.

[255] Chalmers (2012), 320

[256] Chalmers (2012), 2 cites Wierzbicka (2009) as claiming that all expressions in all human languages can be analyzed in terms of 63 'semantic primes'.

**4.1.3 Stage Three: Ramsification.** The ideal reasoner then takes his list of all of the functional roles of each given concept, and quantifies over them in order to form a Ramsey sentence.[257] "Beliefs are characteristically caused by perceptions and combine with desires to generate actions and . . . " would translate into "There is some x such that x is characteristically caused by perceptions and x combines with desires to generate actions and . . .". The same process would be repeated for every occurrence of other concepts within this concept of *actions*, *perceptions*, *desires*, and so on. The result will be a sentence which begins with an existential quantifier and contains only primitive concepts, logical operators, and variables. The ideal reasoner will continue until he has obtained Ramsey sentences for all of the non-primitive concepts which make use of the primitive concepts alone.[258]

**4.1.4 Stage Four: Hunting.** The ideal reasoner now turns to the scrutability base he has been given. For Jackson (1998), the scrutability base is the set of fundamental physical facts – Jackson's interest is in physical-to-phenomenal cosmic hermeneutics. For each of the sentences obtained at Stage three, the ideal reasoner now hunts around in the domain of physics to find which physical properties (or relations, mereological sums of objects, kinds, etc.) satisfy these sentences. For example, the ideal reasoner discovers a bundle of physical properties which is "characteristically caused by perceptions and combines with desires to generate actions . . ." (where 'perception', 'desire', 'action', etc. are all replaced with their

[257] Papineau (2009)

[258] Jackson (1998), 140-145

Ramsified translation), and the ideal reasoner is now able to identify this physical bundle as what a "belief" is.

It may be that no bundle of physical properties perfectly satisfies the concept's functional role, but that there is a bundle of physical properties which *mostly* satisfies the concept's functional role and which stands out distinctly from the others. In this case, the bundle should be still identified as what the concept refers to. This means the process can be revisionary. It may turn out that out that many of our folk concepts are a bit off and need to be corrected by further theorizing.[259] It may turn out that some of our ordinary concepts stand for a disjunction of two or more natural bundles of physical properties. It may turn out that some of our ordinary concepts have conditional intensions – they pick out one bundle of properties if the world is one way and a different bundle of properties if the world is another way. And, of course, it may turn out that some of our ordinary concepts are not even close to being filled by anything in the physical world, in which case our beliefs and thoughts about them are false.

If the scrutability base contains all and only the fundamental facts, then the ideal reasoner can come to know all of the true propositions expressed in human thought and language through these four stages. This process required only *a priori* reflection on the meaning of the non-primitive concepts and a knowledge of the scrutability base. The things humans talk about which are deducible from the

[259] Chalmers (2011)

fundamental facts will enter by entailment[260], and the things humans talk about which are not so deducible will be eliminated.

Suppose there are some primitive, unanalyzable concepts left over after Step 3, but these concepts do not occur in the scrutability base. We are supposed to conclude from this that these concepts are either mistaken, or that they refer to fundamental properties which were not included in the scrutability base. For instance, if the scrutability base consists only of physical facts, and there are primitive concepts which do not occur in physics, then either the primitive concept is a mistaken one (and should be eliminated) or else some non-physical facts must be fundamental which match up with the primitive concept.

Suppose there are some non-primitive concepts which we have a conceptual analysis for, but nothing in fundamental physics comes close to satisfying the role played by these concepts. If this is so, we must either choose to hold that nothing satisfies propositions with those concepts in our world, or else we must hold that there is something non-physical but fundamental which satisfies these propositions.

Jackson holds that Physical-to-Phenomenal Cosmic Hermeneutics is possible by this method; because of this, phenomenal properties are metaphysically nothing over and above fundamental physical properties. Chalmers (1996) holds that Physical-to-Phenomenal Cosmic Hermeneutics is not possible along this method, because we can clearly conceive of physical duplicates of ourselves with absent or inverted phenomenal experiences. He concludes from this that phenomenal properties must be metaphysically fundamental.

---

[260] Jackson (1998), 25-27

## 4.2 Challenges to Conceptual Analysis

Ned Block and Robert Stalnaker (1998) challenge the conclusion that the lack of a conceptual analysis for concepts of consciousness in physical terms entails that consciousness cannot be explained in physical terms – the so-called "explanatory gap", which suggests the "metaphysical gap" between conscious properties and physical ones. One might put their objection in this way: while *conceptual-analytic scrutability* might fail for *A* from *B*, it might nonetheless be that *A* is *inferentially scrutable* from *B* by the ideal reasoner. For Block and Stalnaker, it is inferential scrutability or its failure, rather than conceptual-analytic scrutability or its failure, which might entitle us to come to metaphysical conclusions. Contrast the following two claims:

> (*Inferential Scrutability from a Physical Base*) Were one to know all of the physical facts about the world, one would be in a position to know all of the facts.
>
> (*Conceptual-Analytic Scrutability from a Physical Base*) Were one to know all of the physical facts about the world, one would be able to deduce all of the facts *a priori* by means of conceptual analysis alone.

Block and Stalnaker argue that for ordinary macroscopic truths about natural kind terms, the facts about things like water and tigers and rivers and mountains, conceptual-analytic scrutability fails. But ordinary macroscopic truths are truths which no one doubts are grounded in the fundamental physical facts, and

which no one questions are fully explainable and knowable given the physical facts. If this is so, then there's nothing special about the failure of conceptual-analytic scrutability in the case of consciousness, and we can't infer from this failure that consciousness is explanatorily or metaphysically irreducible to physics.

Consider the well-known example from Kripke (1980) and Putnam (1973) of water and $H_2O$. Water is identical to $H_2O$ – they are one and the same thing. This is a metaphysically necessary identity. However, the concept of "water" is not identical to the concept of two hydrogen atoms bonded to an oxygen atom; at least, it certainly wasn't the same concept in the era before water's chemical composition was known. Nothing about the concept of water – that it is the sort of thing which is in the lakes and streams, which refreshes thirst, and falls from the sky when it rains – suggests that water is $H_2O$.

It is true that, given all of the physical facts in the world, we (and the ideal reasoner) are justified in believing that Water = $H_2O$. We are in a position to know that Water = $H_2O$, given only our concept of water and the physical facts. However, our justification cannot be *a priori* in this case. Why not? For one thing, consider that we can't rule out from any of the information we've been given the strange and spooky scenario in which, for every molecule of $H_2O$, there is a unit of non-physical 'ghost' water located nearby, so that the term "water" in fact refers to both $H_2O$ and to ghost water. We can rule this out empirically: we have no evidence for ghost water, and our evidence is that $H_2O$ is enough to fill the role supplied by water. So, we can know confidently, given only our limited physical description of the world, that Water = $H_2O$. Perhaps the ideal reasoner could be even *more* certain than we

are. But neither we nor the ideal reasoner can rule out this scenario simply by reflecting our concept of "water".

Block and Stalnaker further suggest that we consider the example of "life". How did scientists come to conclude that vitalism was false, and to close the gap between the physical and the biological sciences? What they *didn't* engage in was an analysis of the concept of 'life'. Consider the functional roles which 'living thing' plays: something like 'being the sort of thing which reproduces, digests, and respires'. These functional roles relate life to other concepts in the same biological family of terms, not to anything in microphysics.[261] Instead, what scientists did involved picking out a few paradigm cases of life, came to understand how life worked physically in simpler cases, and extended the explanation upward from the simpler cases to the complex paradigm cases.

Presuming that one does not want to hold that 'water' or 'life' or other ordinary macroscopic truths are ontologically irreducible to the microphysical, we can't infer from their lack of conceptual-analytic scrutability that there is either an explanatory gap or a metaphysical gap between them and microphysics. If this is so,

---

[261] Block and Stalnaker also express concern that these functional are neither sufficient nor necessary conditions for life – nothing in the concept of life rules out non-reproducing beings; a moving van also excretes wastes. However, as noted earlier, nothing in the conceptual-analysis model of Jackson and Chalmers requires that one find an *definition* of a concept in terms of necessary and sufficient conditions.

we must follow the same policy in the case of consciousness – unless someone offers an argument for why we should regard consciousness differently.

### 4.3 Totality, Identity, and 2-Dimensional Semantics

Chalmers and Jackson (2001) offer a number of responses to the objections raised in Block and Stalnaker's article against the conceptual-analytic scrutability of ordinary macroscopic truths. I will focus on three of them.

First, in response to the worry about the epistemic possibility of 'ghost water' preventing the *a priori* deduction that water = $H_2O$, Jackson and Chalmers acknowledge that part of the ideal reasoner's scrutability base must be *T*, the so-called "totality condition". Condition *T* states of everything else in the scrutability base, "this is all there is!" – in other words, that every truth is scrutable from the facts in the base. Chalmers and Jackson acknowledge that *T* itself doesn't follow from the other fundamental facts. However, when added to the base for the ideal reasoner's deduction, the union of the physical facts with *T* allows the ideal reasoner to rule out the possibility of "ghost water", *a priori*. In fact, *T* will successfully allow the *a priori* deduction of all of the other negative truths from the fundamental facts – not only that there is not ghost water, but there are no ghosts. So, when added to the other fundamental facts *T* will be sufficient to rule out competition to the physical scrutability base when hunting for physical properties to fill a conceptual role.

Second, Jackson and Chalmers add to the deduction base two further sets of facts which are supposed to make the possibility of conceptual-analytic scrutability more plausible for ordinary macroscopic truths. The first are the *indexical* truths, *I*: the facts about who's me, where's here, and when's now. The second are the

phenomenal truths, $Q$. If Jackson's physicalism is correct, then $Q$ should be scrutable from the physical truths $P+$ the totality fact $T+$ the indexical truths $I-$ so, if one takes $PTI$ as the scrutability base for some ordinary macroscopic truth, then one gets $PQTI$ as a scrutability base for free. If, instead, Chalmers's dualism is correct, then $Q$ is not scrutable from $PTI$ and $Q$ belongs alongside $P$ in the set of fundamental facts. Either way, Chalmers and Jackson are licensed in allowing the ideal reasoner to use $PQTI$ as the scrutability base for the ordinary macroscopic facts about lions and oceans and tables and chairs.

Third, given this expanded scrutability base, it is arguably more plausible that an ideal reasoner given access to $PQTI$ would be able to deduce the macroscopic truths. The ideal reasoner now knows not only the physical arrangement of the atoms in the ocean, after all. The ideal reasoner also knows what it's like for someone to see the ocean – what it's like for every member of a speech community to think about or conceive of an *ocean* – and what physical phenomena the various members of the speech community were referring to when one of them first used the word 'ocean' and others adopted this use. The same applies to "water" – and the ideal reasoner will see that the watery stuff that the speakers of English call "water" is in fact actually $H_2O$.

Perhaps the methodology of science does not involve conceptual analysis when concluding that, for instance, there is a reductive explanation of life as a physical process. But while scientists may be better than average reasoners, they are not ideal reasoners. For scientists, identities like water = $H_2O$ remain *a posteriori* in large part because there are *other* physical facts about the world which are unknown. Against the backdrop of our ignorance of the world, our knowledge of

some part of it will remain *a posteriori* and subject to doubt, but for an ideal reasoner who knows everything fundamental *and* knows that he knows everything fundamental (given the totality condition) identities like water = $H_2O$ will be entirely certain. Block and Stalnaker wrongly suppose that because identities are metaphysically primitive and "rock bottom", the identities must also be explanatorily primitive – once one gets an identity, there is nothing further to explain. However, there is still an explanatory story to be told about by what means and on what evidence one came to conclude that an identity holds and whether that conclusion is justified. In the case of the scientist, this story will involve *a posteriori* justification only, but in the case of the ideal reasoner, Chalmers holds, if *p* is necessary then *p* is *a priori*.

This is a difficult and controversial view to defend, given the popularity of anti-descriptivist causal-historical accounts in the philosophy of language, and the widespread acceptance since Kripke (1980) that there are *a posteriori* necessities. To try to ease the burden, Chalmers offers a defense of two-dimensional semantics, which is supposed to offer a way in which accepting that there are genuine *a posteriori* necessities can be made consistent with holding that in some cases a lack of ideal *a priori* deducibility amounts to a lack of metaphysical necessity. This use of two-dimensional semantics is controversial, but if it is successful it provides a way in which at least in some cases a failure of Cosmic Hermeneutics is a reliable guide to the failure of metaphysical necessitation, and thus a failure of ontological reduction.

### 4.4 Chalmers's 2-Dimensional Semantics

Two-dimensional semantics was developed early on by Kaplan (1979) and Stalnaker (1978), as a way to account for apparent changes in the content of an

assertion depending on the context in which it occurs. Stalnaker understands the content of an assertion to be a narrowing of the field of epistemic possibilities – by asserting that *p*, I rule out some number of possible worlds, and this is the meaning of my utterance *p*. In most cases, the content of an assertion of *p* will be identical to the semantic content which *p* has in our language: the worlds which I rule out by expressing *p* will be the same worlds which the semantic content of *p* in our language rules out. When I say "The car is in the driveway", there is no difference between what my sentence means and what I mean to communicate. The worlds I rule out epistemically are the same as the worlds at which the sentence is false – ones at which the car is not in the driveway.

However, there are situations in which the content of my *assertion* that *p* cannot plausibly be identical to the semantic content of *p*. Suppose that you and I are watching a hockey game, and a player comes quickly from out of our line of sight and checks another player into the boards. You complain, "*he* doesn't play fair". Neither of us knows, however, whether the player referred to by *he* is player #48 or player #52. Suppose the player who ran the other into the boards is actually player #52. The *semantic* content of your sentence was thus "Player #52 doesn't play fair." However, this can't be the content of your assertion, because nothing ruled out, in our context, the epistemic possibility than player #48 is the one who ran the other into the boards. In this case, my assertion that *p* narrows the field of epistemic possibilities which for the semantic content of *p* aren't possibilities at all.

To take another example, suppose I assert that "Hesperus is Phosphorus". The semantic content of my utterance is a necessary truth – it doesn't rule out *any* metaphysically possible ways the world could be. However, my utterance clearly did

communicate something and did rule out a number of *epistemically* possible worlds: namely, those in which the star one sees in the morning is distinct from the star one sees in the evening.

Chalmers (1996) adapts two-dimensional semantics into an account on which every proposition is associated with two (rather than one) intensions.[262] The *secondary intension* is the set of worlds at which the content of the proposition is satisfied, where any rigid designators in the proposition (like 'water', or 'Aristotle') retain their actual value. The secondary intension of "Aristotle drank water" is the set of worlds in which the person who is Aristotle in our world drinks $H_2O$, even if in those worlds Aristotle is illiterate rather than the teacher of Alexander, $H_2O$ is a rare and valuable substance, and XYZ is the stuff in the lakes and streams. The *primary intension* is the set of worlds such that, were the world considered *as though it were actual*, then the characteristic descriptions associated with the terms in the proposition would justify us in believing that the proposition was satisfied in that world. The primary intension of "Aristotle drank water", in other words, is the world in which the Greek philosopher who was the teacher of Alexander the Great drank the clear, odorless, colorless substance which flows in lakes and streams and falls from the sky when it rains.

What is the upshot of this distinction supposed to be? According to Chalmers, in many cases, the primary intension of a concept will diverge from its secondary intension for ordinary reasoners. "Water" is such a case, as mentioned above: the secondary intension of water picks out $H_2O$ in every world, whereas the primary

[262] also see Chalmers (2010).

intension picks out whatever it is in that world which most closely matches with the characteristic features of our own concept of water. However, the primary intension of a concept will not diverge from its secondary intension for an ideal reasoner.

Why not? Recall for a moment why there are *a posteriori* necessities in the first place. They exist because some terms in our language like proper names, natural kind terms, and demonstratives are *rigid designators*: they designate the same individual in every possible world, without regard to the conceptual or descriptive properties we may happen to associate with the term. Part of what it is to be water is to be the *actual* watery stuff; part of what it is to be Aristotle is to be the *actual* teacher of Alexander the Great, as opposed to someone else who could have been his teacher. This '*actual*' operator is part of the secondary intension of these terms. However, epistemically, we're in a non-ideal spot: we don't know which world is actual. When we consider or imagine various scenarios where our concept of watery stuff applies, we have to treat the scenario we are considering as if it were actual, but also have to admit that we could be seriously wrong about which world is the actual one. We (epistemically) could be wrong that the thing we've been calling water is $H_2O$, although if we are right, then it couldn't (metaphysically) be anything other than $H_2O$.

On the other hand, an ideal reasoner knows which world is actual, and knows everything in *PQTI* for our world. The ideal reasoner does not have to wonder what the thing people refer to by 'water' is: he can see what they are thinking about when they say 'water' and he can see what physical entity it is which causes their talk about 'water' and gets used whenever they say 'water'. Because an ideal reasoner has no epistemic limitations when it comes to determining which world is actual, the

primary intension and the secondary intension will be the same for every term. So, given *PQTI,* an ideal reasoner will be able to deduce the ordinary macroscopic facts, even though we cannot hope of every doing so.

According to Chalmers, consciousness is supposed to be a special case, where we can be confident that the primary intension of our concept of a "conscious experience of pain" lines up exactly with the secondary intension of our concept of a "conscious experience of pain". As Kripke (1980) argued, anything which seems like pain simply *is* pain. So, for consciousness, the primary intension will pick out the same worlds as the secondary intension – anything that has the characteristic features of consciousness is consciousness. So, even though we don't have the mind of an ideal reasoner, we are justified in concluding that even an ideal reasoner could not deduce *Q* from *PTI*.

To put it differently, Chalmers argues that ideal conceivability entails epistemic possibility, and that in the special case of consciousness epistemic possibility entails metaphysical possibility, offering a bridge from conceivability to metaphysical possibility.[263] Any epistemically possible world where P & ~Q is necessarily a metaphysically possible world.[264]

---

[263] Chalmers (2010): Section 3

[264] For example, it is conceivable that Mary, a neuroscientist who lacks color vision, could know all of the facts about neuroscience and color vision, and yet not know what it is like to see the color red. (Jackson, 1982) It follows that there is a 1-possible world in which Mary knows all the facts about neuroscience and color vision, but yet does not do anything qualitatively similar to knowing what it is like

## 4.5 Challenges to 2-Dimensional Semantics

Scott Soames (2005), Block and Stalnaker (1998), and others have raised a number of challenges to Chalmers's project of 'ambitious' two-dimensional semantics. I will mention only three here, as well as the relevant replies by Chalmers (2010).

**4.5.1 Constancy.** First, it's not clear what should be kept constant. Suppose the conceptual analysis of water tells us that water is the clear odorless liquid which fills lakes and streams, but doesn't tell us what water's chemical composition is. This is a fairly rich description, and we can imagine a scenario *s* in which a clear odorless liquid fills lakes and streams and has a very different chemical composition than water does. However, this isn't quite fair: we're relying on a rigid interpretation the terms "clear", "odorless", "liquid", "lakes", and "streams" in the description which picks out *s*. Each of these takes a conceptual analysis and has a primary intension. Suppose we replace all of the concepts in our descriptions with their respective analysis, and remove any rigidifying expressions. What will we be left with, if anything? Will it be enough to still pick out *s*? It would be absurd if the answer is

---

to see the color red. (That is to say, were this world considered to be actual, it would be true that "Mary does not know what it is like to see red"). But any (1-possible) world in which Mary do anything qualitatively similar to knowing what it is like to see the color red is a (2-possible) world in which Mary really does not know what it is like to see red, insofar as anything qualitatively like a conscious state is a conscious state. So we can move from the conceivability of a case like Mary's to the possibility of a case like Mary's.

that we are just supposed to keep fixed orthography, but there seems to be no other clear answer.

In reply, Chalmers emphasizes that it clearly isn't orthography which is held constant – evaluating the primary intension of water does not involve evaluating what "water" would mean in the mouths of speakers of another language in another scenario.[265] Primary intensions are evaluated in terms of the epistemic properties of a concept in our world. However, this still leaves the question of whether there will be much of anything left to pick out *s* when those epistemic properties are traded in for their non-rigid conceptual-analytic description.

**4.5.2 Ascriptions of Belief.** Primary intensions are supposed to capture the epistemic features of our expressions. Soames raises a concern that "*S* believes that *F is G*" ought to be true, on a two-dimensionalist account, if and only if the content of *S*'s belief matches up with the primary intension of "*F is G*". This would entail that the content of "*S* believes that *F* is *G*" is identical to the content of "*S* believes that *the actual F* is *the actual G*". This leads to a contradiction. It is a necessary truth that, if *S* believes that *the actual F is the actual G*, and that belief is true, then *the actual F is the actual G*. But it is not a necessary truth that, if *S* believes that *F is G*, and that belief is true, then the *actual F is the actual G*.[266] Were two-dimensionalism correct, these two sentences would be equivalent.

Chalmers responds by denying the original charge – he rejects the claim that ascriptions of belief that *p* are true if and only if the primary intention of *p* is the

[265] Chalmers (2010), 561

[266] Soames (2005), 272.

content of the belief. Soames's objection is a straw-man, an attack on a position which may seem *prima facie* intuitive for a two-dimensionalist but which neither Chalmers nor any other two-dimensionalist has held. In fact, Chalmers holds that it is the secondary and not the primary intension which is relevant in belief ascriptions.[267]

**4.5.3 Intuitions aren't Fine Grained.** Intuitions about the application of a concept to various scenarios aren't likely to be fine grained enough that one will successfully be able to find a match in *PQI*. While the ideal reasoner's concept of water may include a characteristic description in terms of lakes and streams, there is no reason to hold out how that pursuing this strategy will lead to an analysis which only makes use of purely microphysical properties. Many of our concepts are vague, they shift over time and between different contexts of use, and few of them invoke quarks. At the very least, it seems unlikely that primary and secondary intensions are all there is to meaning, because primary intensions can't distinguish between pairs of propositions which are *a priori* (consider logical and mathematical truths, which have different cognitive significance and different meanings, but will have identical primary intensions).

Chalmers accepts the latter point – that primary and secondary intensions need not be all there is to meaning. However, the robust two-dimensionalist's argument only requires that the primary intension be some part of meaning, not all of it. It is also clear that the output of a conceptual analysis is unlikely to be in terms of quarks and leptons. However, this is why the role of $Q$ is so essential to the

[267] Chalmers (2004)

project of cosmic hermeneutics. Although quarks and leptons may not plausibly be part of the canonical analysis of ordinary macroscopic truths, collections of perceptual experiences may play this role.

In his most recent work, Chalmers (2012) moves away from concerns about conceptual-analytic scrutability towards a much broader thesis, that of the *a priori scrutability* of all of the truths from *PQTI*. Although it is clear that he thinks of conceptual analysis as continuing to play a significant rule in the *a priori* deduction of the macroscopic facts from *PQTI*, other sorts of *a priori* reasoning might be acceptable as well. I will turn lastly to Chalmers's argument for *a priori* scrutability, through what might be called the "Cosmoscope" approach to Cosmic Hermeneutics.

## §5 Constructing the World

### 5.1 The Cosmoscope

We have been supposing that the ideal reasoner has "access" to all of the truths in *PQTI*, by which we must mean that the ideal reasoner has access to representations of these truths in some form or another. Conventionally, these representations have been thought of as sentences, perhaps in some formal language, and the ideal reasoner has been thought of as methodically attempting to derive one sentence from another. However, Chalmers offers the ideal reasoner a particular device – a "Cosmoscope" – which offers a representation of *PQTI* through *perception* rather than through language.

The Cosmoscope is loaded with all of the information in *PQI*, infinite storage and processing power, and an excellent virtual reality program. It is supposed to allow the ideal reasoner to explore all of the information in *PQI* as though he were watching a three-dimensional movie (or four-dimensional, since one must count the

distribution of *P* and *Q* across all of time). It allows the ideal reasoner to "zoom in on" any region of the world, no matter how remote, and to "zoom into" the phenomenal states of other minds in the world through a kind of virtual reality in order to experience them first hand – all the while marking these movies with a stamp which lets the ideal reasoner know the relevant indexicals in *I*. It even allows the ideal reasoner to simulate counterfactual scenarios. The only thing the Cosmoscope doesn't do for the ideal reasoner is tell him whether a sentence is true or false. The reasoner looks at the world through the Cosmoscope to verify propositions, but isn't given verification.[268]

The Cosmoscope is useful for showing the plausibility of the *conditional scrutability* of all of the positive macrophysical facts from *PQI*. (The negative macrophysical facts require the addition of *T*). Take any sentence in the world that you like – for example, my utterance of "the dog is hungry" earlier today. The Cosmoscope will allow the ideal reasoner to see, first of all, what was in my head when I uttered this sentence, and what I meant by 'the dog' (namely, my own dog) and by 'hungry'. If meaning ain't all in the head, then it will also allow the ideal reasoner to see what is the heads of everyone else in my linguistic community associated with the words 'dog' and 'hungry', and the complete causal history of speech behaviors involving 'dog' and 'hungry' going back to various Indo-European tribes and including plenty of instances of doghood and hunger along the way, and what any authorities I defer to for the meaning of 'dog' would decide were they hard-pressed with the question of whether or not a dingo is a dog. The ideal reasoner will

[268] Chalmers (2012), 108-110

learn from our concept of dog both what sorts of phenomenal experiences dogs characteristically play a role in and what sorts of physical structures are dogs – the indexical facts will be helpful for locating the experiences where the physical events are. The Cosmoscope will allow the ideal reasoner to see my occasion of utterance, and the presence of my dog in the room when I said "the dog is hungry", and our convention for referring by definite descriptions. It will allow the ideal reasoner to experience first-hand the conscious states of the dog, and to know whether or not the dog is experiencing something relevantly similar to the experience I call hunger. (Although, it may cause the ideal reasoner to experience what it's like to be a dog by *imagination* rather than getting full experience, since the ideal reasoner needs retain the ability to reason while observing). At this point, the ideal reasoner can offer a judgment as to whether my sentence was true or false.[269]

A similar process can be performed in a "Twin Earth" scenario:

> If a Cosmoscope tells us only that there is a watery liquid made of H2O, we cannot thereby conclude that water is H2O, as we cannot rule out a hypothesis on which the Cosmoscope is showing us a distant planet with H2O and on which water (our water) is XYZ. However, this sort of hypothesis will be ruled out by indexical truths fixing our relation to the objects. Given these truths, we can determine that the H2O we are seeing is in our own

[269] ibid., 110-125

environment and that H2O has been the relevant liquid in our environment all along. [270]

Chalmers believes it is plausible on this basis that all of the ordinary macro-scale truths are conditionally scrutable from *PQI*: the ideal reasoner can use the Cosmoscope to determine the meanings of various sentences, can calculate what would follow on the assumption of a given *PQI*, and can look through the cosmos to see if those sentences are verified. Notice that this process appears to be partly *empirical* – the ideal reasoner looks and "perceives" what follows from the assumption that *PQI* is the case. In itself, it not *a priori* scrutability. Yet it is *a priori* scrutability which is of metaphysical interest to Chalmers. So, how do we get from conditional scrutability to *a priori* scrutability?

## 5.2 From Conditional to *A priori* Scrutability

Chalmers offers two arguments meant to bridge the gap between the conditional scrutability of macroscopic truths from *PQTI*, which he justifies by means of the Cosmoscope thought experiment, and the *a priori* scrutability of macroscopic truths from *PQTI*.

**5.2.1 The Argument from Suspension of Belief.** Suppose that the ideal reasoner uses the Cosmoscope's ability to process the information in *PQI* into an imaginative experience, but is able to disregard entirely the way the world actually is. The ideal reasoner engages in a kind of Cartesian skepticism about the actual world and ceases to believe that *PQI* are actually the case. Conditional scrutability

[270] ibid., 124

works the same way as before: the ideal reasoner concludes on the basis of the Cosmoscope's rendering of *PQI* that *if PQI*, then *M*, for the set of positive macroscopic truths *M*. However, nothing in the ideal reasoner's process of reasoning appeals to perceptions of the actual world. It appeals to the experience of imagining of what would be the case if *PQI*, much like an armchair thought experiment turned into immersive virtual reality. The ideal reasoner's justification for the conditional belief that *if PQI* then *M* can't be empirical, Chalmers argues, since it's not perceptual. So, it's justified *a priori*. So, *PQI* → *M* is *a priori*, and *M* is *a priori* scrutable from *PQI*. (The negative macroscopic truths are then scrutable *a priori* from *PQI+M+T*).

**5.2.2 The Argument from Frontloading.** Suppose instead that there is some empirical evidence which justifies the ideal reasoner's conclusion of *M* from his conditional assumption of *PQI*. Often our background beliefs about the world or our contingent epistemic position as human reasoners may lead us to feel that certain inferences are "intuitive" or "natural" or "automatic", and thus we wrongly conclude that they are *a priori*, even though they are dependent for their justification on past experiences. For instance, it is 'intuitive' and 'obvious' that when I drop a coin, it will fall down towards the earth. However, my justification is not *a priori*, but rather a variety of empirical evidence which I have obtained over a lifetime of living on the surface of an object with substantial gravitational pull. Likewise, it is 'intuitive' that living things move. However, my justification is not *a priori* from the concept of 'living thing', but empirical from observations of living organisms. Perhaps the ideal reasoner relies on evidence like this when discerning the macroscopic truths through the Cosmoscope. If so, call this evidence *E*.

If it is the case that the conditional belief that *if PQTI then M* is justified by *E*, then from *PQTI+E* the macroscopic truth *M* should follow *a priori*. *PQTI* → (*E* → *M*) is equivalent to (*PQTI* & *E*) → *M* by exportation. The evidence *E* needed for the scrutability of *M* from *PQTI* can thus be "front loaded" into the scrutability base, and *M* will be *a priori* scrutable from this newly expanded base.

However, Chalmers argues, empirical evidence is necessarily evidence obtained by conscious experience. Thus, *E* is already part of *Q*, the totality of phenomenal properties in the cosmos. The ideal reasoner should be able to deduce *E* from *Q*, meaning that *PQTI+E* follows *a priori* from *PQTI*. So, it follows that *M* is *a priori* scrutable from *PQTI* alone.

### 5.3 Hard Cases

To sum things up so far, Chalmers (2012) has proposed that we adopt a model on which Cosmic Hermeneutics amounts to *a priori* scrutability. He has argued that his thesis of the *a priori* scrutability of all of the facts from *PQTI* follows from accepting the less controversial claim of the conditional scrutability of all of the facts from *PQTI*. In the Cosmoscope, he has provided a thought experiment which makes the conditional scrutability of all of the facts from *PQTI* for an ideal reasoner seem plausible.

Chalmers's notion of *a priori* scrutability is broader and more permissive than the notion of *conceptual-analytic* scrutability discussed in Section 4. This can be seen most clearly when Chalmers handles certain "hard cases", like mathematical truths, normative and evaluative truths, ontological truths, intentional truths,

counterfactual truths, and social truths.[271] It is agreed by many that all of these truths are *a priori* given *PQTI*. For some cases, like certain universal normative truths (if any) and mathematical truths, these truths might be *a priori* for the ideal reasoner given *any* scrutability base.

What is far more controversial is Jackson (1998)'s claim that these truths would be *analytic* from *PQTI*, obtainable by means of conceptual analysis. So, Chalmers's move away from conceptual analysis alone and towards a broader notion of the *a priori*, which includes *a priori* synthetic truths, allows him to justify the *a priori* scrutability thesis. Chalmers does suggest that conceptual-analytic scrutability is possible if the base is expanded from *PQTI* to include a much richer set of truths, such as mathematical, spatiotemporal, nomic, ontological, and normative truths.[272] However, without the addition of these to the scrutability base, analytic scrutability is likely not possible.

Chalmers is aware that his project is concerned only with the *conceptual* structure of reality, rather than its metaphysical structure.[273] Cosmic Hermeneutics conceived of in terms of *a priori* scrutability will tell us which concepts are conceptually fundamental, but it will not tell us which entities are *ontologically* fundamental or answer the question of what grounds what. Yet it is ontological analytic scrutability, not sentential *a priori* scrutability, which is most likely to tell us about the metaphysical structure of reality. Whereas *a priori* scrutability from

---

[271] See Chalmers (2012), Chapter 6 for these topics

[272] Chalmers (2010), 390

[273] ibid., 442

one set of sentences or propositions to another tells us about the fundamental "relations between ideas", if our goal is to understand the relations between the natures of things, our concern should be with what would be *ontologically* and *analytically* scrutable from the fundamental facts: that is, which are wholly contained in the others and thus ontologically reducible to them. I will develop this idea of ontological analyticity further in Part B of this chapter.

## 5.4 Preliminary Conclusions

Chalmers also offers an account on which $P$ and $Q$ are clearly equal partners in the project of cosmic hermeneutics, in notable contrast to the older tradition on which $P$ alone is supposed to play this role. It does seem plausible to me that, given the addition of $Q$, Cosmic Hermeneutics would be possible. So long as Chalmers limits his project to telling us the structure of our concepts, everything will be okay. However, ambitious metaphysics might be tempted to identify the ideal reasoner's *a priori* deduction of the world as telling us which facts are fundamental or emergent, and which are reducible. For three reasons, this seems to me to be a mistake.

First, discerning macrophysical objects at various scales above the level of microphysics depends on the Cosmoscope reaching the right 'zoom' level and the ideal reasoner's use of $Q$ to 'perceive' the distinctive phenomenology of observing chairs, elephants, or so on, at their respective levels. Why does the ideal reasoner care about the aggregations of physical pieces which we call chairs and elephants? Why not half-chairs, or elephant-pairs? Why wouldn't the ideal reasoner consider these as equally interesting as a rather arbitrary mereological sum of physical particles distributed across space and time? Chalmers must appeal to the ideal reasoner's interest in *our* interest, as speakers of a language with 'chairs' and

'elephants', and the ideal reasoner might rely on our phenomenal states when we *perceive* chairs and elephants to discern them through the Cosmoscope. Yet the notion of *perceiving* macro-scale objects as such (as opposed to perceiving them as aggregates) suggests that there is something *real* and intrinsically interesting occurring at the macro-level of organization, a "joint in nature", a "perfectly natural" property, a nature over and above the nature of its parts.

Second, Chalmers's notion of the *a priori* is so broad that it risks including any truth which the ideal reasoner would find "intuitive" or "obvious". However, there are good reasons to be skeptical about a notion of the *a priori* which involves only "intuition". Someone might charge that it is intuitive that one ought to bring a gift to a wedding worth more the cost of one's share of the meal at the reception, or intuitive that objects with heavier masses will fall faster than objects with lighter masses, or intuitive that men by nature have short hair and women by nature have long hair. These intuitions are not a very good guide to truth. We are justified in believing them only insofar as they involve veridical experiences of the actual world. Again, so long as we limit ourselves to discerning the structure of our concepts, perhaps there is nothing wrong with a reliance on "intuitions" like this. However, if our purpose is to learn something about metaphysics, then the sort of *a priori* scrutability which tells us about the structure of reality must be more narrow and not simply a matter of seeming to be intuitive.

Third, it's not clear what the metaphysical significance of cosmic hermeneutics with such a permissive notion of the *a priori* is supposed to be. If the only purpose is (in conjunction with two-dimensional semantics) to allow us to conclude that macro-scale objects are real and that they are metaphysically

necessitated by the distribution of properties in *PQTI*, then the project succeeds. However, it seems that we are being encouraged to infer that there is something metaphysically fundamental about *PQTI* and something non-fundamental about everything which can be deduced from it. Whereas definitional scrutability and analytic scrutability seem more clearly connected to the notion of an ontological reduction, *a priori* scrutability does not seem to offer us an ontological reduction. The constraints Chalmers puts on the *a priori* are not "meaning constraints", to use Horgan's (1982) expression. Consider the following observation by C. D. Broad (1930), in discussing the distinction between naturalism and non-naturalism in ethics:

> It is very common to find that the following two propositions are not clearly distinguished from each other, viz.:
>
> (a) "The ethical characteristic E synthetically entails and is entailed by the non-ethical characteristics N1, N2, . . ."; and
>
> (b) "The ethical characteristic E is analyzable without remainder into the conjunction of the non-ethical characteristics N1, N2, . . ."
>
> Many moralists are liable to think that they believe (b) when they really only believe, or only have produced reasons for believing (a). Non-naturalistic theories can, and generally do, accept some propositions of the first kind. … [but deny] that to be good can be analyzed into containing a positive balance of happiness.[274]

[274] Broad (1930), 258

The ethical non-naturalist accepts the *a priori* entailment of ethical properties by the descriptive properties in the world (both physical and phenomenal), much like the naturalist. Yet the non-naturalist holds that ethical properties are fundamental in a way in which the naturalist does not, because the non-naturalist holds that there is no *analysis* of the ethical properties in descriptive terms.

If the project of Cosmic Hermeneutics is to have any ontological significance in distinguishing emergent from non-emergent properties, as Broad originally conceived of it, then it must involve a much more narrow sort of scrutability than Chalmers, Jackson, or Lewis conceived of it has having. Chalmers's *a priori* scrutability is too broad, as I have discussed: it includes ethical truths, even if non-naturalism is correct. Lewis's *simulative* scrutability is also too broad, I will argue, and suggests a kind of emergence. Carnap's older *definitional* scrutability is too narrow, since we rarely have definitions for relevant terms.

Jackson's *conceptual-analytic* scrutability is much closer to the mark of an ontologically significant program, but it is concerned with the wrong subject matter (propositional instead of ontological), bound up with a defensible but controversial use of two-dimensional semantics, conflates the ontological grounding relationship with supervenience, and remains committed to the 'analyticity' of intuitions about meaning despite recent evidence that philosophers ought to be more skeptical about the universality and objectivity of their intuitions.[275]

[275] See Stich, S. (2000) for a critique of Jackson; see Knobe and Nichols (2008) for a critique of philosophers' intuitions.

I will argue instead that, if cosmic hermeneutics is to be a guide to fundamentality, then cosmic hermeneutics must be conceived in terms of ideal ontological mechanical-analytic scrutability.

## PART B

## §6 Reconceiving Cosmic Hermeneutics

### 6.1 Introduction to Part B

In part A, I discussed a number of historical approaches to the question of how an ideal reasoner might come to know all of the truths from a base of some concise category of truths, or the question of "cosmic hermeneutics". I distinguished varieties of what Chalmers (2012) calls *scrutability* in terms of:

1. The scrutability base: whether the ideal reasoner must start from the physical truths, the phenomenal truths, and/or some other fundamental or non-fundamental truths.

2. The form in which the base truths are represented: sentential, propositional, or ontological.

3. Cognitive resources: which patterns of inference the ideal reasoner is permitted to use, and whether these are limited to *a priori* truths and/or *analytic* truths.[276]

---

[276] That is, whether conclusions must follow *a priori* from premises without *a posteriori* bridge statements, and whether conclusions must follow analytically from premises without synthetic bridge statements.

Jaegwon Kim (2010) and others have held that a truly "ideal" reasoner would seem to have *a priori* access to *all* of the necessary truths. I call this *entailment scrutability*. If cosmic hermeneutics is understood in terms of entailment scrutability, then the distinction between what an ideal reasoner can and can't deduce from a given base simply collapses into the distinction between those truths which are or aren't metaphysically necessitated by that base. If this is so, then cosmic hermeneutics is not of any special philosophical interest, except as an illustration of what metaphysically necessary supervenience might mean.

David Lewis's project of Radical Interpretation involves what I call *simulative scrutability*, a case where the ideal reasoner has access to all inferences which are analytic *a priori* justified, mathematics, and a subcategory of synthetic *a priori* justified inferences which involve running a simulation of what might be happening within a system without being able to directly perceive that system. Whether the ideal reasoner runs this simulation within his or her head, or whether the ideal reasoner relies on a "Cosmoscope"[277] or other device to run the simulation and observes the output, it makes no difference. For Lewis, the simulation involves the application of Davidson's "principle of charity" in order to deduce the beliefs and desires of a subject, Karl, which give rise to Karl's speech.

In his recent work, Chalmers (2012) argues for *a priori* scrutability on the basis of *conditional* scrutability. For Chalmers, when an ideal reasoner is given access to all inferences which are justified *a priori*, all truths are scrutable on the

[277] The reference is to Chalmers (2012); see the discussion in Section 5 of this Chapter

basis of *PQTI,* which consists of the fundamental physical truths, the phenomenal truths, the indexical truths, and the "totality condition" that all truths are scrutable from *PQTI*. Chalmers's project provides insight into the relationships between our concepts, and *a priori* scrutability (or the lack thereof) can be used to argue (somewhat controversially) for metaphysical necessity (or the lack thereof). However, because *a priori* scrutability is so permissive, it fails to distinguish between the claim that one domain is necessitated by another and the claim that one domain is *grounded in* another.

A closely related project involves *conceptual-analytic scrutability*, where an ideal reasoner's resources are limited only to those inferences which are *analytic a priori* justified, including inferences from the meanings of terms which come from intuitions about the application conditions of various concepts. Because this project is more restrictive than *a priori* scrutability, since the ideal reasoner is limited to 'analytic' resources alone, it comes closer to supporting an ontological thesis. If our concepts for macro-level properties can be given an analysis in terms of a functional role which is filled by physical properties, then this gives us an account of how the macro-level properties *really* are just physical properties, though it hardly entails it.

However, even this is still short of the role which cosmic hermeneutics is supposed to play in the definition of emergentism, on which emergent properties are those which are not ideally deducible from the basal conditions which give rise to them. The absence of a conceptual analysis in physical terms for some property is not enough to justify a belief that the property is ontologically something over and above the physical: a conceptual gap, or a gap in the sorts of scenarios people are

able to form meaningful intuitions about, does not entail a metaphysical gap.[278] Common physicalist responses to dualist arguments emphasize that the same phenomenon might be apprehended under two different concepts (one first-person or phenomenal, and the other third-person or physical), or call into question the reliability of the intuitions on which dualism is built: it might be that our intuitions about inverted or zombie worlds are the product of a kind of cognitive illusion.[279] Conversely, the existence of a conceptual analysis for some property in terms of a functional role which some physical property can fill suggests, but does not guarantee, that the property is grounded in the physical property.

In this section, I offer a competing proposal, on which the ideal reasoner's deduction base is more restrictive still. On my account, cosmic hermeneutics should be thought of in terms of *ontological mechanical-analytic* scrutability for the purpose of distinguishing emergent from non-emergent properties. On my account, cosmic hermeneutics is supposed to tell us whether the *nature* of some phenomenon is fundamental or non-fundamental.

In Section 7, I will extend this account of scrutability to allow for stronger and weaker kinds of emergence, depending upon which sorts of ontologically

---

[278] At least, not without extensive argument, as discussed in Section 4 of this chapter.

[279] For two examples of an extensive literature, see Tye (1999) and Loar (1997). For critiques of this strategy, which I will not discuss in detail, see Stoljar (2005) and Horgan and Tienson (2010).

synthetic *a priori* inferences are needed for a deduction of an emergent property from its basal conditions.

In Section 8, I will attempt to further clarify the relevant notion of "analyticity" for discussions of emergentism and the analytic *a priori*. It is unlikely that there is one single analytic/synthetic distinction. Rather, there are a series of contrasts to be drawn between various forms of reasoning and justification. Emergent phenomena may be understood as those which are not ideally deducible from their basal conditions by means of mechanical procedures alone.

Finally, in Section 9, I will suggest some surprising consequences of understanding ontology in terms of cosmic hermeneutics. In particular, I will conclude that that (i) emergentism remains plausible for consciousness even if the "epistemic gap" is merely a feature of our contingent cognitive limitations, and conscious facts are *a priori* deducible from physical ones, (ii) emergentist accounts are motivated by the failure to find an analysis of emergent phenomena rather than by philosophers having or lacking certain intuitions, and (iii) emergentist accounts might apply to a much wider variety of phenomena across the special sciences than generally assumed – a point which I will explore further in Chapters 5 - 7.

## 6.2 Grounding and Deducibility

**6.2.1 Does Grounding Entail Deducibility?** Recall that C. D. Broad defined an emergent phenomenon as one which was not predictable from a knowledge of its basal conditions, even by the Mathematical Archangel, except by actually observing the correlations between the two properties. Compare this to the definition of an ontologically emergent phenomenon which I offered in Chapter 2, on which basal conditions must be sufficient for the emergent phenomenon, but the emergent

phenomenon is nonetheless ontologically distinct from the basal conditions, defined in terms of grounding:

> (Distinctness) If a phenomenon *e* is ontologically emergent from basal conditions *b,* then *e* is not wholly grounded in *b*.

What is the relationship between ideal deducibility or scrutability (an epistemological question) and grounding (a metaphysical question)? The biconditional doesn't likely hold: it is not the case that E is ideally deducible from base B if and only if E is grounded in B. After all, grounding is usually conceived of as being irreflexive and asymmetrical, but every truth is ideally deducible from itself. Further, consider that it is ideally scrutable from the fact that *s* knows that *p*, that *p* is true, given that the ideal reasoner has access to the definition of knowledge – but *p*'s truth is not grounded in *s*'s knowledge. However, perhaps the following one-way conditional claim is worth exploring:

> (GD) if *E* is grounded in *B*, then *E* is ideally scrutable from *B*.

It could turn out that (GD) is trivial. If grounding is not distinct from metaphysical necessitation, and ideal scrutability is interpreted as *entailment* scrutability, then the two collapse into each other. However, we will assume that grounding, as a type of asymmetric ontological explanation, can be distinguished from mere necessitation. Consider then two different ways in which (GD) might turn out to be defective: (i) the grounding relation could fail to be represented in the ideal

reasoner's representational system, or (ii) we could choose the wrong scrutability relation, so that the resources given to the ideal reasoner could be either too weak or too strong.

**6.2.2 Representational Failures.** It might be none of the cases in which $E$ is grounded in $B$ are cases in which $E$ is ideally deducible from $B$, because the form in which the truths in $E$ and $B$ are represented is not the sort of thing whose deductive relations can be used to represent truths about grounding. We know that well-formed sentences can fail to express any meaningful proposition: *colorless green ideas sleep furiously*. We also know that sometimes the concepts in our propositions fail to match up with the natures of things, like the proposition that *phlogiston is released during combustion*. We can recognize in history that many concepts in the past were loaded with racist, sexist, or culturally-limited biases, and we are likely often blind to the same errors in our own concepts today. In a worst-case scenario, all of our concepts and all of our sentences could be error.

Of course, to borrow a theme from Burge (2010), I suspect the fact that some *errors* exist is good reason to think that, on the whole, our sentential and propositional representational systems do successfully represent the world. To be in error, a system must represent to begin with. A system which *never* represented successfully couldn't represent at all, and thus couldn't be in error. Nonetheless, we could still be wrong in many cases, especially difficult cases – a point the phenomenal concept theorist exploits.

For this reason, it seems to be preferable to me to interpret the representations with which the ideal reasoner is supposed to work as ontological rather than sentential or propositional: that is, things will represent themselves. We

might imagine (granting ourselves a heavy bit of metaphor) that the ideal reasoner starts out speaking the 'language of ontology' in which every real thing is in one-to-one correspondence with the terms of his language, and that the ideal reasoner has access to a dictionary in which the natures or essences of all of the properties are written out as 'real definitions'.[280] Of course, the ideal reasoner will still have the task of deducing from this base all of the true sentences (e.g., those that come out of Karl's mouth) and the true propositions (e.g., those in Karl's thoughts), but what he starts out with will be a form of representation which we stipulate aligns with the way the world is perfectly, or at least to the maximum extent that the world is in a way which can be represented at all.[281] The grounding relations between things will then appear within the ideal reasoner's representation of the world.

---

[280] I recognize that many philosophers will cringe at this open discussion of essences. Yet it seems to me that disguising a discussion which we all know to *really* be about essences by putting it in terms of linguistic practices or conceptual necessities is generally unhelpful, except on occaision as a way of reminding ourselves how philosophy got back into the metaphysical discussion to begin with.

[281] We are assuming that the world is represent*able* – that is, it has the sort of structure which can be represented by our concepts or our language. It need not be that the world is representable. It may be only that some portion of what is real is representable – this will be what we indicate by "the world". It could be that the world is empty, in the sense that nothing of what is real is representable – that is to say, the world may be transcendent. We are assuming that this is not the case; if it

It remains an open question as to how much *we* can know about whether something is ideally deducible or not. Our concepts may not align with the ideal reasoner's concepts. So, our intuitions about whether or not a concept applies in a given thought experiment will only be as good as a guide to ideal scrutability as the degree to which our concepts accurately represent the natures of things. Choosing an ontological base (as opposed to a sentential or propositional base) in no way eliminates the problem, but locates it safely outside of the mathematical archangel's task. It would certainly *not* be ontological interesting if cosmic hermeneutics failed only because the ideal reasoner's representational system failed.

**6.2.3 Wrong Resources.** Our scrutability relation might be too restrictive: it might be that some of the cases in which $E$ is grounded in $B$ are cases in which $E$ is not ideally deducible from $B$. Alternatively, our scrutability relation might be too permissive: some of the cases in which $E$ is not grounded in $B$ are cases in which $E$ is ideally deducible from $B$. If we give the ideal reasoner too few resources to perform the deduction of $E$ from $B$, then a commitment to (GD) will lead us to conclude that $E$ is emergent even in cases in which it is grounded in $B$. On the other hand, if we give the ideal reasoner too many resources to perform the deduction, then a commitment to (GD) will lead us to fail to recognize that $E$ is emergent even in cases where it is.

An obvious case of too *few* resources is minimal logical scrutability. Here, the ideal reasoner is limited to those inferences which are justified in classical first

---

is, then what we say will apply to that portion of the world which is representable, or will apply (trivially) to the empty set.

order logic with identity.[282] A great many cases of grounding are not likely to be minimally logically scrutable. For instance, it is not minimally logically scrutable from an object's being red that the object is colored. Nonetheless, assuming determinables are grounded in determinates, an object's being colored is grounded in its being red.

A case of too *many* resources, discussed earlier, can be found in *a priori* scrutability, assuming that there is such a thing as the *synthetic a priori*.[283] For instance, assuming there are normative ethical facts of some sort[284], the ethical facts will be *a priori* scrutable by an ideal reasoner from a physical description of the world, given a sufficiently robust conception of the *a priori*. Naturalists, who believe that the ethical facts are grounded in the physical facts, will accept this. Non-naturalists, who believe that the ethical facts are not grounded in the physical facts,

---

[282] Or, if one prefers, some named alternative to classical first order logic with identity.

[283] Of course, someone might hold that the *a priori* is coextensive with the analytic and deny that there are synthetic *a priori* truths. In this case, the *a priori* may be sufficiently constrained to be ontologically significant.

[284] If someone holds that there are no such facts, a similar problem can be raised for the metaphysics of causation or claims about the ontology of mathematics. Platonists and anti-Platonists agree that the mathematical truths are *a priori*; two philosophers could agree that causal facts are *a priori* knowable given the regularities in nature, but one holds that they simply are regularities, while another holds that causation is something over and above regularity.

will equally accept this – on their view, it is supposed to be *a priori* that pain is morally bad, or *a priori* that one ought according to a maxim which one can will to be a universal law. Yet, if non-naturalism is true, then it should be possible to perform *Modus Tollens* on (GD). The word 'derive' in "you can't *derive* an 'is' from an 'ought'" presumes that there is a sense of scrutability which is more restrictive than the synthetic *a priori*, a sense which the non-naturalist takes to indicate that ethical facts are not grounded in descriptive facts.

This isn't to say that *a priori* scrutability tells us nothing about metaphysics. If E is *a priori* deducible from B, then this fact strongly suggests that, assuming the base B is real, E is also real.[285] Certainly, if B is an accurate representation of the world, and an ideal reasoner can deduce *E* from *B*, then *E* must also be a true representation of the world. We might be tempted to adopt a claim like this:

> *(DER) Deducibility Entails Reality*. If E is ideally deducible from B, and B only refers to real objects and properties, then E only refers to real objects and properties.

For instance, from "Tom is a bachelor" one can deduce "Tom is unmarried"; from "Roses are red" one can deduce "Some flowers are colored." It would be strange to hold that bachelors are real, but marriage is an illusion. It would be strange to hold that roses and redness are real, but flowers and colors are not. Granted,

---

[285] In the everyday sense of 'real' identified earlier, as opposed to the sense in which 'real' is synonymous with 'fundamental'.

depending on the view one takes of truth in literary fiction, there may be counterexamples to (DER).[286] However, whether (DER) is true or not, *a priori* scrutability at most tells us what properties or entities our ontology should preserve, or what facts we are committed to when we are committed to a certain base of facts, but it does not tell us about the relations between these facts or what grounds what. A failure of *a priori* scrutability may tell us that a phenomenon is emergent, but the success of *a priori* scrutability is not enough to tell us that a phenomenon is not emergent.

There are several options between *a priori* scrutability and minimal logical scrutability. Which one is relevant to Broad's test for an emergent property? Which one should we adopt for our interpretation of (GD), so as to weed out all and only the non-emergent properties? I will turn to this question next.

### 6.3 Simulative Scrutability?

Slightly weaker than *a priori* scrutability is the relation I called *simulative scrutability*. While simulative scrutability does not allow the ideal reasoner to have

---

[286] Suppose that the ideal reasoner is given a sentence which contains a complete description of the life of Leo Tolstoy, including which words he wrote while authoring of *Anna Karenina*. From this, the ideal reasoner can deduce that the sentence "Anna Karenina was the sister of Stepan Oblonsky" is true. Nonetheless, unless modal realism is true, Anna Karenina is not real, nor is Stepan Oblonsky. Or, suppose that one holds that abstract mathematical properties are not real, but that mathematics is *a priori*. Then mathematics could be deduced from any sentence $S$ – including Tolstoy's life story – but one would not want to hold (DER).

*a priori* intuitions about normative truths (to extent they might exist), it does permit the ideal reasoner a number of other *a priori* resources, including mathematical and analytic *a priori* inferences. In addition, the ideal reasoner is permitted to engage in a kind of perspective-taking or simulation, in which he replicates (or uses a computing device to replicate) the macro-scale behavior of some rule-governed system, and notices certain patterns which re-appear regularly in the system. In some cases, the ideal reasoner is not able to deduce the patterns which appear given the rules and initial conditions alone without actually applying them step-by-step (as in the case of so-called "artificial life"). In certain very complex cases, the ideal reasoner makes inferences about the intentions and goals of the system, including that it composes a reasoner or an agent, guided by a principle of charity.

Some may regard it as debatable whether such simulation counts as *a priori* rather than *a posteriori* – it certainly involves some form of 'observation' – but it is not empirical in the traditional sense of running an experiment or perceiving the deduced states of affairs directly. Conversely, others may regard it as debatable whether simulation of this sort qualifies as non-analytic, as I have defined it, since the running of a computational simulation requires only that a system act on predictable set of rules. However, while some computational simulations are the sort where one can predict the output simply by knowing the rules, others are of the sort which one can only predict by actually mentally running through a step-by-step application of the rules: running, as it were, a mental simulation of what the computational simulation will do. Compare this to having to determine which is the "shortest proof" of some conclusion from a set of premises and rules of inference. While the rules of the game are fully specified, and determining the outcome only

requires applying the rules, it might not be possible to determine the shortest proof without actually running through various possible proofs to see which is shortest, because the conclusion is about the system of rules itself. Simulation seems to occupy an intermediate position on the border between the analytic *a priori* and the *a posteriori*.

Someone might also be concerned that the sorts of phenomena which humans are inclined to "deduce" by observing simulations – whether computational simulations of artificial life or our own psychological simulations of what someone else must be thinking or feeling – are cases in which a certain evolutionary history and biological make-up predisposes us to finding certain 'patterns' particularly interesting or certain conclusions about someone's mental states particularly 'obvious', whereas there is no reason an ideal reasoner would find these patterns particularly interesting or notable in comparison to other patterns. However, while it is reasonable to believe that an ideal reasoner would notice far *more* patterns than we do in the world, and would find interesting patterns in the world which do not strike us as interesting or obvious, this does not undermine the possibility that the patterns we notice are interesting to us in part because they are *real* patterns: our evolutionary history and biological composition has been molded by the structure of reality, thus permitting us to (fallibly) notice some (but not all) of the patterns in nature. While our biological purposes might be different than the ideal reasoner for noticing these patterns – we're interested in survival, he isn't – ideal reasoner might still find patterns like life or psychology interesting for the purposes of understanding the world (perhaps the reason why these patterns are useful for survival is precisely because they are real). There is no guarantee that this is so, and

it is respectable to believe that the patterns we notice are interesting only relative to our biological purposes, not ideally, but nothing in an evolutionary account of human psychology requires that we believe this to be so.

Bedau (1997) has argued that this sort of "deducibility *only* by simulation" constitutes a distinctive sort of 'weak emergence'. We might interpret Bedau's definition as indicating as weakly emergent those phenomena which are simulatively scrutable from their basal conditions, but *not* scrutable from any strictly weaker set of resources. Thus, to find the form of scrutability relevant to (GD), we will need to reduce the resources available to the ideal reasoner.

## 6.4 From Conceptual Analysis to Ontological Analysis

I will pass over what I have called *mathematical* scrutability[287], and turn next to *conceptual-analytic* scrutability. When the scrutability base has a sentential or propositional form, then the topic of conceptual-analytic scrutability aligns closely with the project of Jackson (1998) and others outlined in Section 4. However, I argued in section 6.2.2 that we ought to take the base for the ideal reasoner's deduction to be ontological – to consist in things structured as their natures truly are, representing themselves. In this case, instead of "conceptual-analytic" scrutability we have cosmic hermeneutics conceived of as consisting in *ontological-analytic* scrutability. The ideal reasoner is not permitted to rely upon intuitions

[287] I included this in the class of synthetic a priori forms of scrutability because of a preference towards this position in the philosophy of mathematics, but don't wish to press the point here, and don't know how much of ontological interest is tied to this.

about the application conditions of various concepts, since the natures of things themselves are sitting right in front of his face. Instead, the resources the ideal reasoner uses in an attempt to deduce the higher-level facts from the base facts will be those which I identified as involved in *mechanical-analytic scrutability*:

> (c) *Mechanical-Analytic scrutability* holds iff *A* is scrutable from *B* when only logic, definitions, and all other analytic *a priori* inferences which involve only the mechanical application of a rule, independent of content or meaning, are permitted.

Here, it seems to me, we have a test for minimally emergent properties which gives us a useful interpretation of (GD): if *E* is grounded in *B*, then the nature of *E* will be ideally mechanical-analytically scrutable from the nature of *B*, since part of *E*'s nature will be its grounding in *B*. However, unlike *a priori* or simulative scrutability, it is not plausible that even weakly emergent properties (those whose natures aren't grounded in their basal conditions) are mechanically-analytically scrutable from their basal conditions.

Why should mechanical-analytic deducibility be relevant to ontological emergence? Consider that on the ontologically "mechanistic" account which the emergentist is opposed to, only basal conditions themselves are fundamental. No new ontological "content" can be added to the nature of the higher-level property beyond what is already contained in the nature of its base, since this new content would then have to be fundamental also. So, any apparent differences between the natures of higher-level properties and their basal conditions must be explainable in

terms of relations which involve no addition of content. For instance, an object's being red grounds an object's being colored because coloredness is *contained in* what it is to be red, where "contained in" is a relation which can be represented in a purely mechanical way with no addition of content.

One might wonder why we haven't chosen the even weaker criterion of *definitional* scrutability. Mechanical-analytic scrutability does allow the ideal reasoner access to definitions (since we are dealing with the natures of things, these take the form of "real definitions"), so it includes anything which would be definitionally scrutable. However, it also includes any deduction by means of a fully determinate mechanical decision procedure where the procedure itself adds no additional content, even if the procedure is not part of classical logic and does not involve definitions.[288] For instance, it is implausible that one can find a *definition* for what it is to be a 'heap' of rice, given the vagueness of a term like 'heap'. However, for any property the heap has (ignoring external relations to outside objects), an ideal reasoner in principle could specify a mechanical procedure for determining the heap's property on the basis of the properties of the rice in the heap and their spacial

[288] Consider Hilbert's *finitism* in mathematics as an attempt at establishing mechanical-analytic scrutability for arithmetic without the logicist's commitment to definitional scrutability. Although Godel's proof put finitism to rest for mathematics, were something else scrutable by means of the sorts of procedures Hilbert would have accepted, it would count as mechanical-analytically scrutable without counting as definitionally scrutable.

locations – the reasoner does not have to rely on intuitions about norms, simulating the structure of the heap, or so on.

If I am correct that it is ontologically analytic scrutability which is relevant to ontological emergence, the emergentist need not be committed to the two-dimensionalism of Chalmers and Jackson or the view that the ideal *a priori* deducibility or non-deducibility of one set of concepts from another informs us about whether a phenomenon is emergent or not. It may in fact be that an emergent phenomenon is *a priori* deducible from its basal conditions, provided that the deduction requires further resources above and beyond those involved in ontologically analytic scrutability.

In the next section, I will expand further on this account of "ontologically analytic" and "ontologically synthetic" properties, and use it to fill out the notion of weaker and stronger kinds of emergence which I discussed above.

## §7. Ontological Analysis

### 7.1 Essences and Ideal Deducibility

In chapter 2, I proposed that an emergent phenomenon is one which (i) wholly depends for its existence (or manipulatively supervenes) upon some set of basal conditions, but nonetheless (ii) does not ontologically depend upon those basal conditions – it is not grounded in them. In this chapter, I have identified this second condition with the question of whether an emergent phenomenon has a nature which is ontologically analytic given the nature of its basal conditions, and is thus analytically scrutable from them.

Being an *essential property* of another thing is one type of grounding relation, though not the only grounding relation. I am here understanding "essence"

to indicate the consequential or constitutive essence of a thing, Although it's contentious whether essences really are all that much like the ontological counterparts of linguistic definitions, particularly given that semantics is not so very concerned with finding definitions for terms, it remains a useful metaphor to think of the essential nature of a thing is something like a "real definition" of that thing – a relation between it and the set of properties in virtue of which it has the identity it has. Of course, something may have multiple true definitions, but has only one essence[289] -- an essential property is what is *common* to all possible definitions of the thing.[290]

I will assume that while all essential properties are modal properties, not all modal properties are essential properties. In this regard, my assumptions align with a more or less Aristotelian approach to essence[291], but differ markedly with the modalist approach towards essence which has gradually become popular over the last half-century or so, on which all necessary relations are supposed to be essential relations. I will assume that the fact that *f* has G does not tell us that *G* is essential to *f*, since it need not be the case that *f* has *G in virtue of* being *f*. Consider the following example from Kit Fine:

---

[289] Kit Fine (1995) offers the following example: for an Aristotelian, it is essential that red be the color of some *x*, and red can be defined as "the color of that actual tomato", but it is not essential that red be the color of that actual tomato.

[290] ibid.

[291] See Corkum (2008)

> Consider two objects whose natures are unconnected, say Socrates and the Eiffel Tower. Then it is necessary that Socrates and the Tower be distinct. But it is not essential to Socrates that he be distinct from the Tower; for there is nothing in his nature which connects him in any special way to it.[292]

Just as necessary relations need not be essential relations between things, so too, as I have discussed in this chapter, *a priori* relations between things need not involve grounding relations between things, provided that non-analytic resources are relied upon in the deduction. Instead of looking to the *a priori* / *a posteriori* distinction to carve up he world, we might instead look to some parallel of the analytic / synthetic distinction. Here's how.

Typically, the analytic / synthetic distinction is applied to sentences or pairs of sentences. The analytic is supposed to be that which is "true in virtue of meaning", a definition which is no longer so clear as it once seemed. For instance, "all bachelors are unmarried" is supposed to be analytic given the meanings of the relevant terms in English. One also may speak of analytic or synthetic propositions, or of a sentence being analytic given the relevant concepts, regardless of what language they are spoken in: for any language which has a term for the concept *blue* and term for the concept *colored* it will be analytic that *all which is blue is colored*. To be analytic isn't to be *trivial*: perhaps we can learn interesting things through careful analysis of the meaning of a term. We might learn through considering the meaning of "harm" that some practice we regarded as harmful is in fact not harmful,

[292] From "Essence and Modality", Fine (1993)

or through considering the meaning of "part" that some things which we didn't regard as proper parts of a whole in fact are. An analysis of rich terms like "marriage" or "person" can teach us something new.

However, as many undergraduates ask, why should philosophers spend so much time investigating *semantics*? Shouldn't we be concerned about real persons and real parts, not what "person" or "part" means? Presumably, it's because we believe that our words and concepts align in some way with the way things really are, and so the meanings of our words and an analysis of our concepts will yield us some information about the "real definitions" or essences of things.[293] This information need not amount to knowledge: it can be prone to serious error and bias, and so it should be subject to healthy skepticism – one should seek more intuitions rather than fewer. One should not be naïve in thinking that a complete semantics for the term "harm" will be sufficient to tell anyone whether a government should legalize prostitution or recreational drug use. Nonetheless, the weight many philosophers put on semantics suggests that at some level many hope that analysis will tell us something or other about essence.

## 7.2 Ontologically Synthetic and Ontologically Analytic

Let's adopt a specialized, *ontological* sense of "analytic" and "synthetic" – where analysis not relative to one's language or set of concepts, but operates within the language of ontology. Using our metaphor of a "real definition", we might say that one thing is ontologically analytic given another if the "real definition" of the one is wholly contained in the other. Alternatively, understanding *grounding* to be

[293] Fine (2009)

broader than essence, we might say that if one thing is wholly grounded in another, then it is ontologically analytic given the other.[294] On the other hand, if one thing is ontologically synthetic given another, then it is not wholly grounded in the other, and it is not wholly contained in its definition.

Consider the well-worn sentence, "Water = $H_2O$." This sentence is *a posteriori* and it is (linguistically) synthetic given the semantics of English. Likewise, take the proposition this sentence expresses, <Water = $H_2O$>. This proposition is also *a posteriori* and synthetic, because the concept of $H_2O$ is not contained in the concept of water. However, consider what would be analytic or synthetic were objects and properties to represent themselves, and their essences represented as definitions – or else, a language used by an ideal reasoner in a one-to-one correspondence to such a representation. (There is no reason to think any real language capable of fitting the bill.) In this language, it would be analytic that "Water = $H_2O$" – because this would just be to say, "$H_2O$ = $H_2O$". In other words, it is *ontologically analytic* that the thing we refer to by "water" is identical to the thing we refer to by "$H_2O$".

One should not over-apply ontological analyticity, and thereby confuse it with necessity or the *a priori*. Laplace held that every past and future fact was necessitated and *a priori* given the present facts, but *not* that the future or past was

---

[294] Note that neither biconditional holds. One thing may be ontologically analytic from another without being contained in its definition; one thing might be ontologically analytic from another without being wholly grounded in it. These are ways in which one thing might be ontologically analytic, not a definition for 'ontological analyticity'.

ontologically analytic given the present: they were ontologically distinct states of affairs.

For contemporary physics, the essential properties of the fundamental particles in one part of the world are ontologically synthetic given the essential properties of some other fundamental particles in some other part of the world, for the most part, except perhaps in special cases involving entangled electrons. However, on most accounts of mereology, for any non-emergent macroscopic entity – say, a heap of rice – the properties of the whole *are* ontologically analytic given the properties of the spacio-temporal parts, even if the whole has different identity conditions than its parts.

### 7.3 Emergence and Ontologically Synthetic Entities

A monist holds that all of the particular entities are ontologically analytic given the cosmos as a whole. An atomist holds the cosmos as a whole is ontologically analytic given the particular entities in the world. In contrast, an *emergentist* holds that there are distinctive levels in nature, and that at certain levels there are distinctive cases where particular facts remain *ontologically synthetic* given the lower-level facts. In some cases, one thing is "nothing over and above its parts", and so it is a mere "resultant", or ontologically analytic given the facts about its parts. In other cases, however, an entity at one level may be "something over and above its parts", and so the entity is emergent and ontologically synthetic given the entities at any other level.

So, a phenomenon $E$ is ontologically emergent from basal conditions $\boldsymbol{B}$ if and only if $E$ supervenes on $B$ and $E$ is ontologically synthetic given $B$. Returning to the task of cosmic hermeneutics, this means that an ontologically emergent phenomenon

may be *a priori* scrutable from its subvening base, provided that for an ideal reasoner the deduction would require an ontologically synthetic bridge premise – that is, the deduction would not be possible by the analytic *a priori* alone. On the other hand, a phenomenon which is not ontologically emergent (though it may be 'epistemically emergent') may fail to be conceptually *a priori* scrutable from its subvening base for an ideal reasoner, because the ideal reasoner lacks access to some ontologically analytic but *a posteriori* necessary identity.

A phenomenon which is simulatively scrutable but not analytically scrutable, as Bedau (2008) holds for "weakly emergent" phenomena in complexity science, would qualify as emergent on this definition. Likewise, a normative or teleological property might be *a priori* scrutable but nonetheless ontologically emergent. Discussions of "weak emergence" have not typically distinguished between ontologically analytic and ontologically synthetic *a priori* transitions. As a result, we may mistakenly infer that when a phenomenon is *a priori* deducible from another it must be ontologically reducible.

## 7.4 The Spectrum of Emergence

My account thus suggests a hierarchy of emergent phenomena, though there may be many grades of emergence in between, with no clear borderline between them. There are weaker forms of ontological emergence, on which an ideal reasoner is able to deduce the emergent phenomena from its subvening base, but the deduction is not ontologically analytic – and then there are progressively stronger forms of ontological emergence, on which an ideal reasoner needs more and more resources in order to be able to perform the deduction – perhaps a kind of

simulation, perhaps a certain kind of normative reasoning, or perhaps resources beyond our comprehension. We might divide things up roughly as follows:

(a) Empirically predictable, but not *a priori* scrutable: the strongest kind of ontological emergence.

(b) Deducible *a priori*, but only by normative intuitions: a kind of ontological emergence stronger than (c), weaker than (a).

(c) Deducible *a priori*, but only by simulation: the weakest kind of ontological emergence.

(d) Deducible *a priori* by mechanical-analytical means alone: Not ontologically emergent even in a minimal sense.

Consciousness would fall into category (a), insofar as it is arguably not *a priori* scrutable even by an ideal reasoner. Of course, were consciousness to be *a priori* scrutable by an ideal reasoner – perhaps the ideal reasoner could simply look at the patterns in a brain and deduce the presence of a conscious mind there by some method beyond us – it might still qualify as emergent, just more weakly so. Normative and teleological properties would fall somewhere into category (b), with certain higher-level patterns in complex systems which are deducible only by simulation falling into category (c).

Someone might wonder how reliable our own judgments as to whether or not something is "ontologically" analytic could possibly be. On the one hand, it is true that the natures of many things are hidden from us until the empirical details come in – as in the case of heat being molecular kinetic energy. One might even be a pessimist and suppose that, even when all the empirical details have come in, there will be some things whose natures will remain obscure or hidden from us – which will appear deducible only by synthetic means, but in fact will turn out to be ontologically analytic. On the other hand, to whatever extent the nature of something *is* evident to us, our judgments as to whether or not something is analytic from it can be regarded as fairly reliable. While the "loftier" forms of the *a priori* are unclear and are the subject of much skepticism – how is it that one *intuits* the badness of pain or the wrongness of torture? – the mechanical-analytic *a priori* is the sort of thing we have confident mastery over, since it involves only content-free processing of the sort a computing device can be programmed to perform.

There may be, however, an abiding skepticism about the whole analytic / synthetic distinction to begin with – even though I have clarified that I am using these terms in a specifically ontological way, with the analytic corresponding to a kind of mechanical procedure, not as involving concerns about semantics or conceptual analysis. Rather than become entrenched in a discussion of post-Quinean attempts to revive the analytic / synthetic distinction or weigh various proposals for these highly disputed terms, in the next section I will try instead to offer an intuitive distinction – between calculating and counting – which I think maps up to the relevant differences between the way an ideal reasoner would deduce an emergent theory as opposed to a mechanistic theory of some phenomenon. Is this the *real*

analytic / synthetic distinction? I doubt it. However, it will suffice to clarify the sense of "analytic" and "synthetic" I have in mind when saying that emergent properties are ontologically synthetic given their subvening base.

## §8. Counting and Calculating

### 8.1 Mechanism

As mentioned, I do not hope to find a definition of the "analytic" / "synthetic" distinction which applies to all domains and debates in philosophy and is free from counterexample. My purpose to try to distinguish those cases in which ideal *a priori* scrutability of A given B justifies us believing that there is an ontological reduction from A to B, and those in which *a priori* scrutability is consistent with emergence.

Recall that the early British emergentists opposed their program not to materialism *per se,* but rather to *mechanism* – the view that the cosmos worked more or less like a piece of industrial age machinery. The ideal deducibility of A from B matters to the debate between the emergentist and the mechanist when ideal deducibility is understood to be a kind of mechanical procedure, an intellectual *model* of the sorts of levers and pulleys which operate in the mechanist's cosmos.

If there are sorts of *a priori* scrutability which do not involve mechanical procedures, as many philosophers believe there are, then these sorts of *a priori* scrutability are compatible with emergentism: one can hold that A emerges from B, even though A is *a priori* scrutable from B, provided that A is not scrutable by means of mechanical procedures alone. So, for my purposes, "analytic" will indicate a derivation which relies upon this sort of mechanical procedure. As mentioned, there are uses of "analysis" in contemporary philosophy which will fall into the category I call "synthetic".

Instead of using the terms "analytic" and "synthetic" to modify a proposition or pair of propositions, suppose instead that we come up with a pair of terms to describe *two distinctive processes of reasoning* which an ideal reasoner might be engaged in. One of these, we will call "calculating" – reasoning which involves only a mechanical procedure, and whose outputs are analytic. The other, we will call "counting" – reasoning which involves something more than a mechanical procedure, and whose outputs are synthetic. I will try to draw an intuitive picture of the many varieties of reasoning process which *counting* or *calculating* might be applied to.

## 8.2 Calculating

Calculating is a *mechanical* procedure whereby a given input is transformed into a particular output as determined by the application of some rule. By a "mechanical" procedure, I mean a procedure which can be reproduced by some in principle constructible mechanism, whether or not the mechanism is constructible in practice. For example, given "2+3" as an input, I apply the rule for the '+' sign and obtain the output "5". Given "~p & ~q", I apply DeMorgan's law and obtain "~(p v q)".

Of course, the plus sign rule and the DeMorgan's law are obtained by similar calculations from the rules of ZFC set theory and first order logic, and these rules are obtained by mere calculation from truth functions – a truth table is a sort of calculation. There is nothing in the nature of calculation which requires that we start from the basis of the axioms of Zermelo-Frankel set theory or that we start from the basis of classical logic. We might start from a non-standard set of axioms, or we might adopt an intuitionistic logic instead. We could then derive from these axioms a different set of rules which lead us to calculate in a different way.

Calculation is the sort of procedure which is entirely neutral with regard to the content of the thing being calculated. DeMorgan's law is not concerned with what *p* and *q* mean. This is consistent with the fact that calculations can be sophisticated and complex, far outstripping the abilities of the human mind. We now have the ability to construct mechanisms which can perform calculations which the mind cannot perform, though these calculators are not concerned with the content of what they are calculating.

Calculators cannot count. They can *represent* counting, in the sense of "represent" in which numerals represent numbers. Given as input a representation of times (t1, t2, t3 . . . ) through an internal clock, a calculator can attach an index to those times, and so produce in time an output which, upon interpretation by a viewer, appears to be a sequence representing ordinal numbers. I might program a function which can "count" the number of characters in some paragraph. But the calculator is no more counting characters in paragraphs than a watch is counting minutes. One counts times *with* a watch; it is not the watch which does the counting.

## 8.3 Counting

**8.3.1 Making a Count.** In contrast, it is hard to say what *counting* amounts to except by illustration. For example, counting is what I do when I define moments in time by assigning them distinct names: "1", "2", "3". Imagine a conductor of a symphony orchestra, who is assigned the task of keeping the whole orchestra in time. He might set a metronome to perform this task, which would calculate the next beat for him. But suppose he starts from scratch. He begins to count "1, 2, 3 . . ." and then the orchestra plays. Once he has defined the times by counting, then the others can begin the process of calculating when one should play. But the rhythm

itself that he established, the count, was not something he calculated from something else.

Suppose there are two conductors, and two orchestras, each playing the same symphony, isolated from one another so that neither can hear the other. Perhaps one conducts a bit fast, and the other is a bit slow. If each conductor has calculated his or her time from the same starting point – something like the metronome – then necessarily there exists some way of bringing the two counts together. If one has set the metronome at 60 bpm, and the other at 90 bpm, then the half notes of one will be the dotted quarter notes of the other.

However, suppose that instead each conductor simply begins counting. Even if each conductor maintains his or her respective count perfectly, there is no immediate guarantee that there exists a means of translate from the time of one to the time of another.

Counting time is one sort of counting. A different sort of counting involves counting objects. I count the flowers in a garden: 4, 5, 6 . . . and so I engage in an activity, assigning a time stamp to each object, where the stamp is determined by aligning my count of time with my perceptual contents.

One could create a computer which given certain light-patterns as inputs would spit out certain outputs which closely correspond to my counting. For that matter, it may turn out that my neurology is such that this computer provides an excellent model of the process that I actually engage in. But this only would qualify as counting given a certain phenomenology – and only if such a phenomenology were to arise in the computer would it be appropriate to say that the computer was

counting. Counting is a creative act rather than a mechanical one – one creates T1, T2, T3… the calculating transforms data into representations of T1, T2, T3…

I don't mean to give the impression that temporal order is the only sort of ordering which counts as counting. Any ordering might do. An ideal reasoner might have many ways to count; perhaps humans have many ways to count. Time is only a clear and obvious example.

*Naming* is another form of counting. When I baptize an object, or a natural kind, or a child, and assign a name to that child, what I am doing is counting a certain word – the name – as a representation of that child. The name refers to the child, the name is connected to the child by our *counting* the name as referring to the child. There may be certain characteristic properties associated with a name, but it is not the properties associated with the representations which refer to the child by the name. Suppose one held it was the bundle of characteristic descriptions which referred to the child, since the child held the properties referred to in the description. Aside from the obvious difficulties identified by Kripke, this only pushes the issue back one step – for how did the predicates come to refer to the properties?

All of these ways of counting so far we might call *making a count*. It is from making a count that we develop the notion of numerical identity. A demonstrative count directly refers to a particular object. given that two demonstratives on two occasions can co-refer, we say the two things are numerically identical – they are one thing. a demonstrative at one time refers to itself – this is also numerical identity. Numerical identity is part of logic but it is not something mechanically derivable from the other axioms of logic. It is not mere assignment of rows on a truth table

(and in fact is needed in order to assign rows on a truth table in a meaningful way). The numerical identity facts tell us precisely what counts as one thing.

**8.3.2 Keeping Accounts.** A different sort of counting might be called *keeping accounts*. Making a count involves assigning names, asking what they refer to, and asking whether the things we have so named refer to one another. Having done so, we might then ask if certain things might be related to one another in some way, *assigned* to one another – we might ask what something *has* or possesses, what its "properties" are, whether an object has a given property in its account or not. (We might in turn ask about the properties of properties, and so on.) It is within the practice of keeping accounts that it becomes necessary to talk about truth. We can create a representation of an object-property pair (a name which refers to each) and then ask whether what is assigned in the representation belongs to the object or not. If we line them up and if the representation matches what it represents, we say they are accurate, veridical, or true.

What is it to match? Again, it is no mechanical procedure to match up a representation to reality. It involves another sort of counting activity: coming to a judgment on whether the representation counts or not. The notion of a *truthmaker* is important to keeping accounts. A truthmaker is that in virtue of which a representation counts as matching. Again, this "in virtue of" is a kind of asymmetric ordering and a kind of counting and not mere calculation. The notion of *qualitative identity* is also connected to the activity of keeping accounts. We say two things are qualitatively identical based on a comparison of what is in their accounts.

To call counting an 'activity' or 'coming to a judgment' is not to suggest that there is no right way to count. There is an *a priori* correct way in which to count. But it is the sort of *a priori* which is not mechanical.

**8.3.3 Giving an Account.** The notion of *ordering* is central to counting. Events may be ordered by the time at which they occur. Objects may be ordered in terms of which are parts of others. Properties and essential natures can be ordered in terms of which ground the others. Possible worlds can be ordered in terms of their nearness or distance given some similarity relation, perhaps in terms of how one should have expected things to go if this or that hadn't occurred.

A *causal* ordering is one such ordering which ranks worlds, explaining the future in terms of the past. A *teleological* ordering is a different sort of ranking of worlds, explaining the past in terms of the future – not "backwards causation" or causation in reverse, but the reverse of causation. Varieties of supervenience and ontological dependence are also sorts of orderings which rank worlds or entities within worlds.

We might call these varieties of explanation *giving an account*. Since Hume, it has been recognized that the act of giving an account of an event is something over and above merely stating the spacio-temporal locations of objects at various successive points in time. Nonetheless, we shouldn't take this to mean that giving an explanation of an event or some other phenomenon is the sort of thing we do arbitrarily, unguided by reason. Like other forms of counting, it is an *a priori* judgment, but one which is not mere mechanical calculating.

**8.3.4 Holding to Account.** There is a fourth example of counting which we might consider – that of holding someone to account. Having given an account of an

event, and have assigned the event (in some sense) to the accounts of some *person* as responsible for it, we might then wish to count higher or lower (to *praise* or to *blame*) a person in light of some role the person played in our account of what happened. To count higher or lower is a kind of ranking or valuing which goes beyond mere calculation. Holding *responsible* differs in that we rank one in virtue of what that person is responsible for. To hold someone accountable gives us a distinctively *personal* sense of identity, for the purposes of reward, punishment, and so on, which is distinct from the qualitative and numerical senses of identity.

How are our judgments of blameworthiness or praiseworthiness to be made? Again, the fact that these sorts of judgments are not calculable by mechanical procedures should not lead anyone to conclude that they need be arbitrary or unreasoning (though perhaps, in reality, they often are), or that there is no fact of the matter about what judgment one should make. These judgments could be guided by a kind of reasoning which is *a priori*.

## 8.4 Counting and Calculating *a priori*

Both counting and calculating can be counted as types of reasoning and inference which confer *a priori* justification when performed properly. One who knows the procedure and the input can be justified in the output. One who engages in an act of counting or calculating need not further appeal to some empirical basis to justify how the counting or calculating occurred. The procedures which are involved in calculating are those which I have in mind when discussing mechanical-analytic scrutability. Not all that passes as analysis is calculating – some of what passes for "conceptual analysis" sounds more like what I have described as counting

than calculating. An emergent property is one which cannot be calculated from its subvening basal conditions.

The procedures involved in counting are the sort I have in mind when I discuss a sense of the *a priori* synthetic. Counting is the sort of thing which requires a conscious mind to perform – and yet this does not mean that "anything goes" in counting, or that it is equally natural to count in any way one pleases. Some forms of counting produce justified beliefs, and some do not. I can hold the wrong person accountable despite knowing all of the descriptive facts; I can give a worse explanation which lines up with the empirical facts just as well as a good one; I can choose to count two steps backwards on the number line for every step forward. But I am in none of these cases justified in believing the result of my count.

I am not committed to the counting / calculating distinction aligning perfectly with a useful notion of the analytic / synthetic distinction for the purposes of linguistic or conceptual analysis. I am only interested in it offering an explanation of what I mean by "ontologically" analytic or synthetic relations between the natures of things, or the "mechanically-analytic" procedures an ideal reasoner might use in a deduction.

## §9. Conclusions

I have argued in this chapter that the project of cosmic hermeneutics is only relevant to the question of whether or not a phenomenon is Emergent when cosmic hermeneutics is understood as involving ontological analytic scrutability, as opposed to (i) conceptual or sentential scrutability, or (ii) *a priori* or simulative scrutability. This distances emergentism from a dependence on arguments based on our intuitions about the applications of phenomenal concepts (or other concepts for

purportedly emergent kinds), two-dimensional semantics, what's *a priori* for an ideal reasoner, or reasoning from an epistemic gap to an ontological gap.

One might wonder if this isn't stacking the deck too heavily in favor of emergence. Analytic scrutability is a high burden. However, consider three ways in which the paradigm case of emergence – phenomenal consciousness – might turn out to be ontologically analytically scrutable from the fundamental facts of physics:

(a) It might turn out that the natures of phenomenal properties are already present in the fundamental physical things, or that the fundamental things have a nature which is both physical and phenomenal. In other words, panpsychism might be true.

(b) It might turn out that there is a reductive analysis of consciousness in fundamental physical terms, though we haven't found it yet. Once we do find it, insofar as we are right in thinking that our concepts align with the natures of things, we'll have reason to think consciousness is ontologically analytic.

(c) It might turn out that, in the case of phenomenal consciousness, our concepts for the world simply don't line up with the true natures of things, and so we are unable to understand how the phenomenal is analytic from the physical.

On the other hand, consider three ways in which consciousness might turn out to be ontologically emergent – two of which are *not* part of the standard conception of emergent consciousness.

(d) It might turn out that there is some *a priori* inscrutable set of conditionals which link up conscious properties to their physics basal conditions – that is, consciousness might be strongly emergent.

(e) It might turn out that conscious properties are *a priori* scrutable by an ideal reasoner given the basal conditions in physics – even though they aren't scrutable for us – because an ideal reasoner has access to a kind of *a priori* reasoning which justifies the conclusion that something with neurological states like ours is conscious. Perhaps this could be something like an extremely powerful version of the reasoning which justifies our own belief in other minds. In this case, consciousness would be more weakly ontologically emergent.

(f) It might turn out that there is a certain perspective or stance such that, were an ideal reasoner to see all of the physical facts from it, the ideal reasoner would put himself in our shoes, and *simulate* all of the experiences we are having by undergoing them himself – perhaps by rewiring his own brain (or whatever subvening base he has for his own consciousness) to match up with our own brains, and seeing what happens. This would involve a kind of simulative scrutability of consciousness given a knowledge of our brain states – but consciousness would nonetheless be ontologically weakly emergent. If there was no analytic way to carry out the deduction instead, consciousness would still be weakly ontologically emergent.

So, it is no challenge to emergentism in the case of consciousness to suggest that (e) or (f) might be the case, since both are cases in which ontological emergence would be true. Notably, this also opens the door to the possibility of emergence for a number of other higher-level phenomena in the world, which are *a priori* scrutable from a microphysical base, but for which the deduction appears to require ontologically synthetic premises, much like (e) or (f). Biological functions, for example, are arguably more like (e) or (f) than like (b).

Consider that, in some senses of "reducible", biology, physiology, forestry, psychology, sociology, economics, etc., are all reducible to physics – the sense in which they are necessitated by physics, ideally *a priori* deducible from physics, and permitting of a functional reduction into physical terms. However, there is also a sense in which these fields might *not* be reducible to physics: it is not likely that they can be given a definition in physical terms on the model of (b) on which they follow by mere mechanical procedures from the truths of physics. If this is the case, some of the distinctive natural kinds in these fields might be considered weakly ontologically emergent on my account. This suggests that instead of a one-time event in nature which occurs only in the case of consciousness, emergence might be a widespread feature of the natural world.

In the next three chapters, I will elaborate on this thought that the world is full of examples of emergence. Although phenomenal consciousness seems to be the most plausible, clear case of emergence, I believe focusing on weaker cases might help address the worry that emergentism about phenomenal consciousness is too *ad hoc* to be a satisfying explanation for our experiences.

Chapter 5

# THE FAMILIAR SMELL OF EMERGENCE

## §1 Introduction

A good explanation has the virtue of applying across a variety of situations, not simply the one at hand. An explanation smells fishy when it applies to one situation only. "The bulb burnt out" is a better explanation for why the lamp won't turn on than "this lamp has *dyslampia*", defined as "the condition which arises if and only if it is today and this particular lamp is not functioning." *Dyslampia* adds nothing; it has no explanatory power.

Is ontological emergence a good explanation for consciousness? The emergentist[295] holds that the mere supervenience of conscious states on physical states does not explain consciousness – some further explanation is needed. The explanation offered is that consciousness *emerges* from the physical, yet is not identical to it, but has a nature which is something over and above the nature of the physical.

In this chapter, I will discuss the most historically influential, widely-cited, and *prima-facie* compelling objection to emergentism: the objection of J. J. C. Smart (1959) that emergentism fails to mesh with a scientific perspective on the world. While Smart's objection can be interpreted in several ways, its underlying concern is

---

[295] "Emergentism" can be applied to a whole range of mutually inconsistent views. In this chapter, I will be using "emergentism" to denote only the form which I have been defending, a kind of "minimal" emergentism which accepts some form of mental-physical supervenience, but maintains a difference in essential nature.

that emergentism is an *ad hoc* explanation which applies to only one case: phenomenal consciousness. So long as consciousness has been considered as the only plausible case of ontological emergence, Smart's objection has proven difficult to argue against.

However, the recent rebirth of emergence as an explanation in the special sciences gives us reason to reconsider Smart's objection. Emergence has become a concept with wide explanatory currency across the sciences at many different levels. These cases are generally categorized as superficial or "weak emergence", irrelevant to the ontological "strong emergence" alleged to happen in the case of consciousness. On the contrary, I believe that philosophers have been too quick to dismiss the use of "emergence" in the sciences as a *merely epistemic* claim. I will argue instead that cases of "weak emergence" may equally be cases of *ontological* emergence: cases where the emergent phenomenon has a nature which is over and above that of its subvening base. On my account, there are many things which are ontologically emergent in our world, and many kinds of emergence – emergent consciousness is only one of them.

If it is reasonable to regard cases of weak emergence in the special sciences as ontologically significant, then emergence ceases to be an *ad hoc* explanation for consciousness, providing an answer to Smart's objection. I will not argue in this chapter that particular cases of weak emergence in the special sciences *are* reasonably regarded as ontological – I do this elsewhere, in Chapters VI and VII. Rather, my task here is to demonstrate how weak emergence might be regarded as

ontologically significant, and how regarding it in this way would remove the most significant challenge to emergentism generally.[296]

I will begin in section 2 by laying out Smart's objection to emergent laws. In section 3, I will attempt to tie together the so-called "weak" sense of "emergence" used in the sciences and the "strong" sense of "emergence" used in debates between dualists and physicalists about the mind.[297] In section 4, I will give an overview of my reasons to think that ontological emergence is a widespread, regular phenomenon in nature and that emergence-like explanations are common in the special sciences – a case which I will build in more detail in chapters VI and VII. In section 5, I will show how these considerations offer a rebuttal to Smart's objection.

## §2 Smart's Objection

### 2.1 Overview

Emergentists have traditionally been committed to the existence of "trans-ordinal laws"[298], laws relating low-level physical properties to high-level conscious properties. These laws allow us to reliably predict conscious states on the basis of

---

[296] Other challenges include concerns about the coherence of emergentism, problems involving causal overdetermination and downward causation, and objections to drawing metaphysical conclusions from conceptual analysis. I address these challenges in chapters II, III, and IV, respectively.

[297] Chalmers (2006) distinguishes "strong" emergence (the notion used in philosophy of mind) and "weak" emergence (the notion used in the contemporary sciences and complex systems theory), citing Bedau (1997) for the distinction.

[298] Broad (1925), 77-78

brain states – to know that activity in one part of the brain indicates that a subject is having a certain kind of experience – even though brain states and conscious states are not identical.

In "Sensations and Brain Processes" (1959), Smart observes that laws relating high-level to low-level phenomena would be completely unlike the fundamental laws which modern physics has discovered. Laws like gravity and electro-magnetism are explanatorily simple because they apply equally to objects at all levels, rather than relating higher to lower levels. The emergentists forget how unprecedented their laws would be in comparison to the laws we know about. They would be "nomological danglers."[299]

Identity theories are simpler and ontologically cheaper than emergent theories, and they don't smell quite so fishy. Identity theories do not require non-physical natures, and never leave us having to accept consciousness with "natural piety."[300] They do not require a complicated backstory for the correlations we observe between brains and minds, invoking fundamental laws that link low-level physical phenomena with high-level non-physical phenomena.

Emergence seems to invoke an entirely novel kind of explanation (emergence) for one lonely explanandum (consciousness) and no other. The physicalist recalls that emergent vitalism in biology and emergentism in chemistry were long ago disproven, and *nowhere else* in our study of the world have we found the sorts of

---

[299] Smart (1959), 144

[300] The expression was famously applied by the early British emergentist Samuel Alexander (1920), though it has its origins in Wordsworth's "The Rainbow."

emergent laws or explanations that we are supposed to accept in the case of consciousness. Emergent laws are just plain *weird*.

An immediate reply to Smart's objection might be that it doesn't apply to emergence with metaphysical necessity. On this interpretation, Smart's objection was historically directed at traditional dualism, in which there were nomologically necessary but metaphysically contingent trans-ordinal laws (or "nomological danglers"), but doesn't apply to metaphysically necessary emergence. However, one can reformulate Smart's objection to apply even in the case of metaphysically necessary emergence. Consider Hume's dictum, that there are "no necessary connections between distinct essences." On one interpretation of "distinct essence", a case in which one thing is neither identical to nor grounded in another qualifies it as having a distinct essence. On my account, emergentists deny the grounding of emergent phenomena in physical phenomena. So, on my account, metaphysically necessary emergence would violate Hume's dictum, where the motivation for applying Hume's dictum is the same as the motivation for Smart's denial of "nomological danglers": namely, metaphysically necessary physical-phenomenal bridge laws which aren't grounded in physics have a suspicious smell to them, and are unlike anything we encounter elsewhere. I will take Smart's objection to apply equally to metaphysically necessary emergence, even if the historical Smart might not have been concerned by it.

A second immediate reply to Smart's objection is what we might call the *weirdness reply*. The dualist might say: *of course* laws linking complex physical states to consciousness would be unique, unprecedented, and just plain *weird*. For dualists, consciousness truly is a unique and mysterious thing, something that

deserves to be treated as a special case if anything does. But the weirdness reply does not go very far. It only reveals the physicalist's and the dualist's differing initial prejudices about the relative "weirdness" of consciousness. It does nothing to assuage the physicalist's concern about the apparent dissonance between emergentism and contemporary science.[301]

Why think emergentism is contrary to contemporary science? In considering J. J. C. Smart's classic work, there seem to be at least four possible interpretations of his objection to emergentism: (i) as an application of Occam's Razor, (ii) as a kind of meta-induction of successful reductive explanations elsewhere, (iii) as the "nomological danglers" objection that fundamental laws only relate fundamental particles, or (iv) as the concern that there is something spooky or suspicious-smelling about emergent laws. I will lay out each interpretation here.

## 2.2 Occam's Razor: Two Versions

Smart writes:

> Why do I wish to resist this suggestion? Mainly because of Occam's razor. It seems to me that science is increasingly giving us a viewpoint whereby

[301] For example, the two-dimensional semantics arguments of Chalmers (2010) depend on the reader sharing the intuition that zombies are conceptual possibilities and consciousness is a very strange phenomenon. While I share these intuitions myself, I worry about making them the basis of an argument for an ontological difference, when many people (philosophers and non-philosophers alike) report not having these intuitions.

> organisms are able to be seen as physicochemical mechanisms: it seems that even the behavior of man himself will one day be explicable in mechanistic terms. There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. [302]

The emergentist claims that emergent phenomena have a different essential nature than physical phenomena do. But the emergentist also acknowledges that emergent phenomena are entirely dependent upon their physical bases, that they can be manipulated through changes in their physical bases, and that they cannot be manipulated through any means without producing changes in their physical bases. In some sense[303], emergent phenomena are *necessitated* by their physical bases, and they *supervene upon* their physical bases. If the emergentist is willing to accept all of this, then why make the further claim that emergents are, as an ontological matter, a different kind of thing entirely from their physical base? Why admit a new category into our ontology – the immaterial psyche – when everything else in the world (including human behavior) can be described successfully in a purely mechanical and physical manner?

---

[302] Smart (1959), 142

[303] Which sense of *necessitated* and thus of *supervenes upon* the emergentist should use is disputed (see Chapter 2), but the sort of emergentism I defend in this chapter is meant to be compatible with metaphysical necessity.

Occam's razor advises us not to multiply entities beyond what is necessary to give a sufficient explanation of the phenomena we experience. One way to interpret Smart's use of Occam's razor would be that he advocates this principle:

> (OR1) If there are two theories of some phenomenon which provide the same[304] explanatory power, then, all other things being equal, we should prefer the theory which entails the fewest commitments to novel properties or entities not occurring in our other theories.

The consequent of (OR1) conflicts with the emergentist's commitment to the ontological novelty of emergent properties or entities. Of course, there is room for the emergentist to debate the antecedent of (OR1). To have equal explanatory power, two theories must be equivalent under rational consideration, not merely *verificationally* equivalent (i.e., equivalent with respect to all empirical tests). The emergentist will hold that emergence *does* provide more explanatory power than physicalism: the ontological gap explains why there is an explanatory gap. The emergentist and the physicalist are thus locked into in a debate over the antecedent of (OR1).[305] The physicalist's position is enhanced by the success of physicalist

---

[304] I assume that in the case Emergentism provides *less* explanatory power and more commitment to novelty, we should not prefer it over physicalism; (OR1) applies in the case where Emergentism provides equal explanatory power.

[305] An emergentist might also reject the consequent of (OR1), but this seems like a weak move to me, and I will not explore this option further.

explanations for previously mysterious phenomena: the triumph over vitalism, for example.

Suppose that a physicalist does not accept that non-empirical considerations, or considerations about the "natures" or "essences" of things, should be relevant to theory choice. One should recognize that this as form of *anti-essentialist* physicalism.[306] The anti-essentialist physicalist cannot appeal to principle (OR1). For the anti-essentialist, if two theories make the same predictions with respect to all empirical tests – if they are verificationally equivalent – then we should remain neutral and follow a principle of *tolerance* with respect to the entities and properties invoked by each. Which theory we adopt in this case becomes a matter of convention. Occam's razor could alternatively be understood as:

> (OR2) If there are two theories of some phenomenon which are verificationally equivalent, then, all other things being equal, we should

[306] I understand physicalism proper to be committed to the claim that all real concreta have a wholly physical nature, whereas anti-essentialist physicalism is a view about the empirical adequacy of physics (perhaps in an idealized future) for generating the truth values of statements. Emergentism for anti-essentialists is the view that some sentences – those about qualia – have a truth value, and yet are not analytically derivable from any set of sentences in micro-physics. An essentialist makes the further inference from this that qualia have a non-physical nature.

> prefer the theory whose explanations are most congruent with those used in our other theories.[307]

Here, the consequent of (OR2) offers a suggestion for how to decide between verificationally equivalent theories: congruence with other theories which are already in use. Suppose that physicalism is congruent with the purely mechanistic explanations utilized in our other scientific theories, whereas emergentism proposes a new type of explanation (emergence). Physicalism seems not to require the same drastic mutilations to our interconnected beliefs that emergentism seems to require, and so it is the preferable theory.

For what it's worth, J. J. C. Smart was no verificationist. His identification of sensations and brain processes was a full-fledged metaphysical claim in the "Australian" tradition: a sensation by nature is nothing over and above what a brain process is by nature.[308] Smart endorsed the essentialist version (OR1) rather than the anti-essentialist version (OR2) of Occam's razor:

> If it be agreed that there are no cogent philosophical arguments which force us into accepting dualism, and if the brain processes theory and dualism are equally consistent with the facts, then the principles of parsimony and

---

[307] This is more of a principle of Conservativism than the traditional form of Occam's Razor.

[308] See Polger (2011)

simplicity seem to me to decide overwhelmingly in favor of the brain-process theory.[309]

Thus, Smart's own view was clearly aligned with (OR1), under which the emergentist's task is to demonstrate that emergentism *has greater explanatory power* than physicalism, but "explanatory power" is understood to include rational considerations ("cogent philosophical arguments"), not simply empirical ones. A philosopher who holds that the only "explanatory power" relevant to Occam's razor is the power to predict empirical data must align with version (OR2) instead, on which the emergentist's task becomes showing that emergentism is *more congruent with our existing theories of the world* than a mechanistic theory.

### 2.3 Meta-Induction

Smart continues:

> So, sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together-roughly, biology is to physics as radio-engineering is to electromagnetism) except the occurrence of sensations seems to me to be frankly unbelievable.[310]

---

[309] Smart (1959), 156

[310] ibid.

While we may not have a sufficient explanation of consciousness in physical terms, we may have reason enough to believe that such an explanation in principle exists. We have a *physicalist portrait* of the world, and nearly all of the canvas has been filled in. At this point, we have reason to expect the blank spots left on the canvas will also be filled in with explanations from physics. We can perform a kind of *meta-induction*: in a universe in which we are continually finding satisfying physical explanations for a host of phenomena, where fields like chemistry and biology are reducible to physics, why should consciousness be the lone exception? Even if we don't have such explanations yet, we have reason to think that we will one day get them. Smart might be understood as adopting this principle:

> (MI) If there is some property that many of our established theories have and not many lack, then when considering non-established theories, we have a reason to prefer a theory which has that property over a competing theory which lacks it.

The physicalist will argue that many of our established theories of the world have the property of being either theories of physics or theories which are in principle fully explicable in the terms of physics. Emergentism is a theory which is not even in principle fully explicable in terms of physics. So, we have a reason to prefer competing physicalist theories over it.

If we find it a challenge to understand how physical structures could ever give a satisfying explanation of consciousness, why appeal to emergence for the

answer – as opposed to our own ignorance? Perhaps the simplest explanation of why any possible explanation of consciousness in physical terms seems inevitably unsatisfying isn't something in the world but something in us: perhaps *we* suffer from some epistemic limitation or cognitive defect.

On this reading, the emergentist's task is to show that *many of our well-established theories are not explicable in terms of fundamental physics*, and that the physicalist portrait of the world is inaccurate.

## 2.4 Nomological Danglers

> Such sensations would be "nomological danglers," to use Feigl's expression. It is not often realized how odd would be the laws whereby these nomological danglers would dangle. It is sometimes asked, "Why can't there be psycho-physical laws which are of a novel sort, just as the laws of electricity and magnetism were novelties from the standpoint of Newtonian mechanics?" Certainly we are pretty sure in the future to come across new ultimate laws of a novel type, but I expect them to relate simple constituents: for example, whatever ultimate particles are then in vogue.
>
> I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles) all put together for all the world as though their main purpose in life was to be a negative feedback mechanism of a complicated sort. [311]

[311] ibid., 142-143

Smart's article is often remembered for his use of Feigl's expression, "nomological danglers", as an objection that fundamental emergent laws seem out of place when compared with the fundamental laws of physics. The objection is not that *all* laws in the sciences must be like those of fundamental physics, relating simple constituents. Certainly there are well-established laws of economics and sociology which do not relate particles to one another. There are also plenty of laws that operate between higher and lower levels: there are predictable relationships between heightened expectations, economic reversals, and political revolutions.[312] Smart's objection can't mean that fundamental laws *only* operate at the fundamental level: the laws of gravity work as well for planets as for particles. It also can't mean that all fundamental laws must be general and operating at all levels: while gravity and electro-magnetism operate at all levels, the strong force operates only between quarks within the nucleus of an atom.[313] Instead, I interpret Smart as advocating a principle like this:

> (ND) We should only accept laws which either (i) are fundamental, relate fundamental particles to fundamental particles, or relate higher-level objects only through relating fundamental particles to other fundamental particles, or (ii) are non-fundamental and fully explicable in terms of the laws in (i).

---

[312] This is commonly known as the "J-Curve" theory of revolutions. See Davies, J. C. (1962)

[313] See "strong force", Encyclopædia Britannica Online Academic Edition (2013)

Gravity and electro-magnetism, like the laws relating the strong force within an atom, relate fundamental particles to other fundamental particles. Gravity and electro-magnetism also relate non-fundamentals, but only in virtue of relating the fundamentals of which those non-fundamentals are composed. There are a plentitude of laws which do not relate fundamental particles, like the laws that enable predictions about chemical reactions, the weather, birth rates, extinctions, voting behaviors, and how much poison it takes to kill a rat. But these laws aren't supposed to be fundamental, or inexplicable in physical terms even in principle. In contrast, the laws of the emergence of consciousness are supposed *both* to be fundamental *and* to relate fundamental particles to complex wholes without relating fundamental particles to other fundamental particles.

The emergentist's task, then is to argue that (ND) should be rejected, because *many of the laws we rightly accept fail to meet either condition* (i) or (ii).[314]

## 2.5 A "Queer Smell"

Recall that many of the early British emergentists could have resisted (OR1), (OR2), (MI), and (ND).[315] C. D. Broad believed the emergence of consciousness was

---

[314] The emergentist could also make an appeal for the case of consciousness to violate condition (i), a version of the "weirdness reply" mentioned earlier.

[315] The British Emergentists could also have more easily resisted the antecedent of (OR1) and (OR2), since on their view there were novel emergent *forces* which interacted with lower forces, with the result that they made distinct empirical claims from the non-reductive physicalist, and hence could hold out hope for the

*not* the only case of emergence, since he accepted emergent vitalism in biology and entertained the possibility that chemistry was emergent.[316] Emergence about consciousness already fit well with their other theories about the world. There were enough possible exceptions to mechanism at the time to block the meta-induction. They had no reason to think fundamental laws couldn't relate non-fundamentals. Soon after emergence about chemistry and emergent vitalism were disproven, emergence about consciousness went into hibernation.

A physicalist may see a lesson for emergentists here. Even if an emergentist *could* develop an account of the world on which emergentism about consciousness didn't conflict with our existing scientific accounts of the world, but in fact was in harmony with them – blocking the arguments in (OR1), (OR2), (MI), and (ND) – even *still* there would remain something suspicious about emergentism. Smart writes:

> Such ultimate laws would be like nothing so far known in science. They have a queer "smell" to them. I am just unable to believe in the nomological danglers themselves, or in the laws whereby they would dangle. If any philosophical arguments seemed to compel us to believe in such things, I would suspect a catch in the argument. [317]

---

empirical verification of their theory. As discussed earlier, the brand of emergentism I am defending does not advocate for the existence of novel emergent forces.

[316] Broad (1925), 72-79

[317] ibid., 143

Smart's objection could be rephrased as this principle:

> (QS) Given any argument for a theory which requires the existence of non-physical properties, it will always be more likely that there is an error in the argument (a false premise, or an invalid move) than it will be that the conclusion is true.

Whereas the prior objections did not presuppose physicalism, objection (QS) does presuppose physicalism. Objection (QS) is the other side of the coin to the dualist's weirdness reply that consciousness is just plain *weird* and no physical account (no matter how moderate or non-reductive) will ever be weird enough to explain it. The physicalist may intuitively link up "non-physical properties" with mental images of haunted houses and horror movies. For a physicalist of any stripe, any dualist explanation is just plain weird (no matter how moderate or pro-supervenience), and no phenomenon will ever be weird enough to justify believing in it. Like the dualist's "weirdness reply", the "queer smell" objection does little more than emphasize the differing initial prejudices of the physicalist and the dualist. The physicalist thinks dualism smells funny; the dualist thinks, *that's just how consciousness smells*. This may be an unbridgeable difference in philosophy, but it is not a philosophical debate.

However, by relating the use of "emergence" in the sciences to the ontological claims made by emergentists in the philosophy of mind, the emergentist may gain some ground against (QS). The emergentist can paint a picture on which nature is a

place where one should expect surprises: against the background of widespread emergence, emergent consciousness should smell familiar, not spooky.

In the next section, I will attempt to tie together the 'strong' sense of emergence, which holds for consciousness, and the 'weak' sense of emergence, which is supposed to hold for a variety of phenomena in the special sciences. I will begin with the case of consciousness.

## §3 Varieties of Emergence

### 3.1 Ontological Emergence and Consciousness

According to the emergentist, phenomenal consciousness is entirely dependent upon its subvening physical base for its existence, but nonetheless has an essential nature which is something "over and above" the nature of its physical base. Conscious experiences arise from brain states in a law-like way, enabling empirical predictions about conscious states on the basis of neurology. We can predict, for example, that a certain drug will interact with a patient's brain in such a way as to ease that patient's experience of pain. However, the emergentist holds that the relation between the mind and the brain is ontologically and explanatorily fundamental. The correlations are not something any of us could have deduced *a priori,* given only knowledge of what it is to be a brain with a particular structure, and what it is to be a mind undergoing a certain experience. Empirical study was required for us to learn these correlations between changes in the brain and the experience of the easing of pain.

A non-reductive physicalist might accept that there can be no *a priori* deducibility of conscious states from brain states, but hold that this "epistemic gap" is not ontologically significant. On this view, given our limited cognitive resources,

the identities between experiences and brain states are knowable by us only *a posteriori*; nonetheless, token experiences are identical to token brain states. The emergentist dualist concludes from the epistemic gap that there is an ontological gap as well: experiences and brain states are *not* identical. While experiences supervene upon complex physical states, what it is to be an experience is not the same thing as what it is to be a complex physical state.

Notice that the non-reductive physicalist and the emergentist dualist both accept some form of supervenience and at the same time accept a kind of "epistemic gap".[318] They differ on the ontological significance of supervenience and of the gap. In contrast, the reductive physicalist rejects the epistemic gap, and the substance dualist rejects any form of supervenience.

### 3.2 Not that Innocent

Traditionally, "Ontological Emergence" is applied only to the case of consciousness above. While it is widely acknowledged that the concept of "emergence" has a place in scientific discussions (for example, in the science of complexity), this sense of "emergence" is widely held to be ontologically innocent. Mark Bedau (1997) writes:

---

[318] This statement holds true only as I have defined each position for the purposes of this chapter. There are emergentist dualists who deny supervenience, like Hasker (1999). According to Stoljar (2009), many non-reductive physicalists prior to Smart (1959) relied on a more restrictive notion of reduction, which did not explicitly address the issue of *a priori* deducibility.

> An innocent form of emergence – what I call "weak emergence" – is now a commonplace in a thriving interdisciplinary nexus of scientific activity – sometimes called the "sciences of complexity" – that include connectionist modeling, non-linear dynamics (popularly known as the "chaos" theory), and artificial life. [319]

The distinction between weak emergence and strong emergence has been drawn in a variety of ways[320], but, generally, "weak emergence" is supposed to capture those cases in which some form of deduction of high-level facts from low-level facts is in principle possible, but would be surprising or difficult[321], and "strong emergence" is supposed to capture any mysterious cases in which no deduction of high-level facts from low-level facts is possible, even in principle.

Notice that these definitions are in epistemic rather than ontological terms. Strong emergence might not be ontological emergence (as a non-reductive physicalist might hold). Likewise, weak emergence might qualify as *ontological* emergence, not merely a reflection of our epistemic limitations. Consider the following options presented in Table 3:

---

[319] Bedau (1997), 375

[320] See Chapter 1. Also see Chalmers (2006) and Bedau (2008).

[321] e.g., perhaps possible only through simulation – see Bedau (2008)

Table 3

*Reduction and Deduction*

| | *A priori* deducibility by pure analysis | *A priori* deducibility by synthetic bridge | A Posteriori deducibility only |
|---|---|---|---|
| Ontological Reducibility | (A) Analytic Reducibility | (B) Epistemic Weak Emergence | (C) Epistemic Strong Emergence |
| No Ontological Reducibility | X | (D) Ontological Weak Emergence | (E) Ontological Strong Emergence |

A reductivist physicalist might hold that consciousness is an instance of case (A).[322] The non-reductive physicalist holds that consciousness is an instance of case (C), and the emergentist dualist holds that consciousness is instead an instance of case (E).

I wish to argue that many of the cases currently categorized as (B) ought to be categorized as cases of (D). That is, I advocate for a kind of ontological weak emergence, where some high-level phenomena have a *distinct nature* from lower-level phenomena, even though they are deducible from the lower level phenomena. This includes forms of emergence found in complexity science and in the special sciences. On my view, weak emergence may not be so ontologically innocent after all.

[322] Arguably, many reductive physicalists actually hold that consciousness is really an instance of case (B), insofar as they view the reduction as requiring something less than a Nagel-style theory reduction on the Deductive-Nomological model.

## 3.3 Deducibility and Ontology

Why do I think we should take weak emergence to indicate an "ontological gap"? Understanding the answer requires understanding the relationship between ontology and the epistemic test of *a priori* deducibility.

In the last chapter[323], I discussed the use of epistemic, *a priori* deducibility as a criterion for drawing ontological conclusions about whether the nature of one thing is wholly contained in the other. I argued that the emergentist need not be committed to the view associated with Chalmers and Jackson's interpretation of two-dimensional semantics[324] that the ideal *a priori* deducibility or non-deducibility of one set of concepts from another, if performed correctly, offers a guide to the necessitation or non-necessitation of one set of properties from another. Instead, I argued that the better question to ask is not whether A is deducible from B *a priori*, but rather one of *what resources would have to be a priori* for a reasoner to make the deduction of A from B.

I worked under the assumption that the "essential nature" of a thing is something like a "real definition" of the thing – a relation between it and the set of properties in virtue of which it has the identity it has. This is different from the more conventional account on which the "essential nature" of a thing is only a

---

[323] See Chapter 4, "Cosmic Hermeneutics"

[324] See Chalmers, D. and Jackson F. (2001). For a case against this use of two-dimensional semantics, see Soames (2005). For the purposes of this chapter, I will remain neutral on the issue.

collection of its modal properties.[325] For example, Kit Fine (2004) holds that Socrates is necessarily distinct from the Eiffel tower, but on his "real definition" account it is not an essential property of Socrates that he be distinct from the Eiffel tower. The merit of Fine's approach is that it explains why we should take *a priori* deducibility to indicate anything about ontology at all: *if* our concepts accurately reflect the true natures of things, then a deduction of one concept from another through analysis alone indicates that the nature of the one thing is contained in the other.

Under this assumption, I defended a specialized, ontological sense of *synthetic* and *analytic* pairs of propositions – not conceptual analyticity, or analyticity relative to some language like English, but rather a matter of whether the "real definition" of one thing is wholly contained in the other.[326] Thus, while it is synthetic in English that "H2O = Water" is true, and conceptually *a posteriori*,[327] it is *ontologically analytic* that water is H2O.[328] If one thing is "nothing over and

---

[325] Robertson, Teresa, (2008). I adopt Fine's account in Chapter 2, because it is helpful to get out of the objection that emergence with necessary supervenience is incoherent.

[326] "Consider of what would be analytic or synthetic were objects and properties to represent themselves, and the essences of things to be identical to analytic definitions in the language – or else, a language used by an ideal reasoner in a one-to-one correspondence to such a language. There is no reason to think any real language capable of fitting the bill." (Chapter 4, section 7.2).

[327] Putnam, H. (1975)

[328] This is just to say that H2O = H2O.

above" another, it is ontologically analytic given it; if one thing is "something over and above" another, it is only derivable from it given ontologically synthetic bridge statements.[329]

Consider what follows from this approach. It follows that we should regard a phenomenon as ontologically emergent – as having a nature which includes something "over and above" that of its supervenience base – *if and only if it is ontologically synthetic given its subvening base*.

An ontologically emergent phenomenon may be deducible *a priori* from its subvening base, provided that even for an ideal reasoner the deduction would involve an ontologically synthetic *a priori* bridge law; a merely-epistemically emergent phenomenon may fail to be deducible *a priori* from its subvening base for an ideal reasoner who lacks access to some *a posteriori* (but ontologically analytic) identity.

So, the question of whether or not a phenomenon is *ontologically* emergent is not to be answered by considering whether or not it is deducible *a priori* from a supervenience base that necessitates it. After all, were an ideal reasoner given all of

---

[329] Thus, Laplace held that every past and future fact was necessitated by the present facts, but *not* that it was ontologically analytic; a monist holds that all of the particular facts are ontologically analytic given the fact about the cosmos as a whole. The essential properties of the fundamental particles in one part of the world are ontologically synthetic given the essential properties of some other fundamental particles in some other part of the world for the most part, except in special cases involving entangled electrons.

the necessary truths as axioms, it would become deducible quite easily, as Kim (2010) notes.[330] The epistemic question of ideal *a priori* deducibility is relevant to this ontological question only insofar as: (i) we have epistemic access to the whole natures of both phenomena[331], and (ii) we limit the tools in the ideal reasoner's *a priori* deduction to analytic moves. We should ask: what cognitive resources or inferential abilities would have to be *a priori* for the ideal reasoner to perform the deduction?

If an ideal reasoner could perform the deduction given only ontologically analytic truths, the phenomenon is not ontologically emergent; if an ideal reasoner could only perform the deduction if ontologically synthetic premises were made *a priori* for that reasoner, then the phenomenon is ontologically emergent. Thus, even if it were *a priori* for a very powerful reasoner looking at a bundle of firing neurons to conclude that an experience of a certain sort was going on, the ideal reasoner's deduction would involve something more than mere analysis.

Suppose that a phenomenon is deducible *a priori* only by simulation, not by analysis, as Mark Bedau (2008) has argued is the case for the "weakly emergent"

---

[330] See Kim (2010), 85-104; for my response, see Chapter 2.

[331] e.g., in the case of a posteriori necessities, like the identity of water and H2O, condition (i) fails for those living in the age prior to the discovery that water is H2O – they did not have access to the whole nature of water, and so did not know that it was in fact H2O. In some cases, a phenomenon which truly is nothing over and above another may be "emergent" for us in the epistemic sense only because condition (i) does not hold, and so it does not qualify as ontologically emergent.

phenomena studied in complexity science. Or, suppose that a normative or teleological property can be deduced *a priori* from a descriptive one, but only by a kind of perspective-taking[332]: as is the case with functional explanations in biology and collective behavior in sociology.[333] On the account I've offered, these "weakly emergent" phenomena will qualify as ontologically emergent.

In most discussions of "weak emergence", no distinction has been made between ontologically analytic and ontologically synthetic *a priori* transitions. This has resulted in a false sense of confidence that a phenomenon which we suspect to be ideally *a priori* deducible from another (since we so clearly "see" how the one necessitates the other) must therefore be ontologically reducible.

### 3.4 Emergence as a Spectrum

But how do we distinguish the emergence of consciousness from cases of ontological weak emergence? For, even if there are "weakly emergent" phenomena which are also ontologically emergent – patterns of waves in the ocean, for instance – it is not clear that they should be thought of as emergent in the same way that consciousness is supposed to be. The way in which consciousness emerges from brain states seems much more striking, or "stronger", than the "weaker" way in which patterns of waves in the ocean emerge. How are we to distinguish them, if not by calling one ontological and insisting the other is merely epistemic?

---

[332] i.e., if a phenomenon can be deduced *a priori* from behavioral facts only by means of an intentional stance and not by analysis, then the phenomenon is ontologically synthetic given the behavioral facts, *not* reducible to them.

[333] See Chapters VI and VII

Consider that an emergent phenomenon is supposed to be a "genuine novelty": there is something in the nature of the emergent phenomenon which is not contained in the nature of its supervenience base. We could take the relative complement of the base's nature in the emergent's nature: the set of all elements in the nature of the emergent but not in the nature of the base. Call the result the "novelty" of the emergent phenomenon, the part of its nature not contained in the nature of its basal conditions.[334]

The relationship between consciousness and the brain is different from the relationship between patterns of waves and the water in the ocean.[335] If the wave is

---

[334] Suppose that a murder is by nature an unjustified intentional killing. Let P be a complete physical description of a killing K and the brain states and social states surrounding it, and M be the murder corresponding to K. M supervenes upon P. Let J be the fact that K was unjustified, which also supervenes upon P; and let I be the fact that K was intentional, which also supervenes upon P. The nature of M = <P, J, I>. The relative complement of P in M contains the unique elements J and I. So, "J" and "I" are the novelty of M, where M emerges from P.

[335] One might propose instead that the difference is all in the head: something like the subjective sense of surprise each causes in a scientific researcher (Ronald, et. al. (2008)), or the difficulty of predicting or explaining them in a way our minds can easily comprehend. But, like most psychologistic proposals, this raises more questions than it answers. Scientists don't get surprised just because they're feeling playful – what about the phenomenon makes it surprising? A phenomenon which is difficult to deduce and predict on a given basis is difficult for some *reason*.

emergent, then its nature is *almost but not entirely* contained in the nature of the water, for the same wave could not be realized by substance too physically different from water; but an emergentist holds that the nature of consciousness is very different from the nature of the brain, and that it could be realized by a substance physically very different from the brain.

I suggest that we consider the various elements in the nature of a phenomenon as weighted – that some hold a kind of ontological priority[336] over the others, as more integral to its identity than others. This "weighting" might be understood as analogous to David Lewis's notion of counterpart relations. Suppose that *A* and *B* are both essential properties of *S*. We might ask this: given a world in which some object had *A* but not *B*, and another object had *B* but not *A*, which would be *S*?[337] Alternatively, we might ask: given world $w_{-B}$ in which the closest counterpart[338] of *S* lacks *B*, and world $w_{-A}$ in which the closest counterpart of *S* acks *A*, *ceteris paribus*, which world is closer to the actual world, $w_{-A}$ or $w_{-B}$?[339]

---

[336] I have in mind here Fine (1995)

[337] Of course, I hardly mean to suggest that my view commits one to a counterpart theory of identity across worlds, as opposed to accepting true trans-world identity. I only mean that the notion of weighting the relevance of various essential properties to the trans-world identity of an individual should not be seen as outlandish.

[338] Accepting that, since both B and A are essential to S, S exists in neither world.

[339] These two accounts may give different results. I am not offering a complete theory of how to determine the weight of a particular essential property relative to

What would determine the answer to a question like this? What would one essential property more significant to the identity of a thing than another essential property? We ought to find the answer in the *explanatory burden* carried by the property. Consider what it is that a pattern of waves in the ocean explains. Most of what is explained by a pattern of waves in the ocean is explained by virtue of its having the essential property, *being composed of water*. *Being composed of water* is very important to the explanations that pattern of waves in the ocean enters into. On the other hand, the property of *instantiating such-and-such an emergent pattern* has only a small part in these same explanations. It carries less of the explanatory burden than *being composed of water*: a different pattern in the same water might figure in similar explanations. Contrast this with *being this conscious experience* and *being realized as this brain state*. For most of the explanations which my thought enters into, *being conscious* is far more important than *being a brain state*. Very little explanatory burden is carried by my thought's *being this brain state*. In a world in which a similar phenomenal experience occurred without the brain state (say, in a silicon chip) and a similar brain state occurred without similar phenomenology, the counterpart of my thought would be the experience in the silicon chip, not the brain state.

So, in principle, an ideal reasoner might calculate the ratio between the novelty of an emergent phenomenon and its nature as a whole. This ratio would allow the ideal reasoner a method of differentiating varieties of emergence on the

---

others, only a sketch of how it might be done, to make the notion of such a "weighting" more intuitive.

spectrum: a "weaker" form of emergence is one where the novelty is a *smaller proportion* of its whole nature, and a "stronger" form of emergence is one where the novelty is a *greater proportion* of its whole nature. In this sense, consciousness might be said to be a stronger form of emergence than patterns of waves in the ocean, even though both may qualify as ontological.[340]

Of course, I don't think it is possible that anyone could quantify over natures in this manner. There are too many vague cases, and the relationship between our concepts and the natures of things is too uncertain. However, there may be clear cases – like the difference between the emergence of consciousness and the emergence of a wave – where the framework I have given allows a principled way to distinguish "stronger" and "weaker" forms of ontological emergence.

To summarize, the clarity of the emergentist's ontological claim has suffered greatly from identification with an epistemic question – whether a phenomenon is or isn't *a priori deducible* or predictable on the basis of a knowledge of its subvening base. Because it's plausible that all of the phenomena in the world besides the mind (normative or otherwise) are *a priori* deducible at least in principle from a knowledge of their physical constituents (apart from the odd cases of *a posteriori* necessities), ontological emergence has been reserved for the mind – a novel case of the emergence of novelty. However, as I will discuss in the next section, I believe there is a sense of "emergence" which applies to whole host of phenomena at many different levels of the cosmos. Far from having a "queer smell", novel trans-ordinal

[340] The use of "weaker" and "stronger" here do not match up with the distinction between "strong emergence" and "weak emergence" offered earlier.

laws should strike us as a familiar part of nature. The real world is a jungle of nomological danglers, and the special sciences are teeming with entities whose natures are not fully exhausted by an account of their physical composition. Emergence at lower levels of nature is not so dramatic as is the case of consciousness. But it establishes a pervasive pattern: a backdrop against which the emergence of consciousness fits especially well.

## §4. Instances of Ontological Weak Emergence

So, do we have any reason to think that cases of ontological weak emergence actually occur? Are there genuine cases, apart from phenomenal consciousness, where it is plausible that a phenomenon is ontologically synthetic given its subvening base? I will give a preview in this section of why weak emergence might apply to a wide array of phenomena.

Examples of phenomena which have been called "weakly emergent" include: the behaviors of cellular automata[341] and other "Artificial life"[342] phenomena in the computational sciences; the waves in sand dunes or bodies of water, the formations of schools of fish, and the mounds constructed by termites[343]; biological functions and norms of success and failure; and the relationship between micro-level and macro-level behavior in sociology[344], especially collective social entities and kinds[345]. While

[341] Bedau (2008), 162

[342] Assad and Packard (1992), 231

[343] Crutchfield (2008), 269

[344] Sawyer (2001)

[345] Blau (1977, 1981); Bhaskar (1982); Archer (1995);

all of these phenomena unquestionably supervene upon their respective underlying physical phenomena, they cannot be deduced by mere analysis from it.

All of these examples have something interesting in common: they are all cases where some *normative* property helps define what the emergent phenomenon is. The normative properties in the world are not limited to the sort discussed in meta-ethics. A normative property is one which relates some actual-world state of affairs to some non-arbitrary possible state of affairs. Norms need not have prescriptive or moral force to qualify as norms. For example, a computational *pattern* – or a pattern of waves in the ocean – is a normative property which relates the actual output of the computer simulation to some possible way in which the simulation "ought" to go in order to keep its pattern. A biological *function* or *role* is a normative property which relates what an organ or organism actually does to what it "should" do (in a weak sense of should). A *natural grouping*, whether a species or a social group, marks off non-arbitrary boundaries in order to create a whole that is more than the sum of its parts. When an abstract mathematical function is applied to a concrete situation – like a macro-economic formula – the actual world is being compared to how the model says things ought to go. The special sciences would be impossible without natural norms of this sort.

Weakly emergent properties have some normative aspect to their nature which is not present in their subvening base. This is why they are deducible *a priori* from their base, but not by analysis. Normative properties are widely acknowledged to be synthetic – one can't derive an 'ought' from an 'is'. They are widely held to be *a priori*. Within meta-ethics, this recognition has lead G. E. Moore and others to accept a kind of non-naturalism about ethical properties: these properties supervene

on the natural, but they have a different nature than the natural. I believe a similar view should be taken towards natural norms: they supervene upon the physical, but the normative aspect of their nature is not present in physics.

The alternative – a kind of anti-realism about natural norms – should not be taken lightly. Researchers from biology to sociology do not think they are inventing "patterns" or "functions" or "groupings" as convenient fictions. They do not believe they are reifying projections of their own psychology, but that these normative "patterns" and "groupings" and "functions" are real phenomena which their research endeavors to describe. They are ineliminable from theories which explain and predict their data.

Of course, at least in principle, cases of weak emergence are explanatorily reducible to their subvening base. But, insofar as they are genuinely normative, they should not be regarded as *ontologically* reducible: there is some sense in which even the pattern of waves is something over and above the water that composes it.

Thus, ontological emergence provides a plausible account of how complex phenomena across a wide variety of fields of inquiry can supervene upon some set of basal conditions and yet qualify as something "over and above" these conditions, or a "genuine novelty" which is not wholly contained in the nature of its subvening base.

I will provide further argumentation in Chapters VI and VII for the regarding weak emergence as pervasive and ontologically significant. However, for the remainder of this chapter, I would like to focus on how ontological weak emergence – if I am correct about it – could provide a response to the concerns raised in Smart's original objection.

## §5. Reply to Smart

### 5.1 Escaping Occam's Razor

Consider, then, the two objections which we earlier labeled (OR1) and (OR2):

(OR1) If there are two theories of some phenomenon which provide the same amount of explanatory power, then, all other things being equal, we should prefer the theory which entails the fewest commitments to novel properties or entities not occurring in our other theories.

(OR2) If there are two theories of some phenomenon which provide the same amount of explanatory power, then, all other things being equal, we should prefer the theory whose explanations are most congruent with those used in our other theories.

To reply to (OR1), the emergentist has the task of defending the view that emergence has more explanatory power than non-reductive physicalism. If emergence would help explain a vast array of phenomena to varying degrees, then emergence may in fact have more explanatory power. Consider that emergence does not conflict with any of the explanatory power of physics, since the sort of emergence I have been arguing for accepts supervenience and denies emergent forces. Consider also that the non-reductive physicalist seems likely to accept that many of the phenomena which I am calling emergent are, at the very least, not theoretically reducible. The non-reductive physicalist has the burden of explaining why, even though our theories of higher-level phenomena are not reducible to our theories of

lower-level phenomena, nonetheless we ought to regard the *natures* of higher-level phenomena as *in principle derivable* from the natures of lower-level phenomena using *analytic* intermediary premises which *we have no access to*. The non-reductive physicalist needs a physicalistically-acceptable explanation of why the supervenience occurs but the reduction doesn't.

The emergentist can explain the lack of reduction as one case of many in the natural world where a higher-level phenomenon has a distinct nature from the lower-level phenomena which it supervenes upon. By relating a variety of phenomena, instead of just one, emergence becomes less *ad hoc* and more powerful as an explanation. Of course, emergence admits a larger ontology than physicalism, but this is a move justified in exchange for other theoretical virtues.

Of course, these are purely rational considerations. Emergentism admittedly makes the same predictions with respect to all possible empirical tests as non-reductive physicalism: they are verificationally equivalent. An anti-essentialist[346] could hold that this suffices to establish equal explanatory power – in which case, the discussion should turn to (OR2). However, an essentialist who backs (OR1) cannot: for the essentialist there is a further category of explanation over which the emergentist and physicalist have a genuine (not merely verbal) dispute: an

[346] By this, I mean a philosopher who accepts only explanations in terms of material and efficient causes, not in terms of essential natures, holding to either an eliminativist or deflationary account of modality, and is thereby able to hold that verificationally equivalent theories are in all respects explanatorily equivalent.

ontological explanation in terms of *what it is to be* the kind of thing that the emergent phenomenon is.

Someone might object that this sort of explanation is not significant enough to justify increasing the size of our ontology, particularly if it doesn't make a practical difference. But it does make a practical difference! Ontological questions help guide our choice of methodology, among other things: they tell us whether the *A*s should be investigated by means of investigating the *B*s, or the *A*s should be investigated in their own right. An emergentist believes the nature of consciousness is immediately apparent to us first-hand, and our inability to deduce this nature from the nature of physical phenomena is *better explained by* an ontological gap between physical and conscious phenomena. The kinds of explanations which consciousness figures in are not the kinds of explanations which brains could ever figure in; hence, consciousness should primarily be investigated in its own right. An essentialist physicalist, on the other hand, must believe that the physical nature of consciousness is not immediately apparent to us when reflecting on our own conscious states – its true (physical) nature is hidden. The kinds of explanations which consciousness figures in could ultimately be filled by brains; hence, the best investigation of consciousness will primarily be through the investigation of brains.

To reply to (OR2), the emergentist had the task of showing that emergence was more congruent with our theories of the world than mechanism. Again, if emergence is not an *ad hoc* response to consciousness, but instead emergentist explanations already have a home in many of the natural and social sciences, then emergentism as a theory is nicely congruent with many of our other theories of the world in a way in which mechanistic explanations are not. Of course, this response

depends upon the likelihood that the natural and social sciences will continue in the direction of adopting emergentist explanations. I've provided an account of why natural and social scientists should not be as hesitant to accept ontological emergence, but I can't predict that they will accept it. It remains to see what forms actual future explanations in these sciences will take.

## 5.2 Letting the Laws Dangle

At this point, a physicalist may respond that regarding as emergent such a wide variety of phenomena commits us to a wide variety of nomological danglers, in violation of principle (ND):

> (ND) We should only accept laws which either (i) are fundamental and relate fundamental particles to fundamental particles, and relate higher-level objects only through relating fundamental particles to other fundamental particles, or (ii) are non-fundamental and fully explicable in terms of the laws in (i).

But why accept (ND) in the first place? Why shouldn't nomological danglers "dangle"? It can't be that the reductivist thinks there are no higher-level laws of biology or sociology, period. Rather, it is that the reductivist believes all of these laws are non-fundamental, derivative from the laws of physics. It is true that this conflicts with emergentism in biology or sociology, since an emergent phenomenon (like a biological function) would be related to its supervenience base by a fundamental law which was not itself a law of physics. However, I believe that there

are two ways the emergentist can reply to the objections of the reductivist on this point.

First, the emergentist can hold that (ND) was only plausible when we were certain that emergence didn't happen at any lower level than the mind itself. If the widespread appeal of the concept of emergence tells us that the world really is a jungle of nomological danglers, then the better move is to reject (ND) rather than reject emergence.

Second, the emergentist can point out that, whereas the fundamental physical laws are causal laws which relate events over time, the synthetic "bridge laws" which the emergentist is committed to are all synchronic laws which relate a supervenience base at one time to an emergent phenomenon at the very same time. While they are "fundamental" in the sense that they are not reducible to purely physical laws, it is not as though the emergentist is asserting that they belong in the same category as the fundamental laws of physics. They are laws about a very different sort of thing. We should not expect (metaphysically necessary) fundamental synchronic laws to be like (metaphysically contingent) fundamental diachronic laws, so we should not compare emergence laws with the laws of gravity or electro-magnetic charge and then react with surprise at the difference. Calling both of them by the same term, "law", is a way of expressing their epistemic position as intermediate premises allowing derivations from statements of initial conditions, not to put them in the same ontological category.

### 5.3 Resisting the Meta-Induction

As I near the conclusion on this chapter, I should quickly review the steps of the argument which I believe lead to a way to resist J. J. C. Smart's objection to

emergentism in the case of the conscious mind, an objection which has been historically influential and remains so for many philosophers.

I began by offering a background of Smart's objection, and attempting to clarify his objection. I then discussed my definition of "emergence", which differs slightly from other definitions that have been offered. I defined emergence *not* as the in-principle impossibility of deducing an emergent phenomenon from its basal conditions, but as the impossibility of deriving the nature of the emergent phenomenon from the nature of its basal conditions by means of analytic transitions alone. I accept that emergent properties supervene in some sense upon their basal conditions: there is no way in which the emergent properties could be different without some difference in their underlying physical conditions.

This definition allowed me to offer an ontological sense in which a phenomenon could be "partly" emergent, or emergent to a degree. A phenomenon is partly emergent insofar as some part of its nature can't be derived from the nature of its basal conditions by analytic transitions alone. I then followed a strategy of arguing that many of the phenomena which philosophers have accepted as "weakly emergent", or emergent in a *merely epistemic* sense, ought also to be accepted as *ontologically* emergent to a weak degree. In these cases, it is possible to deduce the nature of the emergent phenomenon from the nature of its basal conditions, but only by means of ontologically synthetic bridge laws.

For specific examples, I discuss in Chapter 6 teleological explanations in biology, specifically those which are partly grounded in a history of natural selection. Teleological explanations have a bad reputation, but I'll offer some reasons why philosophers should fault the past misuse of teleological explanations for this and

not the teleological category itself. I'll argue that biological functions are weakly ontologically emergent, insofar as the *normative* part of a biological function is something whose nature is not contained in the evolutionary history and environmental circumstances on which it supervenes. I'll then turn in Chapter 7 to a number of other possible cases of emergence in our world: emergent patterns, other cases of emergence in biology, and emergence in the social sciences.

If we view the emergence of consciousness against this backdrop – widespread emergence up and down the board – then emergence gains in explanatory power (instead of being a mere "unexplained explainer") and the concerns about emergent laws of consciousness being nomological "danglers" seems much less compelling, since lots of emergent laws dangle.

This background of widespread emergence also blocks Smart's objection by Meta-Induction (MI). Smart saw the sciences as continually discovering new ways to reduce complex phenomena to physics, which then offered justification for the conclusion that a similar reduction would someday be found in the case of phenomenal consciousness. In contrast, I have offered a picture on which the sciences are continually finding the limits of ontological reduction, discovering emergent phenomena or patterns which arise unpredictably out of complex systems. Consider, then, principle (MI):

> (MI) If there is some property that many of our established theories have, then when considering non-established theories, we have reason to prefer a theory which has that property over a competing theory which lacks it.

My account of emergence is in line with (MI), since the property of utilizing explanations in terms of emergence is one which many of our established theories already have. Emergence in the philosophy of mind – even if it is to a much more extreme *degree* than lower-level emergence – is nonetheless the same kind of explanation which many lower-level phenomena in fields from sociology to biology take.

## 5.4 The Familiar Smell of Emergence

I also suspect that this offers a response to Smart's objection that emergent properties simply have a suspicious smell or "spooky" feel to them, or (QS):

> (QS) Given any argument for a theory which requires the existence of non-physical properties, it will always be more likely that there is an error in the argument (a false premise, or an invalid move) than it will be that the conclusion is true.

This objection to non-physical properties is intuitive for Smart because it comes from a set of background assumptions – assumptions on which all of the phenomena in the sciences are regularly reducible to physics and talk of "non-physical" natures of things is limited to people telling ghost stories. However, if emergence is a widespread category of explanation, and if I am correct that it should be regarded even at lower levels as a kind of ontological emergence, then there are frequent cases in which complex phenomena have non-physical natures or bear properties which are something over and above the properties of physics. It is nothing suspicious or spooky at all. There is no need to appeal to the weirdness of

consciousness to justify the weirdness of emergence, because emergence is too common to be "weird".

Emergence should not smell so strange to the philosopher. By the time we get to the emergence of consciousness from its basal conditions, we should already have encountered a variety of types of emergence in the natural world. When we face a situation in the case of consciousness where the nature of a phenomenon is clearly something over and above the nature of its basal conditions, then we should recognize the familiar smell of emergence.

## §6 Preview

In Chapter 6, I will discuss in more depth how biological functions, while partly grounded in evolutionary histories, qualify as emergent properties. Bio-functional properties supervene upon the set of properties involving an organism's composition and natural history, and they are likely deducible from them *a priori*. However (pan-functionalism[347] aside) they are only deducible through *a priori* synthetic premises – premises which go above and beyond the elements in the essential definitions of their physical composition and natural history. The property of being a "goal" or an "end" is essential to a biological function. It is not essential to evolutionary history or fundamental physics. This gives us a (defeasible[348]) reason to

[347] That is, someone who holds that functional or teleological properties are part of fundamental physics, natural history, and everything else.

[348] If we gain new evidence that we do not understand the real natures of "goals" or "ends", and that they are in fact ontologically analytic from some conglomeration

conclude that the goal-directedness of biological functions is emergent from the composition and natural history of the biological function, and so biological functions are partly emergent.

In Chapter 7, I will discuss in greater detail how the weaker kinds of emergence can be found in various computational, natural, and social phenomena – domains characterized by the presence of some *normative* element in their nature. It seems plausible to me that cellular automata, ocean waves, species, ecosystems, social movements, and multi-national corporations are in an ontological sense weakly emergent. Some phenomena which have thus far been accepted as epistemically emergent relative to our human epistemic position make plausible cases for partial ontological emergence as well. I will also make the case that three of the hard problems which contemporary philosophers acknowledge – how abstract properties come to be instantiated in a concrete world, how moral norms can supervene upon the natural world while being distinct from it, and how it is that biological and physical states can give rise to the *about*-ness of meaningful language and thought – ought to be accepted as cases of emergence.

These two chapters propose that the emergence of novelties in nature may not be all that novel. We ought to expect it. Consciousness is just the same old new thing. Obviously, phenomenal consciousness is emergence of a stronger kind than, say, biological functions. Most of the nature of a biological function, for example, *is* part of the nature of the organism's composition and evolutionary history – it's just

---

of physical and historical facts, we would no longer have reason to believe biological functions are emergent.

the “goal” part that is supposed to be emergent. On the other hand, very little if any of the nature of phenomenal consciousness is given in the nature of the brain. Consciousness supervenes upon the brain, but not much of what it is to experience the sweet smell of a rose is contained in what it is to be a mushy pink substance housing a series of electro-chemical reactions.

I hope that in these remaining chapters, a picture emerges on which the emergence of unexpected novelties in nature is neither unexpected nor novel.

Chapter 6

# EMERGENCE AND BIOLOGICAL FUNCTIONS

## §1 Introduction

The heart has the function of pumping blood. The skin of a chameleon has functions like signaling to other chameleons, regulating body temperature, and, as is well known, camouflage.[349] These functions, in turn, serve the larger function of promoting the evolutionary fitness of the chameleon. Biological functions in terms of fitness are widespread, serious, respectable forms of explanation in the life sciences. Biological functions are also *teleological*: they explain a state of affairs in the present in terms of being *for* a possible future state. But teleological explanations are not generally considered scientifically serious or respectable.[350]

Two ontological questions come to the forefront. First, should biological functions be accepted as *real*, or are they merely a kind of psychological projection or convenient fiction? Second, if they are real, should biological functions be regarded as *reducible* to some non-teleological story?

Many philosophers and biologists have made strong cases on both sides of each question over the years, and there is an extensive literature on how best to account for biological functions.[351] I do not expect to add anything new to the debate. Instead, I plan to stake out one plausible, well-established position on these

---

[349] Stuart-Fox, D., and Moussalli, A. (2009) offer camouflage, communication, and thermoregulation as "the three primary functions of animal colour patterns."

[350] Neander, K. (1991)

[351] See the bibliography given in Allen, C., (2009)

questions – realism and anti-reductivism about functions – and to show how this position could be regarded as a kind of emergentism about biological functions. If a phenomenon at one level is inexplicable in the terms of a phenomenon at a lower level without loss of content, then we have a reason to accept that the nature of the higher phenomenon is not wholly contained in the nature of the lower phenomenon, but is in fact a kind of novelty.

I will begin with an overview of the non-reductive realist position on biological functions that I am adopting here. I will consider in what sense functions can be considered normative, what might ground the truth of claims about them, what empirical criteria there might be for adopting them, and what relationship they have to natural selection. I will then reflect on seven common fallacies in teleological reasoning which I intend to dissociate the emergentist position from, and suggest that it is these *abuses* of teleological explanation which led teleology to be scientifically discredited in the past, and not anything about teleological explanations as such. Finally, I will explain why the non-reductive realist position on biological functions is best considered to be an *emergentist* position.

## §2 Realism and Anti-Reductivism

### 2.1 Realism

Up through the 1970's, there was considerable hand-writing among philosophers of biology about the ontological status of teleological explanations. Was the "function" of the heart in pumping blood *really* a property of the heart, or was it a property of the head of the researcher? Successful accounts in micro-biology of the physical composition and causal origins of various organisms and their organs were sufficient to explain what hearts were made of and why they happened to beat – no

teleological claims were needed to answer those questions. Nonetheless, it was extraordinarily inconvenient to try to do serious biology without talking about what hearts and habitats were *for*.[352]

In general, it is far more common now for biologists to be realists about biological functions.[353] Why? First, it was recognized that contemporary bio-functional realists were not asserting that biological functions exercised causal forces or were needed to fill some "gap" in material-causal accounts of biology.[354] Second, it was accepted that philosophical dogmas ought not to interfere with serious scientific inquiry, and discussions of functions were a serious prediction-enabling part of biology.[355] Third, it was recognized that any proposed "analysis" or "translation" of functional language into non-functional language, even if it got the necessary and sufficient conditions right, would inevitably either *leave out* the "functional" part of functions (the part that did the *explaining*) or else would import new functional concepts with it into the non-functional language.

For example, suppose we were to translate "the deer's heart is for pumping blood" into a very complex sentence about the chemical processes in the deer's DNA, the causal account of the origin and survival of hearts due to their pumping blood and the enhanced fitness of organisms with them, an evolutionary history of how the deer's heart has adapted over time to the needs of the deer's environment, and so on.

---

[352] Ayala (1999)

[353] Allen (2009)

[354] Ayala (1999)

[355] Milikian (1989a)

One of two things will happen. On the one hand, we may begin to intuitively *read into* this account a teleological *sense* of "adapted" and "survived" and "fitness" and "selection" – a sense which carries with it an implication of being goal-directed – and in so doing, by the back door, acknowledge realism. On the other hand, we may remain anti-teleological purists, and compile a complete causal and material account of the heart without any implication of goal-directedness – in which case, we've failed to give a translation of anything that answers the question: "what is the heart for?" If teleological talk ("x is for y") is merely shorthand, then either it's shorthand for something itself *teleological*, or else it's shorthand for a massive *non-sequitur*, irrelevant to the question the biologist is asking.

If there are no are teleological facts to make biology true, then we seem almost forced to adopt an *error theory* of biology[356]: scientists who talk about the functions of hearts and chameleon skins have invented a kind of collective illusion, a convenient fiction. Physics is the one true science, and biology is a means of fitting physics to questions that only arise from a human[357] mode of thought or interpretive stance. This view has the advantage of ontological parsimony, but is paid for by a

---

[356] I omit as an alternative *non-cognitivism* about biological functions; i.e., that "the chameleon's skin changes in order to conceal itself from predators" is an expression of the researcher's feelings of pity towards the chameleon – because I am not aware of anyone who holds this view.

[357] Talk of this sort often seem to me to carry with it a very *dualistic* connotation that humans are somehow alien or unnatural, prone to asking unnatural questions or to leaving the engines of language idling.

loss in explanatory range and power (by eliminating truths about biological functions[358]) and a lack of coherence with current scientific practice.

These are the considerations which lead me to believe that realism about biological functions is a respectable and plausible view, and to adopt it for my purposes here. But perhaps an anti-realist can, by translating my teleological claims, also successfully translate my "ontological" claims about emergence into some appropriately purified language, and potentially embrace my thesis in this translated way.

I should note that I recognize that there might be a sophisticated form of nominalism about biological functions, on which there are true statements about biological functions, which are true in virtue of the way things really are, but that biological functional properties are not themselves real. Assuming the nominalist is able to translate my other arguments adequately, this sort of nominalism can be regarded as "realism" for my purposes here.

### 2.2 Against Psychologism

I should acknowledge a strong competitor to this sort of realism. It is sometimes claimed that we get our idea of goal-directed explanations in nature by analogy with the goal-directed aspects of own psychology. For example, our idea of the sense in which Smith's eyes *aim* at seeing seems analogous to the sense in which Smith himself *aims* at getting a promotion, or the sense in which Smith *aims* his

---

[358] Except as a kind of truth about the fiction: "the heart is for pumping blood" is true of our fiction in whatever sense "Dr. Watson is the friend of Sherlock Holmes" is true.

own eyes to see an attractive potential mate. It doesn't seem obvious to me that the order of explanation has to go from the psychological sense of "aim" to the biological sense of "aim", and perhaps one could make a case that Smith's representing an aim in his own psychology presupposes that the possibility of correspondence to real aims out in the world – including biological aims, like potential mates. Nonetheless, suppose the account is right and that Smith's private *ends* are conceptually prior for him to the notion of a biological *end*. What should we make of this?

I think we should resist any conclusion that psychological aims being conceptually prior to biological aims entails that biological "aims" must be merely psychological projections onto nature and not part of nature. Consider that a similar claim could be made about causal explanations. An agent's notion of agent-causation (the sense in which she causes her hand to move) may be conceptually prior to any notion she has of causation in the world. Her idea of how a hurricane causes damage is something formed by analogy to the idea of what happens when she causes damage to something. (She might even resent the hurricane). But I do not think we should conclude that causation is merely a psychological projection upon nature. How would agents cause things if we lived in a world without real causes? Similarly, I think we can resist the conclusion that biological functions are the products of our own psychological projections without having to engage in debates about conceptual priority.

## 2.3 Anti-Reductivism

Anti-reductivism is a far more controversial thesis than realism. Biologists generally accept that the facts about biological functions *are* reducible to some set of non-functional facts.[359] But in what *sense* are they reducible?

A biologist likely has in mind a kind of *epistemic* reduction. In this sense, our body of knowledge about biological functional properties can be reduced to our body of knowledge of some lower-level account of evolutionary histories, roles in promoting fitness, selection pressures, underlying physical and chemical processes, etc. The epistemic reduction could result from a *theory reduction* from our theories of biological function to the theories in the corresponding low-level account, on the model given by Ernest Nagel, in which the bio-functional theories can be *deduced* from the underlying theories by means of bridge laws.[360] However, over the last 50 years, Nagel-style theory reductions have proven difficult or impossible in practice.[361] More likely, the epistemic reduction results from a kind of *explanatory*

---

[359] Allen (2009)

[360] Nagel, E., (1961)

[361] For example, Kitcher (1984) argues that there are no bridge laws allowing the derivation of classical genetics from molecular biology in the manner of a theory reduction, and hence there can be no reduction of one theory to the other. Waters (1990) argues in opposition to Kitcher, holding that the theory of classical genetics is indeed reducible to the theory of molecular biology, but concedes that this requires "reformulating the postpositivist conception of theoretical reduction" away from Nagel-style theory unification.

*reduction*, by which every causal process that is explained by biological functions is equally explained by the corresponding low-level account. Each bio-functional property might be "translated" into a sentence about its causal role[362] relative to other properties (i.e., by a kind of Ramsification), and then it might be shown that the lower-level account fills each of the bio-functional property's causal roles.[363] For example, insofar as the heart's function is *being that which makes blood circulation happen*, the underlying account of the heart (including its composition, structure, and entire natural history) fills the role of *that which makes blood circulation happen*. I see no immediate reason to be an anti-reductivist about biological functions in this epistemic sense: it seems plausible that what biological functions explain causally could in principle be equally explained by an exhaustive low-level account.

Suppose that a biologist has in mind a stronger claim, that there is a *metaphysical* reduction of biological functions to the underlying account. Perhaps the biological functions of the organism are metaphysically necessitated, given the underlying account of an organism's evolutionary history, as well as its current environment and the laws of nature. If one accepts a sense in which there are facts in the *present* about the evolutionary history of an organism through the distant past, then we might say that the biological functions of an organism *supervene upon* the present facts about its evolutionary history: it is impossible that a heart have the

---

[362] Putting aside for the moment that "causal role" is itself a functional, normative category.

[363] Kim (1999), 132-133 gives a good sketch of how this might work.

natural history and composition and structure that it does, and nonetheless fail to be *for* pumping blood, and instead be *for* digesting food.[364] Once again, I see no reason to be an anti-reductivist about biological functions in this sense. The emergentist about biological functions also accepts that, once the underlying conditions were in place, nothing else is needed to add in the biological functions.

The claim that I do not suspect biologists generally mean to make – and the claim which I mean to deny – is that there is an *ontological* reduction of the functional properties in biology to some underlying, non-teleological account. In this sense of reduction, it is ontologically analytic, given the properties of the underlying account, that the teleological properties in biology are what they are. I do suspect that some philosophers of biology hold this view. However, reductivism in this ontological sense leads to an unsavory dilemma.

**2.3.1 Option 1** An ontological reductivist in biology could accept a radical view, much like Nagel's panpsychism in the philosophy of mind[365], that teleological properties *are themselves part of the underlying account*. On this view, the heart's property of being *for* pumping blood is nothing over and above the account of its evolutionary history *because* evolution itself is a teleological process, and this in turn is nothing over and above a microphysical account because microphysical properties can be teleological or goal-directed.[366]

---

[364] Thus, unlike moral norms, biological-functional norms do presuppose a regular pattern of success at following the norm in the actual world.

[365] Nagel, T. (1979).

[366] Hints of this approach are in Nagel, T. (2012)

**2.3.2 Option 2** An ontological reductivist can hold three beliefs in tension: (i) that bio-functional properties are real, neither eliminable from biology nor a convenient fiction, (ii) that their nature is nothing over and above the nature of the properties in the underlying account, which is ultimately a physical one, and (iii) that there are no normative, teleological, or functional properties in the underlying account of microphysics. Natural norms like biological functions are *ontologically analytic* given microphysics. There is a correct "analysis" of relations like *x is for y* and *x occurs in order for y to occur*, at least within the domain of biology, into relations between various microphysical objects properties.

I reject both Option 1 and Option 2. I reject Option 1, because I see no evidence for normative properties in microphysics. I reject Option 2, because it has three unpleasant consequences. *First,* it entails that there exists a proper analysis of an "ought" in terms of an "is".[367] Given the unsuccessful record of the positivist program, and the utter failure over the last three centuries since Hume to derive *ought* statements from *is* statements, for there to be an ontological analysis of functional "oughts" in non-functional "is" terms would require that the true natures of functional properties be drastically unlike the ordinary concepts we've been mistakenly using to represent them.[368] *Second*, it entails that the true natures of

[367] Albeit in the language of ontology rather than in any natural language

[368] The inability of material-causal concepts to answer teleological questions seems as well established as the inability of teleological concepts to answer material or causal questions. For the reductivist to be correct, teleological properties must be real – we successfully refer to them with our teleological concepts – and yet unlike

biological functions are massively disjunctive, given their multiple realizability, considering, for example, the many physical differences in organs for seeing or for hearing.[369] *Third*, it entails that there exists in principle a Nagel-style theory reduction from biological functions to microphysics,[370] even though contemporary reductivists in biology rarely promote this idea. While it is widely accepted that there is an *explanatory reduction* from biological functions to some underlying account[371], Nagel-style *theory reductions* have ceased to be a primary focus for reductivists in biology.[372]

So, it is in this distinctively ontological sense of "non-reductivism" that I am staking out a position as a non-reductivist about biological functions. It should be understood that when I say the emergentist offers a "non-reductive" account of biological functions, I do not mean to deny epistemic or metaphysical reducibility,

---

how we represent them as being, since they are really just material-causal properties.

[369] Fodor, J., (1974)

[370] In other words, if it is ontologically analytic that functional property F = microphysical property P, then an ideal reasoner could deduce F from P for every instance in the theories in which it appears.

[371] That is to say that every event explained by biological functions can be explained by some lower-level causal story, and every biological function can be explained by some lower-level account in terms of composition, structure, and natural history.

[372] Brigandt, I. and Love, A. (2012).

but only that the natures of biological functions are nothing over and above the natures of their underlying account.

## §3 Normativity and Grounding

### 3.1 Functions as Natural Norms

I understand biological functions to be one category of "natural norm", a normative property instantiated concretely in the natural world. Natural norms are not prescriptive *imperatives* that regulate human action in the sense that norms of moral behavior do. They are normative only in the broad sense, that they specify one particular possibility out of many possibilities in a non-arbitrary way, to be compared with actual states of affairs.[373]

Functional explanations in biology sometimes use the same sort of normative language that norms of human behavior do. People say that the heart *ought to* pump blood, and that the temperature of the human body *should be* around 98.7 degrees. Sometimes they use normative language that is more specifically teleological. People say that herds migrate *in order to* find water, that flocks travel south *so as to* avoid the winter, that prey *aim to* evade their predators, that the *function* of predators in an ecosystem is to prevent overpopulation of prey, and that the *purpose* of a cat's

[373] In other words, a natural norm is a modal property with a modal logic whose syntactic rules are equivalent to system D, the deontic logic of moral norms; but whereas the semantics of moral norms specify ideal worlds on the basis of features like intrinsic goods or fulfillment of obligations, natural norms specify ideal worlds some alternative but non-arbitrary basis – for instance, as the worlds in which a certain pattern continues or a system functions properly.

grooming behavior is to promote hygiene. Only slightly more disguised, teleological explanations can speak of the *role* of a flower to attract bees to pollinate a plant and the *job* of the pigment arrangers on a chameleon to blend into its background. Biological functions may specify what is *normal* or *healthy* as a contextual backdrop to counterfactual causal explanations: in asphyxiation, the cause of death is the lack of oxygen, not the lack of CO2 → O2 conversion capabilities in the lungs, even though it is true that, had either been present, death by asphyxiation would not have occurred.

What we should not infer from this language is that biological functions or other natural norms carry weight in our moral deliberations. What we should infer is that the weakly normative "ought" quality of biological functions is irreducible to descriptive facts in the same way that the stronger moral "ought" is often recognized as irreducible to descriptive facts.

### 3.2 Grounding

I have accepted that biological functions are real, ontologically irreducible, and essentially normative. One might ask: what grounds the truth of claims about them? That is, in virtue of what features of the world are some claims about biological functions true, and others false? One position is that all true claims about biological functions are partly grounded in the *fitness* of an organism given as a background its evolutionary history and the process of natural selection. Another positions acknowledges that, while some true claims about biological functions are partly grounded in *fitness*, there are other true claims about functions which are instead grounded in an organism's *success* in some non-historical sense.

Consider a colony of birds which use the roof of an old mission for their nests. Is the *function* of the roof to serve as a nesting area for the birds? In one sense, yes – the birds are successfully using the human artifact for nesting. In another sense, no – there is no evolutionary history linking the use of the roof with the well-adaptedness of the birds.[374] The name *spandrel* –after the architectural features in San Marco where birds have taken up just this sort of nesting behavior – has been given to any functional biological structure which did not originally serve the function of enhancing fitness, but which nonetheless resulted from constraints imposed by other biological structures which did serve the function of enhancing fitness.[375] *Exaptation* is the more general case where a structure which evolved to serve one function is co-opted to serve some other function. For example, it may be that feathers evolved at first as a means of regulating body temperature, and only later were exapted for the function of flying.[376] For example, a present function of the human tongue and vocal chords may be to support language, but language may not have been the driving force which lead to the evolution of the tongue and vocal chords just as they are – they may have originally evolved to serve other fitness-

[374] Gould, S.J. and Vrba, E.S. (1982).

[375] ibid.

[376] Gould, S.J. and Lewontin, R.C. (1979). Note that Gould sees exaptations as *non*-functional. But I do not see why one can't simply say that exaptations are functional in virtue of their present usefulness, just not in a way which is grounded in evolutionary history.

enhancing functions, some of which may for modern humans be less fitness-enhancing than linguistic abilities.

I accept that there are many different senses of "function" which are useful for biological explanation. Some of these do refer to an evolutionary history and make reference to natural selection. Some of these do not, and only refer to a role that something plays in the present success of an organism. Consider the flippers of a turtle, which it uses to bury an egg on a sandy beach.[377] In the narrow evolutionary sense, the flippers are for swimming, but not for burying the egg in the sand. In the broader sense of biological function, the flippers are for both. If an animal adopts a non-innate behavior in order to survive, the behavior is functional in the broad sense but not in the narrow sense. An artificial heart, made of plastic but functioning within a human body, functions to pump blood in the broad sense but not in the narrow evolutionary sense. There may be borderline cases between the two.

Because I don't want to take further sides in this debate, I will focus in the remainder of this Chapter exclusively on biological functions whose explanations appeal to the evolutionary history of an organism, since these are the sorts of functions most generally accepted. But this doesn't mean I presume that all functional explanations must work in this way. I will address the possibility of biological functions which aren't grounded in evolutionary histories in Chapter 7.

[377] ibid.

## §4 Empirical Criteria for Teleological Explanations

There must be empirical criteria for assigning teleological explanations to phenomena. Not all phenomena have teleological explanations – unlike Aristotle, I do not want to say that a rock falls *in order to* hit the ground. At the same time, I do want to say that a bird has wings *in order to* fly. This means I need some means of distinguishing the rock's falling from the bird's wings. It may be intuitive to me that the bird's wings are for flying but the rock's downward motion is not for the purpose of hitting the ground. But it is more likely that I have these intuitions because I have internalized some set of empirical criteria than it is that I have *a priori* knowledge of the natures of rock-fallings and bird's wings.

Under what conditions can we say that one event occurs *for* another event to occur, at least in the sense used in biology? I will abandon all hope up front of finding a counterexample-free set of necessary and sufficient conditions for *A is for B*. I am working off the assumption, already discussed, that the "for" relation is irreducible to more basic explanatory relations. Nonetheless, some discussion of the empirical criteria we use for assigning teleological judgments may provide an illuminating non-reductive explanation of what biological functions are. First, I will consider two proposals for empirical criteria, and note their deficiencies. Then, I offer a third proposal for what criteria to use when judging whether something is or is not a case of a biological function.

### 4.1 Rejected Proposals

#### 4.1.1 Proposal 1: *A is for B means if B, then A.*

For example, the heart is for circulating blood. Every vertebrate circulates blood. A vertebrate circulates blood only if it has a heart. So, if a vertebrate

circulates blood, then it has a heart. However, as mentioned earlier, every vertebrate makes a regular beating noise in its chest only if it has a heart. Nonetheless, the heart is not *for* making a beating noise.[378]

#### 4.1.2 Proposal 2: *A is for B* means that *A plays a causal role as part of a system of which B is a regular outcome*.

For example, the gills of a fish play a causal role in a system of which respiration is a function. So, in fish, gills are for respiration. However, the notion of a causal *role*, as opposed to a mere causal influence, *depends upon* the notion of a teleological explanation. The ability of a fish to respirate also causally depends upon the fish remaining in motion, which depends upon its having the flexible skeletal structure it has. But the skeletal structure of the fish does not play a causal role in respiration precisely because the skeletal structure is not *for* respiration. The gills of the fish are also partly responsible for the mass of the fish being what it is. But the gills do not play a causal role in the mass of the fish, *because* they do not function to promote the mass of the fish. Alternatively, consider the gills of a dead fish. The gills do not presently cause the fish to respirate – the fish is dead. Do the gills play a causal role in a system of which respiration is the outcome? If one says yes, then one can only say so in light of what the gills are *supposed to do* – that is, what their function is.

### 4.2 Another Proposal

What criteria do we use to judge that something is a case of a biological function? I understand our judgment that *x* is the function of *y* to work as a kind of

[378] This example is from Milikian (1989b)

inference to the best explanation: it is not that we observe functions themselves, or that we derive them from our observations, but rather that we infer them in situations where they help explain the regular patterns we observe. I'd propose the following:

**4.2.1 Proposal 3***: if event A occurs in order for event B to occur, then generally the following conditions hold:*

*(i) Possibility*. Given event A, B is a distinct non-arbitrary future possibility, but not the only possibility.

This condition reflects the normative character of a biological function. The heart is for pumping blood. Given the event of the heart's existing, one particular future possibility is that the heart pumps blood. It need not be actual – the hearts of dead animals do not pump blood. It is not the only possibility – the heart might pump kool aid, or it might explode. A norm selects non-arbitrarily one future possibility, that of pumping blood.

*(ii) Manipulability.* There is a possible intervention upon A which would lead to a manipulation of the state of B.

This condition reflects that there was some truth in the "causal role" account rejected earlier. That the function of the gills is respiration tells us that, while the gills may not actually *cause* respiration (for example, in a dead fish), there is a possible intervention upon the gills which would lead to a change in the state of respiration – for example, in the possible world where the fish is alive, shutting down the gills would stop respiration.

*(iii) Stability*. For some range of possible interventions upon A, it is the case that some new event C occurs in order for B to occur with its prior value.

The causal patterns which support biological functions must be *stable* in the sense that in many cases the final cause (B) would have been brought about by some other mechanism (C) even if its original cause (A) had not occurred. For example, the nose plays a causal role in breathing. The nose also plays a causal role in the growth of nose hairs. But the nose is for breathing, not for growing nose hairs. Why? At least part of the explanation can appeal to the fact that, in certain situations where the nose is blocked, an organism may continue breathing normally through some other mechanism – the mouth. However, there are no situations in which a nose cannot grow nose hair, but the nose hairs still grow by an alternative means. There are not many cases where the heart of an organism breaks down and, nonetheless, the blood is still pumped. But there are some cases – like humans who perform CPR – where an event may happen to pump the blood in place of the heart. Should the beating noise of the heart stop, nothing will "step in" take its place.

Stability is most significant when considering evolutionary explanations over time. Suppose an organism has some adaptation which serves one function, but changes in the organism's environment prevent it from serving that function. We could predict some new adaption over time (if the organism is to survive in the new environment) would arise to take the place of the prior adaptation. We would not predict the same response for some non-functional property of an organism.

*(iv) Failure-Explicability*. If A occurs and B does not occur, then necessarily some C has occurred and has prevented B from occurring.

As mentioned, that B is the function of A does not guarantee that A actually causes B. However, if B is the function of A, then there must be a causal explanation of what has *interfered*, or gotten in the way of the process of A causing B. One

purpose of the gall bladder is to store bile and secrete it into the small intestine in order to promote digestion, to promote the overall fitness of the organism. The gall bladder *just happens* to produce gall stones. If a gall bladder fails to produce gall stones, we have no right to expect an answer to the question "why *didn't* it produce gall stones?" beyond, perhaps, a purely statistical or probabilistic explanation. If a gall bladder fails to secrete bile into the small intestine, we do have a right to expect an explanation of what happened.

I am certain that there are other generally true qualities of functional explanations. As I mentioned, my purpose has not been to undertake a reductive analysis of what a functional explanation is. I've only wanted to provide a few examples of empirically investigable criteria for functional explanations, to establish that the criteria for functions are not purely a matter of "intuition" or the special quirks of human psychology, but the same sorts of interests in prediction and manipulation of the world which lead us to seek causal explanations of the phenomena we encounter.

## §5 Natural Selection

I have said that I am discussing cases of biological functions partly grounded by evolutionary histories: cases where *x* is the function of *y in part* because *y* promotes the *fitness* of the organism in which *y* occurs by means of producing *y*. (By "in part", I indicate that functions are not *wholly* grounded in fitness; they are emergent.) For example, birds' wings promote their fitness by enabling flight. The wings' design has been *selected* for fitness. But "selected" is teleological language. In what sense is fitness a *telos*?

We don't want to risk anthropomorphizing natural selection. There is a danger in the language of analogy. It is not as though nature "selects" a feature for its fitness in the same sense in which I select the chicken over the beef on the menu for its tastiness. Nature does not desire fitness in the way in which I desire chocolate pie.

Of course, we also shouldn't make too much of the disanalogy. I *do* desire chocolate pie in part because the organism I am desires chocolate pie, which is in part because chocolate is full of the sugars and fats that have tended (until recently) to promote the fitness of organisms like me.[379] So, it is not entirely off to say that there is something in common between the way I desire chocolate pie and the way nature selects fitness as an end.

In what sense, then, is fitness an *end* for organisms, or whole species? Again, clearly it is not an "end" in the sense that it has imperative force or moral weight. Birth control may not necessarily enhance fitness, but that doesn't mean it doesn't achieve a legitimate end. Birth rate is not a measure of *eudemonia*.

If fitness is not analogous to a moral end, and it is not analogous to an end that *somebody* has in mind, in what sense is fitness nevertheless the end or goal of biological functions? A better analogy might be the sense in which a clock's hand at noon points towards the number XXII, or in which a person's outstretched arm might point at a nearby rock even if he does not intend it to. The hand of the clock is

---

[379] I say in part, because if you dare try to reduce my desire to some abstract formulation about fitness, then it's clear you've never had the phenomenal experience of chocolate pie eating like I've had.

directed at XXII, and the action of the person's hand is directed towards the rock. Similarly, a biological function is a kind of *directed action* towards the end of fitness (or, more properly, towards a variety of other ends which have reproductive success as their end), in a sense that does not imply a corresponding mental act. [380]

## §6 The Seven Deadly Sins of Teleological Explanation

### 6.1 Teleology and its Abuses

So far, I have staked out a position as a realist and non-reductivist about biological functions, and have chosen to focus on biological functions which are partly grounded in and supervene upon evolutionary histories.

Nonetheless, at this point, I suspect many readers will still harbor skepticism about biological functions, period. Teleological explanations and "natural norms" have a bad reputation among many philosophers. I think this bad reputation is partly justified – not because there is anything wrong with teleological explanations as such, but because teleological explanations have been so badly misused in the past. Before I proceed with a realist account of teleology and natural norms generally, it will be necessary to distance my own understanding of natural norms from these past abuses.

By a "biological function" or "teleological explanation in biology", I mean a partially normative explanation of some phenomenon in terms of its function, or

---

[380] One difference is that the rock a demonstrative act refers to is in the *present* whereas the biological function's end of enhanced fitness lies at least partly in the future. Another difference is that pointings and clocks include a history of consciously intentions by agents.

what it is *for*. Because I will be arguing that natural norms are emergent, I will hold that they emerge at different levels – the individual organism level as opposed to the organ, or the individual as opposed to the social group. The natural norm applies only at its respective level: it is a norm which is internal to its level and does not indicate anything about norms at a higher or lower level. So, what something is for at one level may not be the same as what it is for at another level, and the teleological explanations at each level may be incongruent. On one level, an individual organism may have the function of surviving and reproducing. On another level, the same organism may have a social function within its pack which requires it *not* to reproduce.

I accept *pluralism* about natural norms – there may be multiple, true, conflicting norms that apply to the same entity or event, whether at different levels or at the same level. When the immune system is hijacked by a virus, the immune system may both serve the function of protecting the organism from viruses and the function of spreading the virus in the organism. This pluralism need not devolve into anti-realism. There are other things true of the immune system that are clearly *not* its function: leukocytes are white-colored, but it is not their function to be white-colored; the heart causes me to hear a sound, but it is not the function of the heart to cause me to hear a sound.

I regard teleological explanations to be a category of explanation on a par with causal explanations, material explanations, and formal explanations. *I recognize sharp boundaries between each category of explanation*, and do not admit that one sort of explanation can substitute for another: no matter how much you tell me about what the heart is for, you will not have told me how it came to be there,

and vice-versa. We are thus justified under the principle of parsimony in being realists about normative or teleological properties and relations so long as they are the minimum necessary for teleological explanations of real phenomena, regardless of how exhaustive our causal and material explanations of the phenomena are. At the same time, no teleological explanation could even in principle suffice to explain a "gap" in our existing causal and material explanations of the universe.

To distinguish this account of teleological explanation from its abuses, I wish to attempt to identify the nature of the most common abuses which have tarnished its reputation or the "Seven Deadly Sins" of teleological explanation.

### 6.2 Investigative Sloth

An especially satisfying teleological explanation can lead researchers to prematurely abandon causal and material explanations. Consider how investigations of the brain have historically tended to focus disproportionately on individuals with psychological disorders: we feel no compelling reason to investigate the causes and composition of a healthy psychology, since we can understand it in terms of reasons. A similar bias lead to the stagnation of scientific inquiry generally in the West up until the last five centuries. We understood that the function of an acorn was to grow into an oak tree – to move from potentiality to actuality – and it seemed that was all there was to say about why the acorn became an oak tree. This hardly scratched the surface of explaining why acorns grew into oak trees, of course.

So, abandoning confidence in teleological explanations at the dawn of the scientific revolution was the only way to shake the West out of its intellectual slumber. Of course, the error was not in affirming that the function of an acorn is in

fact to grow into an oak tree, but in the sleepy satisfaction that this was all there was to say.

## 6.3 Explanatory Adultery

In the humanities, a history is an artful weaving of efficient-causal explanations with teleological ones – the purposes of Napoleon are thwarted by the Russian winter, and weather patterns save Europe. Natural history is tempted to follow suit. A bare, purposeless, efficient-causal history of the natural world lusts after a meaningful teleology, and teleology longs to make itself impure by adulterating itself with efficient-causal claims. This unrighteous union of teleological and causal explanations is what I call "explanatory adultery."

Consider the first case: struck with wonder that the complexity of life has arisen from purposeless matter, one infers that teleological principles of order must have intervened and placed a causal role in producing over time the complexity of life. Into the causal story of the events leading up to life, it is inserted that these events were caused to happen because they were for the purpose of bringing about future life. Notice that the error is not in the sense of wonder, not in holding biological life has *teloi* that matter does not, and not in holding that part of what makes life what it is are these *teloi* and hence what life is cannot be reduced to purposeless matter. Rather, the error is in substituting a teleological explanation in the middle of a series of causal explanations to fill an efficient-causation-shaped gap. We force into the picture a piece from a different puzzle: *d* happened because *c* happened, *c* happened because *b* happened, and *b* happened because *d ought to* happen. This implies a kind of backwards-causation, with future ends producing

past events. The resulting explanation never lasts long, but it often gives us temporary satisfaction.

Consider the opposite case: in the middle of an explanation of what war is for, one inserts an explanation from natural history – the survival of phenotypes with greater fitness – and so, war to purge the weak from a strong society is justified by inserting an efficient-causal claim into a teleological account. Similar accounts might be offered to justify eliminating charitable social programs or international humanitarian aid. One might invoke the evolutionary account of human origins to justify a controversial medical experiment as the next step in human evolution. Again, one has placed a piece from the wrong puzzle: how something came to be does not in itself tell us all there is to what it is for, any more than what something is for tells us all there is about how it came to be.

### 6.4 Hasty Judgments

When we are considering human practices, there is a case to be made for giving serious regard to human intuitions – regarding the appropriate use of language, the principles of ethics, or so on. But when we are not considering human moral norms, or semantic norms, but natural norms in the world, an intuitive judgment is a sign of impatience. It may seem intuitive that the brain is for heating the blood, not for directing action. It may seem intuitive that, because most organs in the body have a function that serves the overall function of the body, every organ in the body must have a function that serves the function of the body. It might even seem intuitive to some that rabbits have ears in order to make them easier for hunters to carry, or that the beavers built the dam in the creek in order to give us a swimming hole (or – the villains! – to flood us out when it rains). Intuitions are not

trustworthy here. We must not be so quick to judge. There must be empirical criteria to tell us when we have a case of a function or other natural norm and when we don't.

## 6.5 Elitist Greed

I have suggested that we should be pluralists about natural norms, and functions, even for functions of the same object, at the same level, at the same time. For example, the intestines have many functions: digesting and breaking down food, reabsorbing liquids and absorbing nutrients, removal of wastes, and so on. This is parallel to the way in which an event may have a plurality of efficient causes. Opposed to this pluralist view are *exclusivism* about functions, and *relativism* about functions. For the exclusivist, it is not enough to say that something is a function or one function of a thing. It is necessary to say that it is the function of the thing, the real function (or at least the only important one) – that the other functions we might have thought it had were more or less illusory. The exclusivist wants to encompass all of the teleological explanatory power: a kind of greedy elitism.

So, one exclusivist says *the* function of religion in society is to reconcile the proletariat to their oppressed state, another that it is to promote social cohesion and to solve prisoners dilemma scenarios, another that it is a form of resistance to the state; someone says that the function of the universities is to pursue knowledge for knowledge's sake, another that it is to prepare a skilled workforce, and another that it is to legitimate the rule of the technocratic elite; someone says that the function of the civil war was to eliminate slavery, and another that it was really about industrialization and the agrarian economy, and another that it was a tragic accident, and so on. The same exclusivistic approach applies when someone says

that organisms are just mechanisms for the transfer and perpetuation of genes, or that social behavior in chimps is just to promote their own long-term welfare, or that altruistic behavior in humans is merely hidden selfishness.

The relativist, upset by this greed, responds by turning profligate. The Every teleological claim obtains its truth value relative to some perspective or other. It is true from one perspective that altruism is only selfishness, and false from another, on which altruism is only altruistic. Nonetheless, the relativist maintains the exclusivist's assumption: teleological claims remain exclusivistic, "just" or "only" claims, albeit within a perspective. The pluralist resists this move. Teleological claims are rarely "just" or "only" claims. One can maintain that an entity or event serves one function without ruling out that it serves other, possibly incongruent, functions.

### 6.6 All-Consuming Gluttony

Sometimes a teleological explanation at one level is not satisfied to remain 'internal' to that level from which it emerged, and endeavors to consume the role of teleological explanation from other levels. One level of explanation attempts to be all-consuming, a kind of gluttony. On the organism level, the function of grooming behavior in a cat is the cat's own hygiene. But on the level of an individual cat's psychology, the function of grooming behavior may be a reaction to ease anxiety in stressful situations. It would be reaching too far to assert that cleanliness must also be the function of grooming in the cats psychology.

We are most apt to confound explanatory levels in a top-down, anthropocentric way – like the hunter who assumes that rabbits have ears so that he can hold them better. As an antidote to this tendency, someone may overreact in

the opposite direction: saying that an animal exists for the purpose of carrying its genetic code around, or that the function of eating is to funnel nutrients to one's individual cells. This is a failure to explain the phenomenon *on the level at which it occurs* – on which eating is to nourish the organism as a whole.

## 6.7 Morality Envy

As I've stated earlier, the "natural norms" and teleological properties of natural objects I have been discussing should not be confused with moral norms – those norms which apply to conscious reasoners. Unfortunately, natural norms begin to envy for moral norms. They want to have the force of imperatives too.

Many people are familiar with arguments from the premise that one thing is the natural purpose of another to the conclusion that the one thing is the moral purpose of the other. Unless it is the case that there is an independent moral principle of the form "follow the natural purposes of things in situations of this sort," there is no reason to accept these sorts of arguments. I accept that one function of an organism is to reproduce. I also had my dog spayed. One may be a realist about teleologies and coherently support monasteries, birth control, equal rights for same-sex couples, the exaltation of perpetual virginity, and so on. Moral norms do not eliminate natural norms – the function of the heart is still to pump blood, even if the pro-death penalty retributivist is right that the serial killer's heart ought to be stopped. Moral norms supersede natural norms. Even though there may be a sense in which a highly infectious disease is very "successful", the disease is not morally good.

## 6.8 Anthropomorphic Pride

The final and greatest temptation when considering teleological explanations in nature comes from the fact that that we ourselves have various purposes and ends that we pursue, and so we are apt to reason by analogy that something must be true of the ends in the natural world because it is true of our ends. But human purposes are very different from natural purposes. We err in thinking so highly of ourselves that the purposes in nature must be like our own. Three examples:

First, human technological artifacts are distinguished by having one all-encompassing telos. The primary function of the hammer is for hitting things, and a light bulb is for giving light. However, as I have argued, natural objects often don't have one all-encompassing telos: is the function of a pigeon to survive and reproduce, to recycle the trash of the city, to play its role in pigeon society or the urban ecosystem, or to incubate worms?

Second, conscious human intentions are typically unified. I cannot at the same moment knowingly intend two contrary things. If I have two contrary intentions, it is because I am not consciously aware of it; if I become aware of it, I must dismiss the one (or, at least, repress or sublimate it) in order to consciously pursue the other. If I have one end, and a choice between two means to that end, then in a state of complete information I will likely choose the more efficient of the two means. However, natural functions need not display this sort of obvious unity. The function of a mosquito that bites a leg is at odds with the function of the leg. The function of a low-heat forest fire in promoting the health of an ecosystem is at odds with the function of the smaller brush killed in it. Further, the various natural functions of organisms may be fulfilled inefficiently with respect to their other

functions – consider the high rate of women who die in pregnancy, as a result of the failure of female anatomy to adapt in sync with increased infant brain capacity. Even if there turns out to be a deep harmony to the cosmos (a final end of all ends), it is presumptuous to think it will involve unifying all of the pieces together efficiently in a manner similar to a conscious human intention.

Finally, teleological explanations are always mixed with efficient-causal explanations in accounts of human action. Suppose that Jenny kisses Jimmy in order to cheer him up. If the observer does not understand that cheering Jimmy up is one of the causes of the kiss, the observer has not really understood what happened. Similarly, histories in the humanities must be of this form: they must cite desires for some end as efficient causes. It is a serious debate as to whether the *cause* of bombing of Hiroshima was in order to prompt Japan to surrender, or in order to impress the USSR. But, while causal and teleological explanations are married in human action, uniting them together outside of human action is, again, an act of explanatory adultery: we cannot cite the desire of the acorn to become a tree as the efficient cause of its becoming a tree. This is anthropomorphizing the acorn. Likewise, we cannot cite a purpose for the cosmos to become a place such as it is, with things like us in it, as the efficient cause of the structure of the cosmos. This is anthropomorphizing.

## 6.9 A Final Diagnosis

In considering these several deadly sins that might be committed with teleological explanations, it does not seem to me that any of them reveal essential flaws in teleological explanations as such. There may be independent reasons for skepticism about teleological explanations in nature, of course (I have asserted that

they must have empirical criteria, but one might wonder what these criteria are). However, skepticism or suspicion of teleological explanations that is motivated by observing the abuses listed above should be reconsidered – the fault is not in the type of explanation, but in the person who does the explaining.

## §7 The Emergence of Biological Functions

### 7.1 Motivations for Bio-Functional Emergentism

By this point, I hope that I have cleared the way for a level-headed discussion of the emergence of biological functions. I have explained how there can be weaker and stronger forms of ontological emergence, and the consistency of both with the possibility of ideal *a priori* deducibility. I have offered reasons to think that realism and anti-reductivism about biological functions are plausible, and not merely cases of observers reading themselves into their observations. I have distinguished the sense of *function* which I am using from many of the past abuses of teleological explanations, and I have described how functions can be grounded in natural selection. It's about time I get on with it – why think biological functions are ontologically emergent?

On the definition of emergence I have been working with, an emergent phenomenon is entirely dependent for its existence upon its subvening physical base (it supervenes upon it), and at the same time its essential nature is something "over and above" the nature of its physical base.

In the case of biological functions, the supervenience claim is not controversial. Both reductivists and non-reductivists accept that any two possible worlds which were exactly alike with respect to their physical environments and past evolutionary histories would be exactly alike with respect to what everything

biological was *for*. There is no possible world exactly like ours in these respects where the heart isn't for pumping blood. Functions depend on their subvening base. Given a radically different environment and history, it might have been that the beating noise of the heart conferred survival advantages leading to the development of the heart and the pumping blood did not. We would then say that the heart was for making a beating noise, not for blood-pumping.

It is more difficult to establish the something "over and above" claim. I have defined this claim in terms of being *ontologically synthetic*: it amounts to the claim that the nature of the heart's function in pumping blood contains some element which could not be derived by purely analytic moves from the nature of its subvening base in physical environment, material composition, evolutionary history, and so on.

Consider that the mammary glands have the function of producing milk. Imagine writing out a list of all of the events in the natural history of mammals which lead to the evolution of mammary glands, including various feedback loops involving the successful production of milk and the increased survival chances of offspring. Then, imagine writing out a list of all of the chemical and biological structures which compose the mammary glands. You might add to this whatever other facts about present and past events involving mammary glands that you prefer, if you believe they belong in its subvening base. Now, go back, and *purge* each of your lists of any unnecessary *teleological* language: talk of *roles* or *jobs* or *purposes* or *means* and *ends* which are not essential to physical and causal descriptions. What remains of your lists will *necessitate* that the function of the mammary glands is to produce milk. Because you are smart, you can also likely

*intuit* that the function of the mammary glands is to produce milk. But could you *derive*, using only the rules of logic and various translations of the terms you've used in your descriptions, that the mammary glands are *for* producing milk? No – you've gotten rid of everything teleological that was in the base. Perhaps you could derive something like: "mammary glands regularly produce milk and would not exist if they didn't", or any number of other sentences involving "mammary glands" and "milk" that suggest, intuitively, that they are *for* milk. However, the derivation could not go forward without some additional *synthetic* premise that included the operator *for* in it. (If you could derive it, then either you didn't completely purge the unnecessary teleological language from your physical descriptions, or, much to everyone's surprise, teleological explanations are essential to physics and causation.)

A similar consideration applies to all biological functions: a derivation of the function from its basal conditions will require some bridge statement relating the non-teleological properties in the base to the teleological properties in the function. This bridge statement may be necessarily true and perhaps even highly intuitive. However, assuming anti-reductivism is true in the sense given in section 3.1.3, this will be a synthetic statement, not an analytic one.

## 7.2 Objections to Bio-Functional Emergentism

There are a number of objections which are likely to be leveled against emergentism about biological functions.

First, an anti-reductivists may be willing to accept that biological functions are weakly emergent (which they understand as merely epistemic), but they reject

strong emergence (which they understand as ontological).[381] They understand "weak emergence" as a kind of inability to derive or predict the emergent phenomenon from the underlying phenomenon except by means of non-analytic bridge principles, like those learned through simulation.[382] Nonetheless, they deny strong emergence, insofar as the derivation is possible given adequate computational and conceptual resources. This is sometimes expressed as the view that emergence is a function of present ignorance – a phenomenon is emergent only relative to the scientist's inability to predict it with the cognitive resources he has. However, on my account, weak emergence need not be merely epistemic – I have advocated for ontological weak emergence. Ideal deducibility does not rule out ontological emergence, so long as non-analytic bridge principles are needed for the deduction – something the weak emergentist admits. On my account, "weak emergence" is just a weaker kind of ontological emergence, where a lesser part of the nature of emergent phenomenon is something over and above its base. I would suggest that these authors should accept the full ontological consequences of their view, and accept that "weakly emergent" biological functions are in fact ontologically emergent.

Second, some reductivist-leaning critics are likely to complain that reductive accounts of biological functions are available in terms much like those I've offered as a subvening base – evolutionary history, composition, environment, and so on. However, I've acknowledged that there are senses of reduction which are consistent with ontological emergence, including epistemic reductions on which the causal role

[381] See Rothschild, L. (2006)

[382] Bedau, M. (1997), (2008)

of biological functions is fully explained and filled by that of their subvening base. What the reductivist has yet to show is that the quality of *for*-ness is ontologically reducible to some physical quality.

Third, the reductivist might attempt to provide an analysis of the concept of biological function, or at least hold out hope that such an analysis is forthcoming. The analysis could show that the concept of a function really was merely analytic given the relevant physical background. I would resist this approach on two grounds. First, the task of providing conceptual analyses for biological functions, as opposed to supervenience conditions, has not met with much success.[383] Second, even if the *concept* of a function were found to be latent somewhere in the concepts of its physical background, this would be insufficient to show that the nature of a *function* is latent in the nature of the physical background – conceptual priority does not always match up with ontological priority, nor conceptual analyticity with ontological analyticity. It is conceptually analytic (and *a priori*) that the standard meter bar in Paris is one meter long. It is also *contingent* that the standard meter bar is one meter long, not essential (or ontologically analytic) to it.

Fourth, there is a kind of *Quinean* physicalism which would reject the whole metaphysical discussion altogether, denying that there is such a thing as the "nature" of a function and the "nature" of a microphysical substance. In place of the "ontological analyticity" and "ontological syntheticity" distinction which I have offered, the Quinean will insist upon "ontological relativity" given the internal ontology of a given language. For the Quinean, function-language is just a kind of

[383] See Rothschild, L. (2006)

linguistic short-hand derived from a series of observation statements, the common source of our physical material-causal descriptions of the world. There is a kind of radical translation (however indeterminate) which is possible between function-language and physical material-causal language. This approach is immune to my argument about the non-derivability of functions from physical descriptions. However, those adopting this view should be cautioned that their view is not a *metaphysical* variety of physicalism: it does not hold that all concrete realities have a physical nature, since it rejects the category of a language-independent "nature" outright. So, it is *no more opposed to emergentism than it is to the metaphysical physicalism* which the emergentist is debating against.

Fifth, one might complain that the talk of "essential natures" in metaphysics is misguided, and there is nothing more to essence than modality. If emergentism about biological functions makes the same exact epistemic and empirical *and modal* claims as non-reductive-physicalism about biological functions, then there is no serious difference between the views. If the way in which bees are needed *in order to* pollinate the flowers follows logically from non-functional facts about flowers and bees, and makes no empirical claims that could not be expressed without the *in order to* language, and yet can't be reduced to non-functional facts, then what is left to argue about? I would reply that talk of "natures" in fact helps clarify the debate between two otherwise indistinguishable views – a debate which has tended, for the sake of distinction, to force emergentists into weaker modal claims (like denying supervenience) or reliance on an epistemic gap (which not even a mathematical

archangel[384] could bridge), both of which are positions that may be plausible for phenomenal consciousness but are not plausible for biological functions. The emergentist's real interest has always been the "something over and above" claim, and the non-reductive physicalist has always insisted on the "nothing over and above" claim. If these are empty claims, at least we know what they are supposed to be about, and how we would go about answering them.

Sixth, a critic might cite Hume's dictum, that "there are no necessary connections between distinct essences." The emergentist seems to be claiming just that: that emergent phenomena are necessitated by their subvening bases, but nonetheless are distinct in essence from their subvening bases.[385] However, Hume's dictum should be understood as ruling out necessary connections between *wholly* distinct essences, not partially distinct essences. There is a necessary connection between being Socrates and being a person, but Socrates is not identical in essence with personhood, though personhood is part of the essence of Socrates. The emergentist is not claiming that emergent phenomena are *wholly* distinct in nature from their subvening base, only that they are *partially* distinct: that there is some additional *novelty* which is part of their essence. This is consistent with Hume's dictum.

Finally, a critic might revert to anti-realism. One might adopt an account on which biological functions are mere metaphor based on human mental states, or

[384] The reference is to Broad (1925)

[385] Some discussion of this can be found in Stoljar (2010), 151

merely in the mind of the observer.[386] While this move is plausible for many of the emergent phenomena which I will be discussing later, it does not seem as plausible in the case of biological functions. (Consider that if biological functions were in the mind of the observer, liver disease would be in the mind of the observer. Medical science would be mere metaphor.) In any case, the fallacy seems to be in reasoning that because human mental ends are conceptually prior to biological ends, then human ends must be ontologically prior as well. However, biological ends are undoubtedly historically and materially prior to mental ends, which often reflect biological ends, and this makes it odd to say that they depend on analogies with human mental ends.

## §8 Conclusions

In Chapter 5, I argued that the ontological question of whether one thing has a nature which is or isn't wholly contained in another is really the question of ontological analyticity – whether an ideal reasoner, given the nature of the basal conditions, could derive from it the nature of some higher-level phenomenon using only analytic moves, like translation sentences and logical transitions. Neither the failure nor the success of ideal deducibility tests in themselves guarantee us any metaphysical conclusions, but they are ontologically significant only when we consider the deductive resources available to the ideal reasoner. Insofar as we are confident that we understand the nature of a higher-level phenomenon, and we are confident that we are not able to deduce it from the set of lower-level conditions which necessitate it using only analytic transitions, then we have prima facie reason

[386] For a description of this view, see Allen, Colin (2009)

to believe that the higher-level phenomenon is emergent, a genuine ontological novelty.

In this chapter, I have argued that Biological functions supervene on the physical, but they are also by nature normative properties – properties which can only be deduced *a priori* from physical properties by non-analytic methods. Unless we wish to embrace anti-realism about biological functions, which comes at great cost, we ought to accept biological functions as a weak kind of ontological emergence.

Much like the case against emergence in any particular domain is most plausible when viewed against the backdrop of apparent reductions across the sciences, so also the case *for* emergence in a particular domain is most plausible when viewed against a backdrop of other apparent emergent phenomena across the sciences. In Chapter 7, I hope to paint this backdrop. I will to consider cases of weak emergence in a variety of other domains in the special sciences, as well as the emergence of representational accuracy and the emergence of instantiations of abstract properties, and argue that – much like biological functions – they should be viewed as ontologically significant.

Chapter 7

# EMERGENCE EVERYWHERE

## §1. Introduction

A recurring objection to emergentism in the philosophy of mind has been that emergence invokes a novel, *ad hoc* sort of explanation for a single phenomenon. For instance, since J. J. C. Smart (1959) argued that fundamental laws linking low-level states to high-level states would be unlike the other fundamental laws in the sciences – the suggestion has a "fishy smell" to it. Holding that consciousness emerges from the physical – that it supervenes upon it, and yet has some novelty in its nature[387] – seems to offer no gains in explanatory power over a simpler identity theory, violating Occam's razor. Emergence is simply not the sort of explanation which regularly appears outside of the philosophy of mind.

This objection may be becoming outdated. Emergence *is* the sort of explanation which appears outside of the philosophy of mind. It is no longer a one-time *ad hoc* form of explanation applied only to the case of phenomenal consciousness, but an explanation used across a wide spectrum of fields of inquiry, from the life sciences to the social sciences. Elsewhere in philosophy, emergence accurately captures mainstream views in meta-ethics and the study of intentionality. Emergence is everywhere. Relating the case of consciousness to these other cases gives "it emerges" greater explanatory weight, possibly enough to justify the addition to our ontology.

---

[387] See chapter 2 for discussion of the relevant definition of 'supervenience', and chapter 4 for discussion of the relevant definition of 'novelty in its nature'.

In general, philosophers of mind have tended to push back against the suggestion that the "weak" sense of "emergence" in the special sciences is an *ontological* concept, in the same way in which the "strong" sense of emergence which is supposed to hold for consciousness is ontological. In Chapter 5, I laid the groundwork for admitting a kind of *ontological weak emergence*, on which the emergent facts are deducible *a priori* from the underlying facts they supervene upon (hence "weak" as opposed to "strong"[388]), but this deduction is not possible by means of analysis[389] alone (for instance, it requires a form of simulation), indicating that some part of the nature of the emergent phenomenon is something over and above the nature of its basal conditions. In this regard, while the emergence of consciousness is of a stronger kind than the emergence of computational patterns, both can be regarded as ontologically significant.

In Chapter 6, I applied this to a particular case, and argued that biological functions, while grounded in facts about their composition and an evolutionary history of natural selection, should be regarded as a case of ontological weak emergence. Insofar as biological functions are regarded as both real and as essentially normative, they should not be regarded as ontologically reducible to the facts about their composition and history. A major section of my argument was

---

[388] Insofar as "strong" emergence indicates the impossibility of an *a priori* deduction, even in principle.

[389] I use both "synthetic" and "analytic" in a specialized, ontological sense, with respect to the essential definitions of things as opposed to conceptual or linguistic meanings. See Chapter 4.

devoted to dissociating the emergentist view of biological functions from common abuses of teleological explanations. For instance, I rejected the claim that teleological properties in nature could play a *causal* explanatory role over and above the causal explanatory role played by their basal conditions. Nonetheless, I held they should be regarded as having a distinct nature from those basal conditions, because they play a role in non-causal forms of explanation which are ineliminable from the practice of biology, a role which their non-normative basal conditions can not fill.

I have two goals in this chapter. The first is to give the reader a picture of how wide a range of phenomena the concept of emergence can explain outside of the philosophy of mind. My discussion will be broad rather than deep. The second is to show in each case that the emergent phenomenon is (i) essentially normative in a way which its basal conditions are not, and (ii) cannot be eliminated from the field of inquiry, and so, insofar as one regards the field as studying real phenomena, as opposed to convenient fictions or projections of the researcher's psychology, the emergent phenomenon should be regarded as real as well.

I will focus my discussion of the role of emergence in the special sciences: computational, natural, and social. Near the end of the chapter, I will reflect briefly on other philosophical debates outside of the philosophy of mind where emergentist thinking holds some promise – such as those over the nature of intentionality, the presence of concrete instantiations of abstract properties, and Moorean non-naturalist views in meta-ethics. Put together, I hope the resulting picture will provide a more accurate backdrop for future discussions of emergence in the philosophy of mind.

## §2. Emergence and Patterns

### 2.1 Computational Patterns

The so-called "re-emergence" of emergence from intellectual dormancy over recent years has been fueled most notably by the success of emergence at explaining phenomena in computational sciences. Emergent behavior has been considered "one of the fundamental characteristics"[390] of an Artificial life system, and is the basis of most Artificial life studies.[391] Emergence is used to describe the appearance of high-level "life like" phenomena in a purely computational system where every event is fully determined by low-level rules.

The most prominent example of emergent computational patterns can be found in cellular automata. Cellular automata were developed by Stanislaw Ulam and John von Neumann in the middle of the twentieth century.[392] Ronald, Sipper, and Capcarrere describe them in this way:

> a cellular automaton consists of an array of cells, each of which can be in one of a finite number of possible states, updated synchronously in discrete time steps according to a local, identical interaction rule. The state of a cell at the next time step is determined by the current states of a surrounding neighborhood of cells.[393]

---

[390] Assad and Packard (1992), 231

[391] ibid.

[392] See Ronald, Sipper, and Capcarrere (2008), 294

[393] ibid.

Popularized by Conway's "Game of Life"[394], cellular automata are noted for giving rise to complex two-dimensional patterns – "gliders", "puffers", "blinkers", and "glider guns" – whose behavior follows predictable patterns at the macro-scale, but is not readily predictable from the states of the cells which compose them and their local interaction rules.

Thus, Mark Bedau (2008) cites many properties of cellular automata as examples of "weak emergence", systems whose behavior "cannot be determined by an computation that is essentially simpler than the intrinsic natural computational process by which the system's behavior is generated."[395] For instance, the property of *indefinite growth* – "glider guns" have this property, since a configuration consisting of only a glider gun will spawn new gliders indefinitely, but a configuration consisting only of "blinkers" will not exhibit this sort of indefinite growth. Nonetheless, there is no way to predict which patterns will or won't exhibit indefinite growth *without actually running the computational simulation*.[396]

Running a simulation and watching the outcome is a synthetic procedure, rather than an analytic procedure.[397] Assuming the natures of computational

---

[394] ibid., 295

[395] Bedau (2008), 162

[396] ibid., 163

[397] See discussion on this topic in Chapter 4. Of course, each step of the simulation is analytic and a priori: it follows directly by application of the rules. However, the process of running the situation in order to see what happens – a kind

patterns are transparent to us, it follows that these patterns are ontologically synthetic with respect to their basal conditions. Thus, by my account, Bedau's "weak emergence" is rightly categorized as a kind of *ontological* emergence: there is some small novelty in the nature of the glider gun which is not present in the nature of its base.

Of course, there is a tendency to identify the source of emergent patterns in the observer, rather than in the phenomenon itself. Ronald, Sipper, and Capcarrere (2008) argue that emergence lies in the phenomenology of "surprise" experienced by an observer who perceives it given a certain epistemic background.[398] However, it may be that these patterns are represented as surprising *because* they are ontologically something over and above their component parts: surprise is the result, not the constitutive cause, of emergence. The pattern becomes unsurprising only when someone is in an epistemic position to have the synthetic bridge laws needed to predict it.

---

of experimentation or "trial and error" – is a posteriori. Having seen how the outcome of the simulation follows a priori from the rules and starting conditions, the belief that the simulation will have such an outcome may *then* be justified a priori. However, the proposition that the simulation has such an outcome remains synthetic: the outcome is not contained in the rules or starting conditions, as shown by the fact that it could never have been learned by any means except an a posteriori experiment.

[398] Ronald, Sipper, and Capcarrere (2008), 299-301

## 2.2 Natural Patterns

Many patterns in nature have been discussed as possible examples of emergence: the waves in sand dunes or bodies of water, the shape of a hurricane, the formations of flocks of birds or schools of fish, or the mounds constructed by termites.[399] Often, it is found that these formations *are* predictable, but only by means of the same sorts of computer simulations used to generate emergent computational patterns. Once again, the procedure by which they are deduced is synthetic rather than analytic.[400]

For example, defending the claim that there are genuine novelties in nature, James Crutchfield cites as emergent "the convective rolls of Benard and Couette fluid flows, the more complicated flow structures observed in weak turbulence, the spiral waves and Turing patterns produced in oscillating chemical reactions, the statistical order parameters describing phase transitions, and the forms appearing in biological morphogenesis."[401]

Even in these fields of study very close to physics where no one is tempted to regard phenomena as remotely "spooky", we find a kind of difficulty in predicting the qualities of the phenomena which arise given only a knowledge of various component parts. While natural patterns are physical things with largely physical natures, insofar as part of what they are is a *pattern*, they may be regarded as

[399] Crutchfield (2008), 269

[400] See Chapter 4 on this topic.

[401] ibid., 270

having some novel element in their nature which is something over and above their physical nature.

## 2.3 Chemical Patterns

In Chemistry, Anderson (1972) discusses the distinction between the sort of reductionism which is accepted for chemical explanations – that everything can be fully explained in terms of physical particles and fundamental laws – and the "constructionist thesis" that one could start from those laws and reconstruct the universe. In fact, Anderson argues, chemistry is not always constructible in this way. For instance, there is a failure of *symmetry* in the distinction between "left-handed" and "right-handed" sugar molecules or the existence of an electric dipole moment.[402] Similarly, Scerri (2006) argues that Brian McLaughlin (1992) has oversimplified the extent to which chemical properties are reducible to quantum mechanics, misstated the importance of quantum mechanical explanations in chemistry to the decline of emergentism, and falsely implied that quantum chemistry enables predictions of how two elements might react together. In the current state of quantum chemistry:

> . . . we can predict *particular properties* such as ionization energies but *not chemical behavior*. In the case of compounds what can be achieved is an accurate estimate, and in many cases even predictions, regarding *specific properties* in the compounds that are known to have formed between the

[402] Anderson (1972). Similar failures of symmetry are discussed in the context of emergence by Morrison (2006).

> elements in question. Quantum mechanics cannot yet predict what compounds will actually form. Broad's complaint about the inability of mechanistic or classical chemistry to predict the properties of elements or the outcome of chemical reactions between any two given elements remains unanswered to this day. [403]

I have no reason to doubt that in a future or idealized chemistry all natural patterns and behaviors of chemical compounds will be predictable on the basis of fundamental physics, given that so many are predictable in this way, even if it is correct that some are not yet. However, if we find that the prediction requires *synthetic* bridge statements, rather than analysis, we may regard novel chemical patterns as a very weak kind of ontological emergence.

## §3 Natural Norms in the Life Sciences

### 3.1 Natural Norms

In one sense, physics is the most fundamental science, and all of biology can be, and is being, reduced to physics. We can explain hereditary genetics in terms of DNA, DNA in terms of chemistry, and chemistry in terms of physics. In another sense, biology is something more than physics. A revolution in quantum physics is unlikely to change biological theories of natural selection. To understand peacock tails, it is more important to understand competition for mating privileges than it is the physio-chemistry of feathers. To understand the physical structure of a neuron,

[403] Scerri (2006), 6. Emphasis added.

it is helpful to know what the brain is *for*.[404] Reducing biology to simple fundamental physical laws does not give us the ability to reconstruct all of biology given only the laws.

Bauchau (2006) has argued that is compatible to be a reductionist about biology in the former sense and an emergentist in the latter sense. If we accept emergentism about the computational patterns in artificial life simulations, and artificial life provides a successful model for biological life[405], then we should regard biological life as emergent in the same way: it can be predicted *a priori* from the fundamental laws only by means of simulation.[406] This is a weaker kind of emergence – a kind which is compatible with a sense of reduction – but one which is ontologically significant: there is something about the nature of biological phenomena which is not merely analytic give the nature of lower level phenomena.

Likewise, Mayr (1996) in defending the autonomy of biology from physics, holds that biological concepts are not reducible to the concepts of physics, particularly given biology's interest in distant historical causes as well as proximate causes, and that:

> Many properties of systems, such as higher levels of integration, cannot be explained by a study of their isolated components. The integration of systems results in the emergence of new properties because "the whole is [often] more

[404] Bauchau (2006), 37

[405] Heudin (2006), among many others, holds that it does so.

[406] ibid.

than the sum of the parts." The emergence of new properties is characteristic of higher levels in any hierarchy of systems, even in inanimate ones.[407]

What is distinctive about these biological properties? In this section, I hope to make the case that it is a *normative* element which is an essential part of biological properties but not present in physical properties. What I advocate have sometimes been called "natural norms".

Natural norms do not depend upon any agent's having established, intended, appreciated, understood, or accepted the norm.[408] At the same time, they are not a purely descriptive notion: they are something over and above a statistical average. Like the norms of biological function discussed in Chapter 5 (e.g., "the heart should pump blood"), these normative properties are generally acknowledged to supervene upon the fundamental physical facts, yet are not likely to be analytically derivable from the fundamental physical facts. Insofar as there is a case to be made that these natural norms are real and necessary for scientific explanations, there is a case to be made that natural norms generally – like biological functions specifically – are emergent.

In her final book, *Natural Goodness* (2003), Philippa Foot notably dedicated a chapter to the topic of "Natural Norms". Foot's discussion focuses on the work of Michael Thompson (1995) on species and generics. Generics are statements like "Tigers have stripes" which give defining characteristics of species membership, yet

[407] Mayr (1996), 102

[408] In this, I align with Burge (2010), 314-315

remain true even when individual members of the species violate the criteria (some tigers don't have stripes). Thompson observes that attributing species membership to an individual requires the individual to meet some, but not all, of these (often changing) criteria, making the judgment a normative one. Foot draws out the conclusion that biology is rich with normative properties which make reference to a role in the *life cycle* of a given species. Thus, an owl should have a certain quality of vision and a deer a certain degree of swiftness – any less is a defect – insofar as these play a role in hunting (for the owl) and fleeing from predators (for the deer), parts of the species' respective life cycles.[409]

The work of Fransisco Varela and others on *autopoiesis* resonates with emergentist themes.[410] Biologists routinely switch between two domains: the domain of physical and chemical laws as such, and the specifically biological domain which selects certain physical and chemical events as especially *significant* (a normative quality). Varela considers a study of a bacterium swimming in a sucrose gradient. The properties studied by the researcher, such as *flagellar beat*, are only interesting because the bacterium as a unit points to these properties as especially relevant; apart from the bacterium as a unit, properties like flagellar beat are no more significant than any other physical or chemical transformations. Selecting one compound of micro-level properties like this to form an especially "significant" macro-level whole is something that some believe is essentially perspectival. Varela

[409] Foot (2003), 25-30

[410] For an overview of the history and motivations behind this approach, see Damiano (2012)

agrees, but rejects the typical psychologistic view on which this perspectival quality is located in the *observer*. Instead, he locates the perspectival quality in the "autopoietic system" itself – roughly, the bacterium is a kind of self-organizing unity which depends upon an underlying set of physical-chemical properties, and yet actively maintains its distinct identity from them through a kind of reflexive feedback.[411] While there are obvious similarities between this view and emergentism, the form of emergentism I discuss here is not committed to biological entities or norms having an essentially perspectival quality.

I will not be drawing further from the literature surrounding the work of Foot or Varela, and leave it for the reader to determine to what extent either of these views are consistent with the form of emergentism I am advocating. One will notice both similarities and contrasts with some of the examples of "natural norms" in the life sciences which I now turn to.

For convenience, I have sorted uses of natural norms in the life sciences into six categories: natural expectations, natural groupings, norms of group membership, natural roles, norms of success and failure, and norms of directed action. Many of these categories arguably apply to higher-level phenomena as well, like traffic patterns and social groups (and perhaps some lower-level phenomena too, like geological or climatic patterns), but I will leave it to the reader to apply these to other domains.

[411] Varela (1992)

## 3.2 Natural Expectations

Suppose a tree is planted by a creek. One year, it rains at normal levels, but the creek's path is diverted away from the tree by natural forces. The tree dies for lack of water. What caused the tree's death? The diversion of the creek, obviously: were it not for that, the tree would not have died. Of course, it's also true that, had it rained more than usual, the tree would have survived. But there is no reason to expect more rain, whereas it is in some sense natural to expect the creek to continue on its path. Causation often seems to depend on a background of *natural expectations* – a way that things are normally "supposed to go", in the absence of intervention or manipulation. These need not be expectations held by a subject. A natural expectation is whatever state of affairs a natural system makes it reasonable to expect, whether anyone expects it or not.[412] Natural expectations are norms which emerge from statistical and probabilistic properties, but are not themselves merely statistical properties.

Recently, Christopher Hitchcock and Joshua Knobe (2009) have argued, using survey data, that that the ordinary folk concept of causation is heavily influenced by normative considerations – where these norms can be part of the way the world is "set up", without any implication of moral blameworthiness. Consider the following case:

[412] Two distinct natural systems may thus give conflicting natural expectations: it is reasonable to expect it to be a sunny day outside, given average climate patterns, yet not reasonable, given a local weather system.

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine. One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit.
>
> . . . people were more willing to say that the red wire touching the battery caused the short circuit than they were to say that the black wire touching the battery caused the short circuit.[413]

Both of the following counterfactuals are true: "were the red wire not to touch the battery, the machine wouldn't have short circuited" and "were the black wire not to touch the battery, the machine wouldn't have short circuited." These represent two equally true causal dependence claims. However, the fact that the black wire is *supposed* to touch the battery and not the red wire means that the red wire is *more causally responsible* than the black wire for the short circuit, leading to the judgment that it is more appropriately designated as the cause.

There doesn't seem to be anything especially "folksy" about this ordinary concept of causation – it is the same concept which figures in many scientific explanations. Whenever it is *more natural to expect* that $p$ will occur than that $q$ will

[413] Hitchcock, C., & Knobe, J. (2009), page 28 of manuscript.

occur, where some effect *r* causally depends on both *p* and *q*, *q* will be deemed more causally responsible for *r* than *p*. An infestation of invasive bark beetles is causally responsible for the death of the trees in the forest, more so than trees' absence of natural anti-beetle defenses. The failure to find shelter is causally responsible for the squirrel's death by hypothermia, more so than the predictable weather patterns of an average cold winter. The absence of insulin is more responsible for the symptoms of type 1 diabetes than the mere presence of glucose in the blood: it is reasonable to expect glucose (a person's got to eat), and reasonable to expect insulin (that is how the body is supposed to function). This would remain true even if, at some future point, a majority of the human population were to suffer from type 1 diabetes.

Again, suppose that there are two variants of a gene, *r* and *r**, and that people with *r** ('non-readers') suffer from a condition that makes them less likely to learn to read than people with *r* ('readers'). Our background knowledge tells us that it is natural to expect that people will have the ability to learn to read if taught. It seems appropriate to say that *r** is the cause of why the non-readers do not read, even in cases where therapy might be available to overcome the effects of *r**. However, it does not seem appropriate to say that *r* is the cause of why the readers read, because there are many other events which are less natural to expect than

having *r*, such as a decent education, which are thus more causally responsible for reading than *r*.[414]

Ordinary causation[415] plays a role in the life sciences. Ordinary causation depends on a background of natural expectations – a sense of what is "normal". Natural expectations are normative properties which emerge from statistical properties, but are themselves statistical properties. So, emergent natural norms play a role in the life sciences.

### 3.3 Natural Groupings

From these causal patterns or natural expectations of a system, there emerge certain natural groupings of molecules into cells, cells into tissues, tissues into organs, and even collections of individual organisms into species. Again, by "groupings", I do not mean to imply that there must be some mental activity of grouping carried out by a subject. I mean that there is a normative property which specifies a certain grouping as the most natural one. In the absence of any norm, there is no rule to specify grouping neighboring cells of a similar sort into a single instance of bone tissue, as opposed to grouping together some of the cells in the bone area with other nearby cells in the bloodstream, or distant cells in the brain. There is no law of physics which demands we consider the stomach lining to be part of the

---

[414] This example is adapted from James Woodward (2010), who attributes it to Richard Dawkins. Woodward's own account of causation, and of this case, is far more sophisticated than the account I offer here.

[415] As opposed to specialized uses of "causation" that involve transfers of force or energy and apply in physics.

stomach, but not the beef tripe being digested in the stomach. Physics gives us no non-arbitrary reason to regard one cat as a member of *felis catus* along with the other cats, as opposed to a member of the species *delphinapterus leucas* alongside the beluga whales. Yet these are not arbitrary choices.

Consider the apparent paradox of cell differentiation: that there exist cells of different types within a single organism, even though each cell contains the same genetic information.[416] Liver cells, red blood cells, and neurons are all very different groups of cells – but the chromosomes within each do not differ. One cannot predict given only the genetic composition of the cell what sort of cell it will become. Rather, the presence or absence of various proteins in the cell will cause different genes in the cell's DNA to be expressed or repressed, and the expression or repression of one will cause the presence or absence of other proteins, which in turn cause the next gene to be expressed or repressed, and so on.[417] In the end, about a third of the genes are expressed. (The red blood cells end up with hemoglobin; the others don't). Cells will change which genes are expressed by responding to external signals. The process which gets us from identical DNA to distinct *groups* of cells with differing biological functions could be modelled as a network of cellular automata.[418] In other words, the nature of a given cell emerges from processes and interactions within the whole complex system responsible for assigning distinctive functions, not simply its

[416] Weisbuch (2006)

[417] See Albert (2002), Chapter 7.

[418] Weisbuch (2006)

material composition. Groupings into liver cells, kidney cells, etc., are a kind of natural norm.

Species differentiation provides another good example of a "natural grouping." Ernst Mayr, who is responsible for formulating the concept of a species which is dominant in modern biology, insisted that species are real, concrete phenomena of nature, the principle units of evolutionary laws, not arbitrary artifacts of the human mind.[419] That species change over time, but are not mere aggregates of their members, is consistent with an emergentist view on which species depend on their members for their existence (a new species can emerge) and yet have a nature over and above them. Thus, Mayr endorses emergentism as one of the "two pillars of the explanatory framework of modern biology," next to genetics.[420]

The same considerations also apply to larger-scale natural groupings. The hierarchy of natural kinds and the divisions into vertebrate and invertebrate, mammal and reptile and so forth, all involve a kind of grouping norm which is not found in any individual species.

Biology and other life sciences could go nowhere if we were not for directing research attention to these groupings – but the groupings themselves are not part of the laws of physics. Some groupings are simply *more natural* than others given our knowledge of the whole system. Grouping things together in a "natural" way is not a

---

[419] Mayr (1996b), 262-263

[420] Mayr (1997), 19. He defines emergence as follows: "that in a structured system, new properties emerge at higher levels of integration which could not have been predicted from a knowledge of the lower-level components.

merely analytic procedure given the underlying causal patterns and qualities of the component parts, but a synthetic one. So, emergent natural norms are part of the life sciences.

### 3.4 Norms of Group Membership

Given these groupings, there arise norms which apply to the particular members of the group. These groupings emerge from the kind as a whole and not from the properties of the individual member. For example, even though a dog may be born with only three legs due to a genetic defect, and nothing in the composition of the dog indicates that it ought to have four legs, we still regard it as a genetic *defect* or *anomaly*, because it is normal for members of the class of dogs to have four legs.

Consider how population thinking has replaced typological thinking in biology.[421] The Aristotelian brand of essentialism about species identification has been abandoned – biologists no longer look for one single trait in common between all individuals within a species, but instead recognize that *the whole population of the species is more fundamental*. Instead of reducing species membership to low-level properties of individual species members, the whole *population* is recognized as having the properties which are characteristic of a species. The whole organized population of a species is the entity subject to evolutionary laws. While the population has the properties it has because of the properties of its individual members, the whole population in turn gives rise to certain *normative* properties for its members. For example, it is *normal* that a particular cat will groom itself, and

[421] see Mayr (1959) and Sober (1980)

that a particular sparrow will fly. This is because cats (as a species) groom themselves, and sparrows (as a species) fly. An individual cat or sparrow may fail to follow the norm. But understanding the norm – which arises from the whole species – is an ineliminable part of understanding why particular species members do the things they do. The norm for species members is more than an approximation or generalization from the properties of the individuals in the species – it's partly definitive of what the whole population of the species is.

The older, morphological conception of a species focused on identifying definitional traits – necessary and sufficient conditions which some organism had to have to be a member of the species. On this conception, it would be *analytic*, given that Charlie is a cat, that Charlie has the various cat-properties. However, this conception does not account for gradual species changes or adaptations. On Mayr's modern definition, "species are groups of interbreeding natural populations that are reproductively isolated from other groups".[422] Notably, "reproductive isolation" does not rule out individual cases of hybridization between members of related species – it is the normal state for the species as a whole. The morphological characteristics of species (coloration, number of legs, keen sense of smell) are normative, but not definitional.

Norms of group membership are a standard part of contemporary biology. The fact that they are fuzzy and that they change over time through evolutionary processes (unlike Aristotelian views), yet are not mere statistical properties of their

[422] Mayr (1996b), 265

aggregated members (unlike reductivist views), supports the claim that they are emergent properties.

### 3.5 Natural Roles

One particular norm of group membership that can emerge is that of a *natural role* within the group. Consider the roles played by mitochondria within cells, the roles an organ within a physiological system, or the roles of a system within the life of the body as a whole. Or, at a larger scale, consider an ant colony. Within the colony, various ants will have different natural *roles*. Ants do not have rich conscious lives, so far as we know, nor do they collaborate together to distribute the various roles. Nonetheless, through a kind of biological signaling system, the colony acts as a whole, distributing roles to the various members.

A "role" in this sense is more than just a series of causes and effects, though it may supervene upon causal patterns. It relates what the particular member does to the activities of the group or system as a whole.

In Chapter 6, I discussed biological functions. The sense of "biological function" I relied upon in that chapter is sometimes called the *selected effect* (SE) model or *etiological* model,[423] grounded in a history of natural selection. For example, the eye is for seeing, and in seeing it fills its role in the body, and this 'for seeing' is grounded in the contribution of seeing (and thus the eye) to the evolutionary fitness of the ancestors of the organism. Biological functions in this sense are one type of natural role property. However, there is a broader sense of "biological function" which is sometimes used, which grounds the function in its

[423] Brandon (2006), 267

*causal role* (CR) within a more complex system.[424] This is a different type of natural role property. As I noted in Chapter 6, the notion of a "causal role" is *dependent upon* some underlying teleological explanation.[425]

For example, features of the Australian landscape play a (CR) function in the life-cycles of its invasive rabbit population, even though they have no (SE) function in their lives, insofar as these landscape features have no historical connection to their evolution. Another example comes from Brandon (2006):

> Consider the characteristic shape of some particular species of tree, e.g. the American beech. That shape is a product of the branching pattern of the tree, in particular, the distance between branch points and the angles of the branches. These in turn are properties of the growth dynamics of the meristem. In other words, the overall shape of the tree emerges from the dynamics of meristematic growth.
>
> Let us suppose that natural selection has played no role in molding the dynamics of meristem growth in the American beech. That being the case,

---

[424] The causal role account of function remains controversial. Advocates of this approach include Brandon (2006), and Amundson and Lauder (1994); the account has its origins in Cummins (1975).

[425] In Chapter 4, section 4.1, I argued that every event causally depends on a whole series of events, but only a few of these causal dependencies count as causal "roles", because they are relevant to the teleology of some larger system.

> then clearly the meristem growth pattern has no SE function, has no evolutionary purpose. But it does have a CR function in determining the overall shape of the tree. That is because à capacity of the large system, the overall shape of the tree, is explained in terms of the growth patterns of the meristems. [426]

At the many different levels at which they occur, natural roles provide examples of "natural norms" which play an essential role in scientific explanations of the world. Because they are so thoroughly *natural* (that is, not the least bit spooky or mysterious) it may be easy to overlook their *normativity*, an essential feature of "roles" but not of physical processes as such.

### 3.6 Norms of Success and Failure

Closely related to the idea of a natural role or function is the notion of "success" at fulfilling that role or achieving that function. This applies whether the "end" at which the function aims is promoting fitness on a particular occasion or not. Success and failure should be understood as criteria of *adequacy* – the system succeeds by doing "well-enough" to achieve its function or role. Obviously, one system might be better than another at fulfilling a specified function. Norms of success and failure exist at every level at which a natural role or function exists.

It is not controversial to hold that there is a sense of "success" or "failure" with respect to biological functions that promote fitness, but it may help to see that success and failure are norms that can arise independent of whether an individual

[426] Brandon (2006), 274-275

organism's fitness is enhanced by success. For instance, at the organ level, the lungs and circulatory system succeed in fulfilling their role in respiration, even if a creature breathes in a toxic gas that is quickly carried by the bloodstream to the brain, damaging fitness. At the level of a group of organisms, among wolves, the role of mating is reserved for the alpha-female.[427] While the alpha-female succeeds at its role by mating, the other females in the pack succeed at their role by *not*-mating – even though this does not maximize their individual fitness.[428] At an even higher level, within a complex ecosystem, the various parts of the ecosystem – the flora and fauna, the river itself – may each play a role in the overall health of the ecosystem, at which they can succeed or fail. A species may *fail* at this role by being too *successful* at enhancing its own fitness: as in the case of the deer which overrun the forest. Thus, there are measures of success and failure at biological roles which do not appeal to maximizing individual fitness on a particular occasion.

The laws governing successes and failures of these sorts can be quite robust and justify many of our predictions. However, categories like "success" and "failure" are undeniably normative. Insofar as we accept laws of this sort as real, not merely convenient fictions, we should accept a kind of emergent natural norm.

### 3.7 Norms of Directed Action

The life sciences do not merely study events. They also study the *actions* of organisms as a whole. Pacific salmon swim upstream in order to spawn. Canada geese migrate north for the summer, and south for the winter. Tyler Burge (2010)

---

[427] Derix, R., Van Hooff, J., De Vries, H., and Wensing, J. (1993).

[428] Of course, given the scarcity of prey, it does maximize the fitness of the pack.

has developed the notion of *primitive agency*: the directed action of a whole organism towards some end, even if the end is in no sense represented by the agent. A whole variety of "natural norms" arise once a primitive agent has entered the scene. For instance:

*Consistency*. The actions of a primitive agent should be consistent. A fox chasing its prey should not stop to take a drink of water. A plant should not grow roots away from the source of water. The salmon do not alternate between swimming up and down the stream at the same time. Organisms which do not act with this sort of "natural rationality" do not often survive to reproduce. Those that do survive, follow this normative standard by acting in a way that does not contradict the goal of their action.[429]

*Success*. Just a thing can be successful or unsuccessful at fulfilling its natural role or function, a primitive agent can perform an action in a way which is successful or unsuccessful at achieving the goal of the action. The salmon may succeed at making it upstream, or run out of energy. The fox may catch its prey, or lose it. Note that an animal's success at performing an action does not imply success at fulfilling its biological function, or vice versa. Suppose the prey is part of a fox-hunter's plot to catch and shoot the fox. In this case, the fox which succeeds at the action fails at its biological function, and the fox which fails the action enhances likelihood of survival.

*Benefits*. We may also quantify over goals of an action or goals for an organism, and ask whether the action was *beneficial* or *harmful* for the organism.

---

[429] Notice that it is not *logical* consistency or *logical* contradiction we are talking about here: there is no contradiction in swimming upstream and downstream.

This is a normative category – though not necessarily one with any prescriptive force. In asking how much the fox benefits from catching the prey, we are not asking how much pleasure the fox experienced from the kill and the taste of the meat, but what functions the fox is more able to fulfill as a result: perhaps more of its young will survive. This allows us to weigh the trade-offs in terms of harm and benefit for an organism and try to predict what strategies it will use to maximize its benefits.[430]

Similar sets of norms apply not only to individual actions, but to coordinated *group actions*, such as the migration of the Canada geese. In these cases, consistency, success and failure, and benefit and harm are measured with respect to the group, rather than the individual.

In all of these cases, we can give a good, naturalistic explanation for why the organism takes the action it does and why the corresponding norm applies to that action.[431] The natural norm itself, however, has a nature which is not to be found in the phenomena which appear in the underlying explanation. For this reason, it can be regarded as emergent.

---

[430] I don't mean this list to be exhaustive. Depending upon the organism, more specific natural norms may arise for the actions of that organism.

[431] One should not misinterpret this as meaning that the norms are just shorthand for the corresponding explanation, or that the normativity itself is naturalistically reducible in some way, or that the normativity is contributed entirely by the observer's interpretive stance (this later view is found in Dennett (1989)).

## §4 Emergence and the Social Sciences

### 4.1 Overview

I wish to turn now to a few examples from the social sciences[432] where emergence can play an explanatory role. Obviously, the social sciences are about *human* society and activity, and human beings are conscious. So, consciousness (which I have been arguing is emergent) may need to be considered among the basal conditions for the emergence of complex social states – the state of a massive panic in a community, for example, supervenes upon the individual conscious panic-states of those in the community. However, the case for regarding social science explanations as emergent does not depend upon consciousness itself being emergent. Rather, emergence finds a place in the social sciences in describing how unintended macro-level group behaviors can arise from micro-level decisions.

I recognize that the concept of emergence is used in different ways in different social sciences, with greater and lesser degrees of controversy, and will not be able to explore debates within each field here. I will give a sampling of a few works in the social sciences and in the philosophy of social science which either explicitly use or strongly suggest emergentist themes.

### 4.2 Sociology and Emergence

Talk of "emergence" is widespread in contemporary sociology. Because explaining the micro-macro link is so essential to the sociological enterprise,

---

[432] In this section, I will only be discussing emergent phenomena in those social sciences which do not involve the study of representational states – thus excluding linguistics and psychology. I will discuss these in section 5.

emergence is naturally invoked to explain the apparent micro-macro gap. However, the term "emergence" is often used in imprecise and conflicting ways.[433] Helpfully, Keith Sawyer, in his article "Emergence in Sociology" (2001), clarifies the use of emergence in sociology and reconnects the discussion with the contemporary discussion of emergence in the philosophy of mind. Sawyer distinguishes two broad camps in sociology which utilize the notion of emergence.

(A) Among collectivist emergentists (so named because they tend to follow methodological collectivism), collective phenomena are "emergent" in that they depend upon the collaborative actions of individuals, but have properties which cannot be reduced to individual properties. Some methodological collectivists hold that only collective properties emerge, while others make the stronger claim that distinctive social kinds and entities emerge.

(B) Among individualist emergentists (so named because they tend to follow methodological individualism), collective phenomena are fully reducible and explainable in terms of the properties of individuals, yet are still in some sense "emergent".

Sawyer recognizes that the contemporary use of emergence in the philosophy of mind is more in line with (A) than (B).[434] He discusses various collectivist

[433] Sawyer (2001)

[434] ibid.

emergentists in sociology who at times have advocated for (A)[435] and suggests his own non-reductivistic brand of emergentism on this model.

To the extent that sociology accepts emergentist accounts of the micro-macro link like (A), it appears to support a strong kind of ontological emergence of collective properties (and, possibly, collective entities and kinds).[436]

To the extent that sociology tends instead towards emergentist accounts like (B), it either supports a kind of *ontological weak* emergence, like that in

---

[435] ibid. The major references are to Blau (1977, 1981); Bhaskar (1982); Archer (1995); additional papers besides these are cited. Sawyer alleges certain inconsistencies in their work, particularly when it comes to whether social entities should be regarded as real, whether they accept supervenience, how they account for causation, and how they understand reducibility. However, Sawyer's own discussion is at times unclear on the relationship between reducibility, supervenience, and the ontological significance of emergence. In more recent work, Sawyer (2005) has given further arguments in favor of emergence, which he interprets as a kind of "non-reductive individualism", but it is not clear that he has resolved these issues. Sawyer does offer the emergence of social networks and the emergence of "group mind" in improvisational theatre dialogues as examples of irreducibly social phenomena.

[436] However, it needs to clarify why social properties are supposed to be *ontologically* irreducible. Is it because they are essentially normative, or involve a kind of social intentionality? Why not say they are merely *epistemically* irreducible, without having a different nature?

computational patterns, or else it supports a merely epistemic kind of emergence. The answer depends on whether or not social properties are regarded as real, and whether they are ontologically reducible or only explanatorily reducible.[437]

While there is a great deal of clarification and precision needed about the role of emergence in sociology, it is a field which has made ample use of the concept. I will discuss some examples of type-A or "strong" emergence, followed by examples of type-B or "weak" emergence.

### 4.3 Strong Emergence in Sociology

Durkheim is best known for his role in founding sociology as an independent discipline, autonomous from psychology. He defended the view that there were *sui generis* social facts, over and above the psychological facts about various individuals in society.[438] He writes:

> [The autonomy of Sociology] implies that collective tendencies and thoughts are of a *different nature* from individual tendencies and thoughts, that the former have characteristics which the latter lack . . . society has no other

[437] If social properties are real, and yet only explanatorily but not ontologically reducible to individual actions, then they may be regarded as weakly ontologically emergent. However, if they are merely analytic constructs out of micro-level phenomena, or deducible from the actions of individuals, using analytic procedures alone, then they are at most epistemically emergent rather than ontologically emergent.

[438] Rosenberg (1993), 130-131

> active forces than individuals; but individuals by combining form a psychical existence of a new species, which consequently has its own manner of thinking and feeling. Of course the elementary qualities of which the social fact consists are present in germ in individual minds. But the social fact emerges from them only when they have been transformed by association since it is only then that it appears. Association itself is also an active factor productive of special effects. In itself it is therefore something *new*.[439]

Durkheim can be interpreted as advocating a kind of strong emergentism on the basis of these and other comments.[440] In the context of debates at the time about the ontological status of social facts, Durkheim seems to be seeking a middle ground between reductivistic individualism and Hegelian organicist holism. In *Rules of the Sociological Method*, he argues for a compromise view, where social facts are determined by the associations of individuals but have a *sui generis* nature. Sawyer (2002) notes that "Durkheim never used the term 'emergence'; rather, his phrase 'sui generis' was used in a sense synonymous with contemporary uses of the term 'emergent.'"

In *Suicide* (1897), Durkheim sought empirical evidence for this *sui generis* character of social facts in a study of suicide rates. Of all things, suicide would seem to be a phenomenon best explained by the reasons given by individuals. However, he

---

[439] Durkheim (1897), 310. Emphasis added.

[440] Sawyer (2002), 227-228 cites a number of sources who take this interpretation of Durkheim.

observed that suicide rates were much higher for Protestants than Catholics, higher for officers than conscripts, and higher for the newly rich than the longtime poor. This remained true, even controlling for other reasonable-seeming factors. The individual reasons provided for suicide were the same in each case and fell in roughly the same proportions: family troubles, pain, love, jealousy, and so on. So, the cause of the difference had to involve something about the group, not the individual reasons given. Durkheim found this in the degree to which individuals were integrated into society and its *social norms*. Both Catholics and Protestants disapproved of suicide, but Catholics were more tightly integrated than Protestants into their system of norms. Army officers were more tightly integrated than conscripts into a system of norms which approved of a kind of self-sacrificial suicide for the good of the unit. And the newly rich found themselves detached from any system of norms at all – they were left normless and disoriented.[441] Understanding the *âme collective*, or "group mind", was necessary to understanding the behavior of the group.

Interestingly, Durkheim appears to have been sympathetic to early British Emergentism at the time,[442] and defended a similar kind of downward causation for sociological entities. It is possible he endorsed a kind of supervenience of society on

[441] Rosenburg (1993), 131-133

[442] Durkheim (1895), 104 states: "there is between psychology and sociology the same break in continuity as between biology and the physiochemical sciences", evidently an illusion to Lewes.

individuals, much as emergentists endorsed between the mind and the brain.[443] If so, sociology as a discipline was founded on an emergentist ontology.

A more contemporary advocate of Type-A or "strong" emergence in contemporary sociology is Margaret Archer. Archer advocates a "morphogenetic" approach to social structure.[444] On her account, social structures are emergent properties – although they are products of interactions by human individuals, they have distinct causal powers of their own. Social structures cannot be reduced to their individual components, and their powers cannot be reduced to the powers of individuals. No individual or aggregation of individuals may have the power to change a social phenomenon, but the social phenomenon most certainly may have the power to change them.[445] David Elder-Vass, in defending Archer's account, gives the following example:

> Customers, suppliers, and others who interact with an organization always do so through the human individuals who occupy roles within it, but the way they interact with these individuals is conditioned by their understanding that the role incumbents represent the organization concerned, that they act

---

[443] Sawyer (2002), 233, cites Durkheim's essay "Individual and Collective Representations" (1898)

[444] Archer (1995). Unfortunately, Archer's account of emergence seems to be taken straight out of Mill's *System of Logic*, including making use of a (long outdated) analogy with the "emergent" chemical properties of water.

[445] Elder-Vass (2007)

on its behalf. Thus the existence of the organization also affects how these external individuals behave towards the individuals who are its parts. [446]

Interestingly, the philosopher John Wisdom also advocated a kind of strong emergence for collective social properties. Responding to the individualism of Karl Popper[447], Wisdom (1970) advocated a brand of "emergentism" in the social sciences as a compromise between the holistic defense of *sui generis* social phenomena and the individualist's opposition to independent social entities. According to Wisdom, the "social pathology of Great Britain" (a kind of collective, post-imperial *ennui*) was a phenomenon whose causes could only be discussed and debated at the level of the nation, not the level of individual psychology. What made this social pathology "emergent" was that it was not merely an unintended consequence of individual

[446] Elder-Vass (2007)

[447] Popper was a strong proponent of individualism in the social sciences: for him, society really was just the sum of its members. While he acknowledged that there could be unintended consequences to the actions of individuals – a kind of "invisible hand" – the social facts were ultimately just collected facts about the members of a society. However, his strong individualism may have been based on a belief that it was the only alternative an extreme holism and "historicism", the Hegelian or Marxist approach to social science on which society was an organic unit following a deterministic historical path, independent of the individuals in it. See Thornton, Stephen (2011)

behaviors, but something whose nature could never have been predicted or deduced considering only the actions of individuals:

> . . . a psycho-social depression, is an example of such an emergent phenomenon. It is not at all just a social reflection of widespread individual depression. It is not just a surprising occurrence that might have been otherwise; it is not the kind of thing at all that might or might not be expected. It is a different order of eventuality. And not only is it unforeseeable but it may even be unrecognized (like a psychological depression) after it has arisen. (And it may exercise some control over our behavior without our being aware of it).[448]

I have categorized these views as advocating "strong" emergence, insofar as they hold that the sociological facts are not ideally deducible *a priori* from the underlying relations between individuals. I am a bit wary of framing sociological emergence in this way – it seems quite plausible to me that an ideal reasoner with infinite computing powers, given a complete knowledge of the underlying facts about individuals (including their mental states), could deduce the social facts by a running through a kind of "mental simulation" of the entire social system. It is more likely that what these authors have in mind is that no deduction of the social facts could ever be performed in practice, because the deduction would require access to ontological synthetic bridge laws which we do not have access to (such as those we

[448] Wisdom (1970), 292

come to by watching a simulation), not a mere analysis of facts about individuals. In this sense, it is characterized properly as a "weak" kind of emergence – a subject which I will turn to next. Nonetheless, I believe these authors are right in recognizing this as *ontological* emergence: the essential properties of social structures and other social phenomena are no where to be found in the essential properties of the individuals which compose them.

### 4.4 Weak Emergence in Sociology

Perhaps one of the most well-known examples of emergence in the social sciences is Schelling's account of racial segregation.[449] In *Micromotives and Macrobehavior*, Schelling applies game theoretic constructs and rules about the behavior of individuals to predict characteristics of the aggregate of people – particularly in situations where the choice one individual makes, like where to sit in a cafeteria, depends on the choices made by individuals before him. Schelling found that very minor preferences on the individual level to be of the same race as a certain number of one's neighbors could lead, on a macro-scale, to sharp lines of segregation. Because no individual making the decision intended maliciously for segregation to be the result, and yet segregation supervenes upon the facts about individual preferences, segregation could be understood as an emergent phenomenon. It is arguable that, even if all racism among individuals in our society ceased, the small-scale interactions of individuals could still lead to a malignantly racist society with a great deal of inequality. This is not merely a case of epistemic difference between our ability to understand macro-scale and micro-scale

[449] Schelling (1978), 135-164

phenomena, but an ontological one, insofar as 'pushing towards segregation' properly explains the phenomena at the macro-scale but does not explain the behavior of any individual at the micro-scale.

Mark Granovetter, in "The Strength of Weak Ties" (1973), similarly undertakes to relate micro-level and macro-level behavior in sociology. Granovetter considers how our weak interpersonal ties (acquaintances rather than close friends) form the basis for the transmission of information (or rumors, or job opportunities) across vast distances and between different social strata. While our closer friends are most likely to have a set of closer friends who overlap with our own, our weakest acquaintances are most likely to have sets of acquaintances which do not overlap with our own. From these two-person dyadic ties, large scale social networks and communication patterns can emerge.

Of course, the sort of emergentism suggested by Granovetter and Schelling's work is much different from the emergentism of Durkheim and Wisdom. Both Durkheim and Wisdom are advocates of *strong emergence* in the social sciences: there are macro-level social phenomena which *cannot be deduced* on the basis of a knowledge of individual psychologies. Granovetter and Schelling's work describe a phenomena much more like the weak emergence which occurs in computational patterns: a higher level phenomenon is deducible and predictable based on a knowledge of the lower-level phenomenon. Nonetheless, these macro-level phenomena should be understood as ontologically emergence, insofar as they have some novelty in their nature which is not included in its base.

## 4.5 Emergence and Social Norms in Anthropology

Much like the "natural norms" in the life sciences, social norms might be regarded as emergent properties. I understand social norms (the norms which govern the behavior of a society) to be distinct from moral norms (norms with prescriptive force).[450]

Social norms (or cultural norms) can include such things as: what clothes are fashionable, how many hours a day is reasonable to work, the respective duties of hosts and guests, and how, at what age, and with whom it is acceptable to marry. Following the wrong set of norms identifies a member of a community as an outsider – someone who can't be trusted to cooperate.

Consider a few of the processes generally believed to be involved in the development of social norms – a process which may occurs without much in the way of individual intentions to do so:

*Cultural learning* is the process by which individual members of a group pick up on which norms to follow.

*Prestige bias* is the tendency to model one's behavior over the most successful individuals.[451]

---

[450] Although moral norms may be instantiated as particular social norms, social norms may also be grossly immoral. Behind apartheid stood a powerful social norm, but not a moral one.

[451] One who wants to be a good philosopher may try to model his tactics on the most successful philosophers (though he is sure to pick up no shortage of bad habits along the way). Prestige bias tends to favor non-cooperative behaviors, whereas

*Conformist transmission* is simply taking those nearby as a model for one's behavior, a process which spreads the behaviors of the majority.

*Punishment* of norm violators, and of those who fail to punish norm violators, aids the survival of cooperative norms.[452]

*Reputations*, a practice of social accounting, make punishment cheaper. By collectively keeping track of when members of the community violate cooperative norms and docking their reputations for it, the costs of punishment can be spread out over the entire community.[453]

Both social norms themselves, and the package of phenomena which provide for the transmission of social norms – reputations, prestige, punishment, and conformity – could be regarded as emergent phenomena. Despite not being *moral* imperatives[454], they nonetheless have an essentially *normative* element in their

---

conformist transmission can support cooperative and even altruistic behaviors, so long as they are not too costly.

[452] Henrich and Henrich (2007), 67.

[453] ibid. 65. For example, those with a bad reputation may be denied the benefits of cooperative interactions with others or assistance in times of need. Of course, communities may become forgiving of certain violations over time or in light of other positive contributions to reputation, but other violations may prompt expulsion from the community or serious physical harm.

[454] That is to say, while social norms may be expressed in terms of a society's *beliefs* about morality, and may prescribe behavior for those in the society, they are as likely to be morally vicious as they are to be virtuous.

nature, and are unlikely to be merely analytic given descriptions of the social behavior patterns which they supervene upon.

### 4.6 Emergence and Collective Actions

Systems of social norms make possible the emergence of goal-oriented collective action. Collective action is how societies effectively get around the "prisoners dilemma" problem, which otherwise would plague us. Rationally, we ought to be much more suspicious of each other, since everyone has a lot to gain by cheating when someone else cooperates, and a lot to lose by cooperating when someone else cheats. However, norms which operate at the social level have produced a society in which most of us choose to cooperate most of the time – and expect others to cooperate most of the time. We each take an individual loss to gain a collective benefit. Because these norms operate at the social level, no individual need be aware of what the group norm is, what a certain pattern of behavior is *for*, or how the cooperative behavior will maximize benefits to the group – each individual need only think to herself "the group is doing it, so I guess I will do it too," or "I'll be nice to people, since they're being nice to me." We might even speak of *group intentionality* on this basis: the group may intend to produce some effect which no individual member does.

In the case of group action, the nature of the group's action may be distinct from the nature of any individual's action in the group, even though the group's

action is fully determined by the actions and interactions of the individuals in the group.[455]

In this section, I have given a variety of examples of cases where weaker and stronger kinds of ontological emergence can prove useful for explaining the world: at the level of pattern-emergence, at the level of the life sciences, and at the level of whole societies. I will now move on to examples involving the emergence of representational states in perception, language, and thought.

## §5 Emergence and Representational States

### 5.1 Emergence and Perception

There are a multitude of accounts of perception in philosophy. "Perception" is sometimes used to discuss (i) any subjective phenomenal state ("I perceived a horrible monster in my dream"), (ii) the mode of presentation of an object in perception ("I perceived a sheep, but it was really a white rock"), (iii) the registration of information even by an inanimate object ("this car perceives your seatbelt isn't on"), (iv) our conceptualized judgments about what we perceive ("I perceived that Wal-Mart was closed"), (v) any causal connection to the sensations of a subject ("in his sweat, he perceived global warming"), and (vi) the conditions under which it is acceptable to say of someone, "he perceives", such as that a person has the ability and right to say "I perceive" (cf. Ordinary Language Philosophy). However, for this section, I will be focused on only one type of account of perception. On this account,

[455] For much more sophisticated philosophical defenses of collective intentionality (and even collective responsibility), see Bratman (1992) and French (1984). Smiley (2011) offers an extensive bibliography.

(i) - (vi) are legitimate but secondary senses of "perception", which depend upon the primary and strict sense of perception, as a distinctive kind of *objective sensory representation*. Though it is distinct from conceptual or propositional representations, like them perception is an intentional mental state – it is about the world – and representational *accuracy* with respect to the world is characteristic of perception.

Tyler Burge has recently given an account on which the notion of representational accuracy is closely tied to the evolutionary history of a whole organism and its capacity for directed action.[456] Perception arises from biological systems – systems for sight and hearing are the most obvious of these systems, though a capacity for proprioception and other capacities may also count as perceptual. At the same time, the nature of perception (its *about*-ness and its representational accuracy) is something over and above anything we find in the nature of biological systems.

Burge's account of perception in animals is non-reductive. He resists attempts to reduce the representational quality of perception to merely the registration of information or some series of causal patterns. Perception aims at *veridicality*. "Accuracy" or "veridicality" is not part of physics, but is an essential part of understanding animal and human psychology. In explaining what perception and perceptual accuracy are and how they link up to the underlying biology, Burge explicitly appeals to the notion of a natural norm. On this topic, he writes:

[456] Burge (2010)

> By 'natural norm' I do not mean naturalistically reducible norm. I mean a level of performance adequate to fulfill a function or a purposiveness, and that constitutes an explanatorily relevant kind, independently of any individual's having a positive or negative attitude towards the function or the norm . . . usually, natural norms are also independent of any individual's appreciating them . . .
>
> There are natural norms constitutively associated with representational functions, as well as natural norms constitutively associated with biological functions. I think that for every function, it is apriori that there are various natural norms associated with it.
>
> There are natural norms for perceptual representation. The primary natural representational norm that is constitutively associated with perceptual capacity is to perceive things as they are – to form *veridical* perceptual representational. Veridical perception fulfills perception's primary constitutive representational function.[457]

Burge connects the notion of an individual function to the idea of "primitive agency", the sense in which a spider *as a whole* jumps on, bites, and eats its prey, in contrast with processes which are merely internal to the spider, like digestion.[458] The phrase "primitive agency" seems to suggest a kind of emergentism.

---

[457] ibid, 312

[458] ibid., 327

Is Burge's account of perception an emergentist account? While Burge does not seem to endorse emergentism directly, his account of perception could easily be interpreted as an emergentist account. By denying the possibility of reducing perception to some non-normative information-state, Burge seems to acknowledge that perceptual accuracy is something more than a mere information-preserving arrangement of physical structures. Instead, Burge holds that information states become representational states when they reach a level of organization which is susceptible to a semantic standard of "being correct" – an appeal to natural norms. Burge's account also clearly accepts some form of supervenience of representational states upon the physical structures of an organism and their relations to the environment (including causal-historical relations that secure reference). The nature of perception is something over and above the nature of the biological constitution of the perceiving animal, even though it is fully determined by it. If one accepts an account of perception which has these features, then one should regard perceptual representations as emergent.

### 5.2 Emergence and Language

Although it is different in many ways from perception, language is also a representational system. A distinctive aspect of human language is that it allows speakers to combine discrete parts in order to construct novel, meaningful wholes. One can utter a sentence which no one has spoken before, and the sentence still be understood by speakers of the language.

There are two ways in which language has been discussed as an emergent phenomenon. The first involves the emergence of language (particularly, syntax) from some underlying neurological structure, in an attempt to explain how it is that

we can acquire language. The second involves emergence as an explanation of the hierarchy of various levels in language.

**5.2.1 Emergence in Language Acquisition**. How did we learn our native languages? Our parents generally did not sit us down and give us language lessons. In spite of this, we attained a level of fluency that it takes someone learning a second language years or more to achieve. Since the work of Chomsky (1959, 1965), the underdetermination of linguistic outputs by stimulus inputs has been recognized as effectively refuting behaviorism. A child learning the phonology and syntax of a language is not engaged in mere mimicry. In fact, young children often utter ungrammatical sentences which they have never heard before in a kind of "overcorrecting", showing that they have mastered one rule but not another. Sometimes known as the "poverty of the stimulus" argument, linguists have observed that of the many possible structures which a language might have, language learners are never exposed to enough information to rule out most of the options.

The Chomskyan interpretation of this data is to conclude that there is a kind of innate knowledge or grammatical theory present in children, a "Universal Grammar" which all human languages in the world have in common. The process of learning a language requires only enough information to select from a small set of options, to indicate that one's own language has this or that feature (for instance, that the language is head-initial rather than head-final). Our knowledge of various aspects of language can be localized as distinct modules within the brain (hence, the differing effects on language ability of differing sorts of head injury), which come "pre-programmed" in some sense to detect and organize linguistic data.

However, emergentist theories within the study of language acquisition have lately begun to challenge the Chomskyian view. MacWhinney (2001, 2006) and Hollich et. al. (2000) are among those who advocate this approach. Rather than viewing the process of learning a language as a bit-by-bit copying of the various rules of the language to particular nodes or modules in a pre-programmed brain (irregular verbs go here, regular verbs go there), an emergentist approach views knowledge of the rules of language as emerging out of an entire neural architecture, context, and history.[459]

The emergentist joins the Chomskyan in rejecting behavioristic, reductivistic, and eliminativistic accounts of language and accepting the existence of innate structures in the brain which constrain language.[460] However, whereas the Chomskyan explains the poverty of the stimulus in terms of the innate representations of Universal Grammar; the emergentist holds we can explain why certain rules and structures exist in human languages, and how children learn

---

[459] O'Grady (2010) makes a thorough but unpersuasive attempt to give an emergentist account of syntax of this sort. O'Grady's account suffers from explaining purely accidental features of English as the necessary result of such features as "computational efficiency"; examples from a broader array of languages would help clarify things.

[460] Thus, while emergentism in other fields appears as an anti-reductionist's response to reductivistic orthodoxies, emergentism in language acquisition is motivated by a desire to link back together high-level phenomena to their low-level base, against a more strongly dualist tendency.

language in spite of the poverty of the stimulus, in terms of a "high-level" phenomenon which emergences from the complex "low-level" system of observed linguistic behavior, the interactions within a neural network, biological limitations, the pragmatics of conversation, efficiency given limited memory, and so on – even though it cannot be reduced to these.[461] Although emergentism in the field of language acquisition remains a minority position, and has failed so far to produce the explanatory depth and breadth of the Chomskyan approach, it remains an interesting application of emergentist thinking to a serious problem in linguistics.[462]

**5.2.2 Levels of Language.** A different way in which to view language as an emergent phenomenon is not to begin with the emergence of language from neural structures and social interactions, but to notice the application of emergentist thinking to the relationship between the "higher level" and "lower level" phenomena in language.

At the most basic level, spoken human languages consist of sounds. Phonetics is the field which studies these sounds as such. But it is not the sounds as such

---

[461] O'Grady (2011). O'Grady notes that emergentism in the field of language acquisition is often associated with connectionist approaches to psychology. Evidence usually cited for emergentism involves the pattern of "exceptions" typically found whenever a feature of universal grammar is proposed; "Universal Grammar" is seen as more of a very strong norm than an exceptionless innate universal.

[462] Notably, MacWhinney (2001) explicitly appeals to the widespread use of emergence in the biological and physical sciences as reason to accept emergence in linguistics.

which are the building blocks of individual units of meaning, but *phonemes*: the "roles" played by the sounds or *functional units* which make a difference to which word is expressed, yet may be realized in many different ways. Two dialects of the same language may have the same phonology, but drastically different phonetics – contrast the many phonetic realizations in different geographic regions of the English phoneme /r/. Two distinct phonemes in one language may be one phoneme in the other. Consider how 't' in 'talk' isn't pronounced the same way as the 't' in 'bat' or the 't' in 'bottle': in English these sounds are all one phoneme, /t/, but in other languages they are distinct phonemes. Phonology depends upon the underlying phonetics and is constrained by it, but the phonology of a language cannot be predicted simply by a list of the sounds which occur in it.[463] The nature of a phoneme depends on its role in the morphological structure of the language, not what it sounds like. During the "Great Vowel Shift" in English – when the /i/ in "wipe" stopped sounding like the "ee" in "weep"– the phonetics shifted, but the phonology remained the same.

[463] The phonology of a given language weakly locally supervenes upon its phonetics – there is no difference in phoneme without a difference in sound – but does not globally supervene on it: that is, there could be a language identical to ours with respect to which sounds were uttered at what times, but in which the phonology of the language is entirely different, given that it has different units of meaning.

In turn, the morphemes of a language – its smallest units of meaning[464] – are composed out of phonemes, but the nature of a morpheme has almost nothing to do with its phonological composition and almost everything to do with its role in the syntax and semantics of the language. A different morphology could been realized identically in a language with very different sounds. Without sounds, there would be no case or tense markings, but a case or tense marker is something with a nature over and above the nature of the sounds which constitute it. So, there is some sense in which morphology might be regarded as emergent from phonology, and phonology from phonetics, insofar as the nature of each higher-level phenomena does not depend on the parts which compose it.

To go down a different track: the semantics of a language – the meanings of the morphemes in a language – supervene upon the actual uses of those words by individual speakers in that language. If, over time, individual speakers adjust their use of a word and begin to mean something new by it – if "meat" is applied only to edible flesh and not to the inner part of a nut, for example – then the meaning of the word as a whole will change. Nonetheless, the nature of meaning is something over and above a kind of averaging-out of the uses of individuals. Individuals use words the way they do because words have the meaning they do in their community. The majority of individuals might even *misuse* a specialized term in a way which is inconsistent with its meaning. So, semantics might be regarded as emergent.

The syntax of a language is at an even higher level of organization, something which can be and often is studied entirely independently of the low-level

[464] e.g., the word *repossessing* contains 3 morphemes, and *possessive* contains 2.

sounds or words which realize the language. A syntax is a kind of logical structure – the order in which subjects, verbs, and other particles stand to each other and combine with each other – which depends for its existence on being instantiated in a meaningful and communicable way, but which has a nature which has little to do with the meanings of the words which are arranged within its structure and even less with the sounds those words are composed of. So, syntactic structures might be regarded as emergent phenomena.

### 5.3 Emergence and Intentionality

A third representational system is thought. Like language and perception, thought is characterized by intentionality: our thoughts and beliefs can be *about* various properties and persons and entities in the real world. Suppose that I am unaware that Mark Twain is identical to Samuel Clemens, and I believe that they are two different men who lived at different times. When I say (or think, or believe) that Mark Twain is not Samuel Clemens, I say something false, even though I believe that I have said something true. This is because my uses of the names "Samuel Clemens" and "Mark Twain" refer to the same individual whether I know it or not.

But how is it that words (or thoughts, or beliefs) are *about* something in the world? Again, this is a tricky and well-worn philosophical problem, and I do not hope to answer it here. However, insofar as many philosophers have concluded that intentionality is something which is over and above any possible physical description of the world, I can advocate that those who accept this view also consider themselves to be emergentists. It is not clear how bodies or brain states can be *about* something in the way that my thoughts are about something: my thought has a very different

nature from my brain state. Nonetheless, were it not for the brain state, I would not be thinking this thought – and I could not think another thought without a corresponding change in my underlying brain state.

One might respond that our thoughts are physical, but their physical composition is larger than the body or the brain – it includes the causal history of our thoughts. Consider how "Aluminum" is about the metal which is actually used in soda cans and foil wrap (even if I think incorrectly that the word refers to shiny plastic). Any world which is an exact physical duplicate of our world would be one in which my thought of "aluminum" would be about the same thing – the facts about the historical origins of the world are sufficient to ground its reference to the metal used in soda cans and foil wrap. Nonetheless, it does not follow that what it is to be the referent of "aluminum" is just this historical origin. In fact, we could imagine a scenario in which a very different historical pattern occurred, but reference was still secured from "aluminum" to the metal. So, the intentionality of our thoughts emerges from our brain states and these causal-historical chains, but has more to its nature than them.

The sense in which my thoughts are *about* various objects or properties in the world could be plausibly described as a case in which the intentional aspect of my words, beliefs, and thoughts *emerges from* certain facts about my neurology as well as the causal-historical chains and facts about my community which link my use of the words or concepts to what they refer to in the actual world, without being reducible to these facts.

## §6. Emergence and Instantiations

### 6.1 Emergence and Abstract Property Instances

I have been discussing cases where emergence is a plausible hypothesis in our scientific investigation of complex phenomena in the concrete world. However, I also want to briefly suggest a few cases in which emergence might be appealed to as an explanation for certain philosophical conundrums which arise for issues of how abstract properties relate to their concrete instantiations in the world. There is, of course, a vast and historically extensive literature already on these questions. I have no hope of interacting successfully with this whole literature in the limited space available here. Instead, my goal is to briefly suggest that one plausible position on these issues is equivalent to emergentism: one on which the instances of these norms depend upon the existence of the physical world despite having a nature which is in no part physical.

**6.1.1 Norms of Reasoning.** Logic captures the truth-preserving relations between various propositions. Logical relations are not physical relations. However, logical relations could be regarded as providing norms of thought and reasoning, or norms of proper inference.[465] The inferences which logical norms of reasoning apply to are events which occur in space and time.

In addition to logic, there are also *epistemic norms* which might be regarded as abstract, necessary truths: there are norms for when one ought to believe an uncertain claim, and when one ought to reserve judgment, or when one ought to

[465] This is a very recent area of debate in philosophy, on which much remains to be said. See MacFarlane (2004).

make an inference to best explanation, and when one ought to seek further evidence. Again, these norms apply to events of reasoning which occur in space and time.

The instantiations of these norms supervene on physical events: the norms of how credible one ought to consider a claim will not change unless the situation presents more evidence; whether an argument a lawyer is making is valid will not change unless the lawyer changes his argument. However, the norms are not in any sense physical events – unlike "natural norms" their nature does not depend upon the nature of the physical world *even in part*.

**6.1.2 Mathematical Properties.** Consider the cases of abstract mathematical properties – the property of *two-ness*, or the property of *being a square*. According to many traditional accounts in philosophy, these are necessary properties which are not themselves located in the concrete world. Nonetheless, in certain situations, they are instantiated in the concrete world: there might be two bananas on a roughly square countertop. How is it that abstract properties can be located in a concrete world? One response is Platonism: the concrete objects *participate in* the form of the abstract object, where this "participation" has a mysterious quality to it. The alternative is some kind of attack on the traditional account of abstract mathematical properties – to identify them as reflecting the limits of our experience (Kant) or truths of our language (Carnap) rather than metaphysical realities.

An emergentist, it seems to me, could plausibly hold that instantiations of abstract mathematical properties were emergent in a way which is consistent with the traditional view, but which does not require Platonic talk of participation in the forms. Instead, the emergentist could hold that instantiations of abstract mathematical properties "arise from" their physical basal conditions, while having

natures which are partly abstract and mathematical in addition to being concrete. From the placement of the bananas and the construction of the countertop, it is necessitated that the bananas have two-ness and the countertop as square-ish-ness. These instances of two-ness and square-ish-ness arise from the physical arrangement of the bananas on the countertop. Nonetheless, their essential natures are partly abstract: the bananas arranged as so have the abstract property of being two in number.

**6.1.3 Unity of a Fact**. Consider the set of problems raised by what is sometimes known as "Bradley's regress", after F. H. Bradley (1893). There is some sense in which the various constituent parts of a relational fact are (i) ordered and (ii) unified into a whole. The fact that Angelina loves Brad is distinct from the fact that Brad loves Angelia, as well as distinct from the set with the elements {Angelina, Brad, loves}. Suppose one holds that for there to be a difference between two things, the two things must have different constituents. It follows that there must be some fourth constituent C which is part of the fact that Angelina loves Brad, which differentiates it from the fact that Brad loves Angelina. Or, by a similar argument, there must be some fourth constituent C in the fact that Angelina loves Brad which is not a constituent of the set {Angelina, Brad, loves}. However, whatever this fourth item is, the same problem arises: if C unifies the fact, then what unifies C to the fact? If C grounds the asymmetrical ordering of the elements in the fact, then what grounds C's asymmetric ordering within that fact? A similar problem arises with propositions (rather than facts) and with objects (when objects are conceived of as bundles of properties).

Bradley's problem is essentially one of how a whole (the relation, relational fact, relational proposition, etc.) can have a property which is not analytic given its constituents. Arguably, an emergentist could respond that nothing further is needed than the constituents of the relational fact for the relational fact, yet the relational fact is not fully grounded in its constituents – and so the fact fits the standard definition of ontological emergence.

**6.1.4 Musical Properties.** Musical properties might also be regarded as abstract properties which come to be physically instantiated, and so qualify as "emergent" in this sense. For example, one might regard "dissonance" as a real property which can be instantiated in physical wave patterns (with a serious physical explanation), a property which exists independent of whether anyone hears the dissonance as dissonant or experiences a feeling of tension or having one's teeth set on edge. On this account, a subject *perceives* the dissonance that is present in a pair of musical tones; it is not created in the subject's psychology. Yet the dissonance is not itself a physical property either – there is nothing in the laws of physics to offer us rules of counterpoint. Thus, some have argued that some musical properties should be regarded as emergent properties – see Wright & Bregman (1987).

**6.1.5 Economic Properties**. Economics is a field in which sociology meets mathematics. Thus, the properties studied by economists might partly fall into the category of sociological properties discussed earlier, in the weak emergence of macro-level phenomena from micro-level phenomena, but might also fall into the category of instantiations of complex mathematical properties or functions. Thus, the market value of a good might be considered as being emergent out of the dispositions of individuals to trade various goods in the marketplace. Rarely in these individual

trade decisions does anyone have the market-value of the good in mind; each individual only has the values that he or she places on various goods in terms of trade-offs for other goods. The market value of a good is something over and above these decisions (and subject to its own economic laws at a macro level) and yet it supervenes upon them. Further, whether or not one feels confidence towards "invisible hands", the market is notable for providing an efficient (in some ways) distribution of goods without any individual engaged in the market having in mind the goal of providing for an efficient distribution. So, the efficiency of markets is also to some degree emergent from these individual decisions, supervening upon them but different in nature.

### 6.2 Emergence and Vagueness

The issue of vagueness in language seems related to the issue of emergent natural patterns. Consider vague predicates like "is bald" or "is a heap". When objects to which these predicates apply are considered to be composed of some denumerable quantity of more basic objects, one can construct a sorites paradox: a man will never go from being bald to non-bald by the loss of one hair, nor will a heap of sand go to being a non-heap by the loss of one grain of sand. However, when objects to which these predicates apply are considered as ontologically basic rather than mere compositions of parts, such that their parts are non-denumerable, sorites paradoxes cannot be constructed (since sorites paradoxes rely on mathematical induction, which applies only to denumerable sets).

We should expect emergent properties to correspond to vague predicates. Since the nature of the emergent phenomenon is not entirely dependent upon the nature of its parts, it is not properly described as constructed out of some

denumerable set of low level-phenomena, but instead the whole is more basic. Yet, since an emergent phenomenon remains partly dependent upon its parts for its existence, there are still clear cases in which the absence or presence of some quantity of low-level phenomena will determine the existence of the emergent phenomenon.

Thus, although it is beyond the scope of my project to do so here, one might develop an emergentist account of vagueness. On this view, vague phenomena have a nature which is not defined by their component parts – rather, their nature is defined by their relation to some greater whole. Nonetheless, they depend upon their component parts for their existence. Thus, when one makes large changes to the component parts, the vague phenomenon ceases to exist, but there is no small change to the component parts which can produce a change in the nature of the vague phenomenon, since its nature does not come from its parts but from the whole of which it is a part.

## 6.3 Emergence and Moral Non-Naturalism

**6.3.1 Moral Non-Naturalism as Emergentism**. In contrast to the "natural norms" and "social norms" discussed earlier, moral norms have a particular kind of prescriptive, imperative force which makes them obligatory – one who fails to keep them is rightfully blamed. Non-cognitivists hold that this "prescriptive force" should clue us in that moral norms aren't statements about the world at all, but rather something more like attempts to get one another to adopt attitudes of approval or disapproval. In contrast, moral realists hold that moral statements have cognitive content which can be true or false in light of how the world is: it is *true* that you ought not to cause suffering to an innocent creature for no good reason.

Moral realists disagree about whether moral properties are natural or non-natural properties. Both sides accept the supervenience of moral properties on physical states of affairs – there is no way the world could be morally different without being physically different. However, non-naturalists wish to say that moral properties are something over and above the physical, whereas naturalists wish to say that moral properties are just complicated bundles of physical properties.

*Non-naturalists* follow G. E. Moore in holding that the "badness" which makes a killing into a murder is not a natural property. Moore's argument for non-naturalism is sometimes known as the "Open Question Argument". Suppose that one offers an analysis of "goodness" in purely natural terms, so that "a more equitable society is a good thing" could be rephrased as something like "a more equitable society is the sort of thing that causes people to smile and results in fewer acts of violence," (or something more sophisticated). Well, then, it would remain an open question as to whether all this was a good thing: "are more smiles and less violence *good*?".[466]

In contrast, *naturalists* hold that badness is equivalent to some physical state. Naturalists can be divided into two camps. The analytic naturalists, following Frank Jackson (1998), hold that the moral facts are synonymous with descriptive facts about the natural world and could be derived through conceptual analysis from these natural facts. The non-analytic naturalists hold that, in spite of the impossibility of such a conceptual analysis, moral facts just are natural facts.[467] This

[466] Dancy (2006), 125-126

[467] ibid.

"non-analytic" naturalist position is committed to the view that no analysis of moral language is possible in purely descriptive terms from our epistemic standpoint, but, given an ideal reasoner, the moral properties would be ontologically analytic given the natural properties. This position differs from Jackson's position in holding that conceptual analysis is an imperfect guide to ontology, but not in denying the possibility of an ideal reasoner's deducing the moral facts from the non-moral facts by purely analytic[468] moves.

It should not to be difficult to see how Moorean non-naturalism qualifies as a form of emergentism about meta-ethical properties, since it accepts supervenience, yet holds that in spite of this moral properties are distinct from descriptive ones. Of course, moral properties, much like the epistemic norms and logical norms discussed earlier, are perhaps best seen as abstract properties whose *instantiations* are the emergent phenomena. A killing is a natural phenomenon located in space and time, but the killing's being subject to moral evaluation, and thus being a *murder*, involves the emergence of a non-natural moral property.

Notably, emergentism offers a new response to a question which has long plagued moral non-naturalists: *why* do the moral properties supervene on the natural properties, if they aren't identical to them?[469] Whereas traditional non-naturalism has been silent on this question, the emergentist non-naturalist can add a bit of explanation: "it emerges!"

---

[468] That is, ontologically analytic: we might not have the concepts to perform any kind of conceptual analysis

[469] see Blackburn (1985).

Of course, how effective an explanation "it emerges!" is depends upon whether or not one sees it as *ad hoc*, or whether one sees it an explanatorily powerful concept which applies to a wide variety of phenomena. I hope that, in the course of this chapter, I've shown how many different kinds of emergence of varying degrees of strength feature regularly in the natural world. We should expect novelties to arise within nature, and moral properties are one of those novelties. Thus, what explains the supervenience of moral properties is that they are yet another instance of emergence. Of course, one might ask a further question: why do things emerge at all? This is a mysterious question, which I can only touch on briefly in chapter 8.

**6.3.2 The Emergence of Tragedy**. There is an interesting idea which is suggested by emergentism about moral norms, over and above non-naturalism as such. In my work, I have been discussing emergence as a *synchronic* phenomenon wherever it occurs. I have not been making any claims about the development of social or biological phenomena over time as emergent, or asserting that "high-level" phenomena are in some way chronologically later than "low-level" phenomena. However, emergence about moral phenomena suggests – though it hardly implies – an interesting thesis for the diachronic emergence of instances of moral properties (that is, the property of being subject to moral evaluation). I will briefly sketch this thesis before concluding the chapter.

Suppose that social structures and conscious states are part of the supervenience base for instantiations of moral norms. While moral norms themselves are abstract universals, whether or not an event in space and time is rightly subject to moral evaluation might depend on whether or not certain complex

structures exist or not. Given structure $s_1$, moral evaluability for phenomenon $p_1$ emerges; given structure $s_1$, moral evaluability for both $p_1$ *and* $p_2$ emerges, and so on. As social structures develop over time, the subject matter of moral evaluation grows progressively larger – not that the society itself subjects more things to practices of moral evaluation, but that features within the society which were previously innocent now have the emergent property of being instances where a judgment of moral rightness or wrongness applies (or vice-versa).

For instance, persons as such might begin as the objects of moral evaluation: one ought to live in accordance with one's proper function as a person. Person is a distinctly moral category over and above that of human, and the well-functioning of a person has moral force in a way in which the well-functioning of a human does not. Given this moral category, and a certain level of organization of one's social group, the subject of moral evaluation might change: the question is not just whether you are a well-functioning person, but whether you are well-functioning member of the group, so that when the group acts, your actions aren't in conflict with it, and so that when other members of the group cooperate with you and offer aid to you, you reciprocate and cooperate back. Given the moral commitment to being a good member of the group and the further organization of the group, there emerges the normative category of a legitimate authority or power over the group, capable of issuing commands the group genuinely ought to follow. From the practice of legitimate commands, there emerges the moral evaluation of what is actually commanded, and whether or not the outcomes or consequences of the command are harmful or beneficial on the whole to everyone in the group. At the same time, the commands might be evaluated in a different way – not on what is beneficial, but

whether it is just or fair. Given the increasing significance of individual identity, morality may become individualized again – there emerges the question, am *I* acting in such a way that it could be justly and beneficially commanded for everyone to act? Or, if relationships to others become more important, the question might be: am I fulfilling my responsibilities? Whom am I responsible for caring for? And beyond these, still other moral categories might be realized.[470]

What is notable about this picture is that, while the emergentist is committed to moral realism, the emergentist is not committed to the view that *ought* implies *can*, or that one could coherently fulfill all of one's obligations. Far from it: one might have an obligation under one moral system, and an opposing obligation under another equally *real* moral system. It is consistent with the possibility of genuine moral tragedies: of cases where one might violate one's obligations in order to fulfill them. Suppose that my relative commits a heinous crime and is unwilling to turn himself in: do I turn him over to the law, or do I protect him from it? The outcome is tragic either way: either it is the tragic story of how justice trumps over individual loyalties, or it is the tragedy of how family loyalties are more fundamental than law. Suppose one is stuck in a trolley-case, and must cause the death of one as a means to saving the lives of many. Perhaps it is best to say that one has two conflicting obligations, and, whichever ends up trumping the other – the more fundamental obligation or the less fundamental one – the outcome will be tragic.

---

[470] This paragraph is very loosely inspired by the work of Jonathan Haidt on moral psychology.

Of course, the emergentist about moral norms is not in any way committed to this *diachronic* account of moral emergence. However, insofar as disagreement about moral questions is often cited as evidence for moral anti-realism, the emergentist responds by providing an account on which moral realism would quite naturally be accompanied by disagreement over which real moral norm to violate, and which to fulfill.

## §7. Conclusions

Emergence is far from being an unprecedented kind of explanation devoid of explanatory power. Emergence is everywhere. In a wide variety of fields, from the computational sciences, natural sciences, and social sciences, through the study of representational systems and philosophical questions about the instantiation of abstract norms, emergence appears as a plausible and well-regarded, if sometimes controversial, thesis. To the extent to which these fields discuss ontological issues, emergence is meant ontologically, to indicate a difference in the nature of the phenomenon from its basal conditions, and it is not primarily being used as a merely epistemic claim. For many fields, debates over emergence are intended to guide methodological questions: should we investigate the higher-level phenomena in its own right, or should we regard the higher-level phenomena as nothing more than our own psychological projection upon a lower-level phenomena and focus our attention there?

In my review of the literature, it was surprising to find that the most common objections against emergence in these fields were *not* that it was somehow unscientific, fishy-smelling, or unlike the sorts of explanations used in other fields. Far from it: emergence tends to be seen as a popular (if sometimes faddish)

explanation which other scientific fields already rely on. Instead, the primary objections to emergence are taken straight out of the philosophy of mind: in my own informal survey of literature on emergence, it seems to me as though Jaegwon Kim's objections to the coherence of downward causation were the most widely cited and discussed objections, followed at a distant second by concerns about the coherence of emergentist claims about non-reducibility, predictability, and supervenience.

The philosophy of mind should pay more attention to what happens in these fields, much as these fields are paying attention to what happens in the philosophy of mind. In no way would I advocate the idea that philosophers should uncritically accept whatever ontology a scientific field hands them (especially when a special science may very well have a motive embrace a richer ontology in order to legitimize its research program). However, when philosophers do develop critical ontologies, they should keep their picture of the cosmos updated – even if it turns out to contain forests with vines descending from the trees, and less of a desert landscape.

Chapter 8

# CONCLUSIONS

## §1 Summary

### 1.1 Overview

In this work, I have defended Emergentism against four objections: (i) the objection that emergence is either incoherent or collapses into non-reductive physicalism or traditional dualism, (ii) the objection that emergence entails widespread causal overdetermination, (iii) the objection that emergence is wrongly defined in terms of the impossibility of the ideal *a priori* deducibility of emergent phenomena from their basal conditions, and (iv) the objection that emergence lacks explanatory power because it applies only in the special case of consciousness.

### 1.2 Coherence

An emergentist says that the world being the way it is physically is *sufficient for* the phenomenal, but the phenomenal remains *distinct from* the physical. I proposed interpretations for "sufficient for" and "distinct from" which distinguish emergence from other views. Traditionally, the emergentist's distinctness claim is interpreted as involving the failure of the *a priori* deducibility of emergent phenomena from their basal conditions. However, to whatever extent the *a priori* is taken to be a guide to metaphysical necessity, this seems to conflict with the emergentist's sufficiency claim, often interpreted as meaning that emergent properties supervene upon their physical basal conditions with metaphysical necessity. The alternative is to adopt a weaker interpretation of the sufficiency claim, but this threatens to collapse into traditional dualism.

I interpret the distinctness claim as indicating that a particular form of ontological explanation, or grounding, fails to hold between physical phenomena and emergent phenomena: emergent phenomena are not grounded in their physical basal conditions. As many have recognized, supervenience is not a grounding relation – one thing can necessitate another without explaining it. A physicalist needs "super*duper*venience" to hold between the physical and the phenomenal: the phenomenal must both supervene upon the physical and its nature must be wholly grounded in the nature of its physical base. However, the emergentist can plausibly deny this grounding claim, distinguishing emergentism from physicalism. The emergentist simply denies both:

(D1) mental properties are identical to physical properties; and

(D2) there is some ontological explanation of mental properties in terms of physical properties, other than identity, such that mental properties are no addition to ontology.

The emergentist is a dualist insofar as the emergentist accepts:

(GD): mental phenomena are not wholly grounded in physical phenomena.

However, to distinguish emergentism from traditional dualism, I hold that the emergentist must reject the claim that phenomenal properties have an

independent existence from their basal conditions, a claim which at a minimum a traditional dualist must accept:

> (PEIP): mental phenomena do not wholly depend for their existence on physical substances.

I then interpret the emergentist's *sufficiency* claim as indicating the total and exclusive dependence of emergent phenomena on their basal conditions. This can be expressed as a kind of supervenience what I call *manipulative* necessity, as opposed to the strictly stronger notion of metaphysical necessity:

> *x manipulatively necessitates y* is true in world *w* iff if there is no successful manipulation *w** of *w* for *y* through any *z,* where *z* is not identical to *x* and not on the path between *x* and *y*.

The manipulative supervenience of the phenomenal on the physical is consistent with the possibility of a world in which the same physical properties are present, but the phenomenal properties fail to emerge. However, it strictly denies that there is any means through which phenomenal properties can be manipulated except by an intervention upon their basal conditions, where the notions of "manipulation" and "intervention" is taken from Woodward (2005) and do not imply an action by some agent or other. Thus, while there may be two physically identical possible worlds which differ with respect to the distribution of their phenomenal

properties, *nothing else* could differ between such worlds, including the set of fundamental laws: neither world would be accessible from the other.

On the resulting definition of emergence, emergent properties are those which (i) wholly depend for their existence on (or manipulatively supervene upon) their basal conditions, yet (ii) have natures which are not grounded in the natures of their basal conditions. Because traditional dualists will reject (i), and physicalists will reject (ii), emergentism can be distinguished from both positions in a coherent way.

### 1.3 Overdetermination

Jaegwon Kim argues that an emergentist's commitment to supervenience, the distinctness of mental and physical properties, and realism about mental causation, entails either epiphenomenalism or widespread overdetermination. His argument relies on two principles:

> *(CCPD) Causal Closure of the Physical Domain*. If a physical event has a cause (occurring) at time *t*, it has a sufficient physical cause at *t.*
>
> *(EP) Exclusion Principle*. No event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination.

Suppose a mental event M at *t* causes a mental event M* at *t+1*, where P is the subvening base of M at *t*, and P* is the subvening base of M* at *t+1*. First, Kim argues that since M causes M*, and M* supervenes on P*, then M must be a

sufficient cause of M* in virtue of being a sufficient cause of P*. After all, had M occurred but P* not occurred, then M* still would have occurred, since M* supervenes on P*; it can't be that M adds something which is missing to P* or that M and P* jointly, coincidentally are sufficient for M*. So, M must be a sufficient cause of P*. Second, Kim argues that since M causes P*, then according to (CCPD), P* must have a sufficient cause at $t$ – namely, P. P* thus has two sufficient causes: P and M. But this is not "a genuine case of overdetermination," so it violates (EP).

I argue that there are two senses of "sufficiency" which should be distinguished: a deductive sense of sufficiency, and a completive sense of sufficiency. On the completive sense, A is sufficient for B if and only if nothing else is needed besides A for B to occur. On the deductive sense, A is sufficient for B if and only if it is not possible for B to occur and A not to occur. Overdetermination is a problem for an event which has two deductively sufficient causes, but not for an event which has two completively sufficient causes. While CCPD might be interpreted as guaranteeing that every physical event has a deductively sufficient cause (a question I remain neutral on), the downward causation by M of P* is only plausibly considered a case of a completively sufficient cause: nothing else is needed but M for P*, but it's not as though one could construct a valid argument to model the derivation of P* from M.

After adding my own response to several other plausible responses to Kim's argument, I argue that an emergentist has many ways of avoiding Kim's overdetermination problem, and need not be lead by it into epiphenomenalism.

## 1.4 Cosmic Hermeneutics

The distinctness of emergent properties has been defined since C. D. Broad in terms of the non-deducibility of emergent phenomena from their basal conditions, even by an ideal reasoner. However, it is not clear why a failure of ideal deducibility should entail anything about the ontological distinctness of the phenomenon which cannot be deduced. After all, (i) many have held since Saul Kripke that the properties of $H_2O$ are not ideally deducible *a priori* from the properties of water, yet water is not ontologically distinct from $H_2O$, while (ii) many non-naturalists in meta-ethics hold that ethical properties are *a priori*, and thus deducible given a description of the natural properties, yet maintain that the ethical properties are ontologically distinct from the natural properties.

Tracing through the recent history of ideal deducibility tests, I attempt to connect this project of "cosmic hermeneutics" with my interpretation of the distinctness of emergent properties in Chapter 2, in terms of emergent properties being those which are not grounded in their subvening base. I argue that, instead of *a priori* scrutability, emergentists should look to *analytic* scrutability as a means of marking the distinction between emergent and non-emergent properties. On my account emergent properties may be *a priori* deducible from their subvening base, but they will only be deducible by means of non-analytic inferential processes on the part of an ideal reasoner. When an ideal reasoner is limited to analytic inferential processes, and the deduction is carried out in the language of ontology (as opposed to a natural language, or a language of concepts), then all of the cases in which one thing wholly grounds the nature of another will be cases in which the ideal reasoner can deduce the grounded thing from its ground. That something is water will be

deducible from its being $H_2O$, since in the language of ontology each will have the same nature. On the other hand, the ethical properties will not be deducible from the natural properties, assuming non-naturalists are correct, because ethical properties do not follow by mere analysis from the natures of descriptive phenomena, even if they are *a priori*.

When a phenomenon supervenes on a set of basal conditions, but its nature is not analytically scrutable from the nature of its basal conditions, then we have reason to think that the non-deducible phenomenon is emergent.

### 1.5 Nomological Danglers

J. J. C. Smart's objection of that emergentism fails to mesh with a scientific perspective on the world is perhaps one of the most compelling and influential objections to emergentism about the mind. On my interpretation of Smart's objection, Smart's complaint is that ontological emergence is an *ad hoc* explanation which applies to only one case: phenomenal consciousness.

However, I argue that emergence has wide explanatory currency across the special sciences at many different levels: at the level of computational and natural patterns, at the level of biological functions and other natural teleological properties, at the level of human psychology, at the level of standards of representational accuracy, at the level of abstract properties and normative epistemic and ethical properties, and at the level of phenomena in the social sciences. These cases are generally categorized as superficial or "weak emergence", irrelevant to the ontological "strong emergence" alleged to happen in the case of consciousness. However, I argue that there is no reason to regard weak emergence in the special sciences as a *merely epistemic* claim. I will argue instead that there are cases of

*ontological* weak emergence: cases where an emergent phenomenon has a nature which is over and above that of its subvening base. On my account, there are many things which are ontologically emergent in our world, and many kinds of emergence – emergent consciousness is only one of them.

When examples of purportedly emergent phenomena in the special sciences are considered on a case-by-case basis against a physicalist backdrop, it is reasonable to assume that some form of ontological reduction to physics nonetheless holds, in spite of the absence of a reductive analysis of the special science into physics, because emergence would require giving the phenomenon under consideration a radically different account than everything else. However, when this whole variety of emergent phenomena is allowed to form the backdrop against which individual cases are considered, ontological reductivism to physics suddenly becomes far less plausible, and emergence becomes a form of explanation which unites our theories of phenomena at a variety of levels to one another. Emergence isn't an *ad hoc* form of explanation which applies only to consciousness, but a form of ontological explanation which goes all of the way up and all of the way down.

## §2 A Final Objection

### 2.1 Mysterious Questions

One challenge remains for emergence which I have not addressed yet. Consider that emergence may be proposed, by a kind of "inference to the best explanation", as an answer to fill the explanatory gap between the phenomenal and the physical. At the same time, in this role, emergence is what Horgan (2009) calls an "unexplained explainer" – while emergence explains why there is an explanatory gap, emergence itself is by definition inexplicable. Someone might sensibly ask: why

should these physical basal conditions give rise to those emergent properties? Why should anything have emerged at all? What fills *this* explanatory gap?

On my account, emergent phenomena *cannot be fully explained by* their underlying physical conditions, and yet emergent phenomena *cannot be explained by anything but* their underlying physical conditions.

Sometimes there arise mysterious questions for which, even were we to know the answer to every other question, we would have no idea how to answer the original question. It is not for lack of understanding the question that we find ourselves unable to answer it based on any other question. We understand the question fully – but we can't answer it, and no further information would allow us to answer it. At best, we can speculate that there is some further fact which would push the question forward another level: if someone asks "why *p*?", we can respond "because *q*", and hope that this buys us some time before we are asked, "why *q*?"

## 2.2 Personal Identity

One such question, as Parfit (1984) has identified, is the question of personal identity. There are answerable questions of personal identity, such as questions about the qualitative differences between persons and questions about under what conditions a person survives or can be held responsible. However, this question is not answerable:

(Q1) Why am I me and not someone else?

The question most naturally expressed by these words is perfectly understandable. Of course, it is *a priori* that I am me and not someone else – for "I

am someone other than myself" is a logical contradiction. It is impossible that I should not be myself. Nonetheless, the question being asked is not the trivial one it appears to be, which should be no more mysterious than why *1 = 1*, but instead a question which seems both contingent and interesting: why should I be this person that I am and not some other person, such as one in another body? One can push back the question: "you are your body" or "you are your immortal soul" or "you are that emergent consciousness for which that body serves as the basal conditions." However, it only stalls a moment before one asks, "but why should it be *my* body or soul and not someone else's?" or "why should I be this body or soul and not an exact duplicate of it?" For any further fact we specify to answer the question, no matter how esoteric or how concrete, the same question will arise for that fact – and so on, endlessly, hopelessly.

### 2.3 Identity

I bring up the example of personal identity in part because it is the sort of question which motivates some traditional dualists, and someone might think that emergentism provides an advantage in answering the question. Since consciousness is often linked to personal identity, and emergentism has been offered as an explanation of consciousness, it might be natural to think of emergentism as offering an explanation for personal identity. We can address the answerable questions about personal identity – survival, memory, and responsibility – with facts about qualitative conscious states, and so emergentism is relevant to these questions. However, the unanswerable questions about personal identity are of the sort which remain even when one has specified all of the qualitative conscious states, and so

they are no more answered by emergentism than they are by physicalism or by traditional dualism.

However, my primary motive in bringing up personal identity is as a particularly salient example of a genuinely mysterious question. There are many other such questions in philosophy, which all have the same character. For instance, in addition to the problem of personal identity, but there is the problem of identity as such:

(Q2) Why is this thing *this* thing and not some other thing?

Such a problem is expressed by Black (1952) and his question of why, given two qualitatively identical spheres, the two spheres are nonetheless numerically distinct. Even though all of the qualitative facts are in, the numerical identity questions remain unanswered. One could invoke a fundamental identity fact, a *haeccity*, to answer the question. This would push back the question a step. However, the question can always be asked, "why is this thing's thisness *this* thisness and not some other fundamental thisness?"

## 2.4 Reference

The problem of identity leads naturally into the problem of reference. On the one hand, what my words or thoughts refer to isn't answered, even if one has an answer to all of the questions about what is going on in my head. I believe Napoleon was the Emperor of France. But my use of "Napoleon" would refer to that particular individual it refers to, even if that individual spent his life as a humble Corsican peasant and someone else became Emperor of France; similarly, "Water" would refer

to that natural kind it refers to in virtue of its causal history, $H_2O$, even if part of my concept of water was that "water is the stuff in our lakes and streams", and the stuff in our lakes and streams turns out to be XYZ.

On the other hand, what my words or thoughts refer to isn't answered, even if one has an answer to all of the questions about what is going on in the world outside of my head. Given only a description of the world's history and the behavior of myself and everyone else in my linguistic community, how would one determine whether my use of "Napoleon" referred to the whole man, or only to a proper part of the man? How would one determine whether my use of "water" referred to $H_2O$ in general, or only to H2O prior to the year 2525, and XYZ thereafter? One might think that the facts about reference are fixed by a certain pattern of causes, for example: our utterances of "water" are caused by H2O. But when we trace the causal history of a term back to its referent, what guides the story we tell? It isn't that these causal dependencies are particularly salient compared to others: our utterances of "water" are caused by a whole host of things, including thirst. The *intentions* of speakers to refer to this or that are necessary to tell us which history of causes and effects to follow.

Consider, then, an example adapted from Evans (1982):

> Suppose, for example, that on a certain day in the past, a subject briefly observed two indistinguishable steel balls suspended from the same point and rotating about it. He now believes nothing about one ball which he does not believe about the other. This is certainly a situation in which the subject

cannot discriminate one of the balls from all other things, since he cannot discriminate it from its fellow.[471]

Contrary to Evans, it has seemed clear enough to Burge (2010) and others than a subject can meaningfully say, "I am thinking about *that* steel ball", and that his use of "ball" will refer to one of the balls – the one he intends to refer to, that he has in mind – and not the other. Yet, if this is so, it is not in virtue of any *phenomenal* difference between the two balls for the speaker, since all will appear the same to the speaker regardless of which ball he refers to, nor is it in virtue of any *behavioral* or *physical* difference in the worlds in which the speaker is referring to one of the steel balls and not the other. Thus, what the speaker is referring to remains a question which can't be answered even if one has access to all of the physical and phenomenal information – unless one is the speaker himself or herself. One can provide *evidence* that a speaker or speech community is referring to this thing or that, but one could have all of the evidence in, and still be left without an answer for why it is that the speaker's thought of "*N*" refers to *N*.

(Q3) Why does "*N*" refer to *N*?

## 2.5 Time

We cannot neglect the "mystery of time", as Bouwsma (1954) put it. Once again, the fact that it is *now* seems to be as good a fact as any other. But why should

[471] Evans, 90

it be now, and not some other time? A four-dimensionalist about time, who holds that other times in the past and future exist in the same sense that the present exists, must offer some further explanation for why it is that we are actually at the time we are at and not at one of the many other times in the past and the future. Yet a three-dimensionalist about time, who holds that only the present exists, doesn't get off the hook – for the three-dimensionalist must explain why it is that of the many possible times it could be, the present is the one which really is. None of the facts about the present are the facts which make it the present. So then, we're left asking without answer:

(Q4) Why is now when it is?

## 2.6 The Cosmological Questions

Then there are those questions identified by Aquinas quite some time ago, where a perfectly good form of explanation which serves us quite well most of the time, if pursued to the very end, turns out to be interminable. Many people intuitively accept that some things cause others and that sometimes people do things for a reason. Many philosophers accept that wholes depend upon their parts for their existence, and that some facts – even necessary facts – are grounded in other facts. Yet pursuing these forms of explanation leads us quickly into the cosmological questions:

(Q5) "What caused the cause of everything?", or "Why did the event which explains every event happen?"

(Q6) "What's that which everything is for, for?", or "What is the point of the point of everything?"

(Q7) "Why does the supervenience base exist?", or "What does that on which everything depends for its existence depend upon for its existence?"

(Q8) "What grounds the fundamental facts?" or, "In virtue of what are the fundamental facts fundamental?"

No answer to these questions will be an answer which is not itself subject to the same question. For instance, for any event which we cite as the cause of some effect, we can ask what *its* cause was. It is insufficient to say that the answer to (Q5) – (Q8) is some necessary fact, since necessities also require an explanation. If we posit some brute, metaphysically necessary, fundamental fact, then this fact will still give rise to (Q8): why is it fundamental?

## 2.7 The Question of Actuality

All of these questions are reflections of the question of actuality: that is, given that the actual world is one of many ways the world could be, why is this world actual and not some other world? Suppose there is some further fact about this world which selects it as the actual way things are instead of just another way things could have been. Such a fact is part of the facts in the world – and so, it is part of the world about which we are asking, "why is it actual?"

(Q9) Why is the actual world this world?

To insist, as some do, that this world is not the only real world, and that there are perhaps other universes which are equally as real as our own, is a move which does nothing to answer the question. I am not having an experience of those other universes right now – it is the actual universe which I am experiencing. So, the question of why one of many realities should be actual has advanced little from the question of why one of many possibilities should be actual.

## 2.8 The Emergence Question

My point is not to suggest a particular answer to any of these questions here, since that would defeat my point. My point is instead that there are many questions of this sort, which remain in spite of the fact that they can never be properly formulated as questions without appearing to be trivial or self-contradictory – what is being asked with the question is perfectly clear, even though the form of the question itself is not clear.

Our question, "why should these basal conditions give rise to these emergent properties?", is a variation on these sorts of questions. We might compile an exhaustive list of all of the physical properties in the world by means of which emergent properties can be manipulated or brought out of existence – and the result will manipulatively necessitate the emergent properties. However, despite having an exhaustive list, we won't have an answer to why the emergent properties must have occurred. If we suppose there must exist some further fact which brings about the emergence of consciousness – as the traditional dualist does – then we are still left

with the question of why such and such a non-physical substance or other further fact should guarantee that an experience like mine occurs.

Again, the emergentist denies that emergent properties are wholly grounded in their basal conditions. Someone might ask, "what grounds the emergent properties, then?" The emergentist responds that they are *fundamental*. Yet it seems like fair game to ask, *in virtue of what* are they fundamental?

The difficulty applies whether one holds that emergent properties supervene with nomological, manipulative, or metaphysical necessity. If emergence is only nomologically necessary, then the question "why are there fundamental trans-ordinal laws?" ought to answer itself – the laws are *fundamental* – but it doesn't seem to. If emergence is metaphysically necessary, then the question "why does phenomenal emerge from the physical" *ought* to be as uninteresting as the question "why is 2 not identical to 3?", yet it isn't. If emergence is manipulatively necessary, the position I am inclined towards, then there is nothing besides the physical which explains why emergence happens, and yet the physical does not metaphysically guarantee that emergence should occur, in that it doesn't rule out the possibility of it not occurring. In this case, "why does the phenomenal emerge from the physical?" is a question for which no answer in the world exists, and yet the question suggests a clearly conceivable, genuine possibility.

So it seems reasonable to say that the following question belongs to the same class as the other questions discussed here:

(Q10) Why do unexplainable novelties emerge?

Since, if one could explain why emergence occurred, then the phenomena in question would neither be unexplainable, nor novel, nor emergent.

### 2.8 *Fortiora te ne scrutatus fueris.*

The attitude one takes towards these questions is to an extent a reflection of temperament. A certain reactionary temperament adopts an attitude which involves pounding one's fist very hard on a podium and insisting that the questions do in fact stop somewhere. A more radical temperament suggests that we dispose of our old concepts of identity, personal identity, reference, time, causation, teleology, supervenience, grounding, emergence, and actuality and replace them with cropped and culled concepts which stop short before leading to interminable questions. Alternatively, a more pragmatic temperament proposes that we have no choice but to work with concepts which apply quite well to the vast middle of reality, and are thus valuable for their utility, but become nonsensical at the edges or when the engine idles.

A more disciplined temperament holds that a question without an answer isn't a real question at all. It is a *confused question*. It is trivially true that it is presently now, and that I am myself. If emergence is ideally *a priori* scrutable, then for the ideal reasoner is it just as obvious why things should emerge as they do as it is that the actual world is the world. We have only an illusion of understanding a question, but nothing in fact has been asked which we could have comprehended. What we think we are trying to ask cannot be formulated in words in the first place – no answer could ever be properly represented – and the question is trivial nonsense.

There is another attitude which one can take towards these questions, including the question of why anything should have emerged at all. What appeared to be confused questions are in fact the most important questions. Yet, that the question cannot be properly represented is perhaps an indication that the answer cannot be properly represented either – how things are in the world will not answer that it exists at all.[472] One cannot answer the question of being with beings. For the temperament I have in mind, these things are to be accepted with natural piety.

[472] c.f. Wittgenstein (1922), 6.4 – 6.522

REFERENCES

Aristotle. *Posterior Analytics*.

Alberts, B., et al. (2002). *Molecular Biology of the Cell, 4th Edition.* New York: Garland Science.

Alexander, S. (1920). *Space, Time, and Deity: The Gifford Lectures at Glasgow. Vols. I & II.* London: Macmillan.

Alexander, S. (1922). Natural Piety. *The Hibbert Journal* 20, 609-621.

Almog, J. (2009). Dualistic Materialism. In R. Koons & G. Bealer (Eds.), *The Waning of Materialism*. Oxford University Press.

Amundson, R. and Lander, G. V. (1994). Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy* 9: 443-469.

Allen, C., (2009). Teleological Notions in Biology. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.

Anderson, P. W. (1972). More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press.

Anscombe, G. E. M. (1958). On Brute Facts. *Analysis*. 18(3) 69-72

Appiah, K. A. (2008). Experimental Philosophy *Proceedings and Addresses of the American Philosophical Association.* 82(2) 7-22

Archer, M. S. (1995). *Realist Social Theory: The Morphogenetic Approach*. Cambridge University Press.

Assad, A. and Packard, N. (1992). Emergent Colonization in Artificial Ecology. In F. Varela and P. Bourgine (Eds.), *Towards a Practice of autonomous systems: proceedings of the First European Conference on Artificial Life*, pp 143-152, MIT Press.

Austin, J. L. (1956). A Plea for Excuses. *Proceedings of the Aristotelian Society,* New Series, 57, 1-30

Ayala, F. J. (1999). Adaptation and novelty: teleological explanations in evolutionary biology. *History and Philosophy of the Life Sciences*. 21(1):3-33.

Ayer, A. J. (1936) *Language, Truth, and Logic*, London: Gollancz, 2nd Edition, 1946.

Barnes, E. (2012). Emergence and Fundamentality, *Mind*. 121 (484): 873-901.

Bauchau, V. (2006). Emergence and Reductionism: From the Game of Life to the Science of Life. In B. Feltz, M. Crommelinck and P. Goujon, (Eds.), *Self-organization And Emergence In Life Sciences*, 29-40. Springer.

Bedau, M. (1997). Weak Emergence, *Philosophical Perspectives*, 11: 375–99.

Bedau, M. (2008). Downward Causation and Autonomy in Weak Emergence. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press.

Bennett, K. (2003). Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It. *Noûs* 37:3, 471-497.

Berkeley, G. (1713). *Three Dialogues between Hylas and Philonous.*

Bernstein, S. (2010). *Essays On Overdetermination*. (Doctoral dissertation). University of Arizona, Tucson.

Bhaskar, R. (1982). Emergence, Explanation, and Emancipation. In Secord, (Ed.) *Explaining Human Behavior: Consciousness, Human Action and Social Structure*, 275-310. Sage.

Black, M. (1952). The Identity of Indiscernibles, *Mind*, 61:153-164.

Blackburn, S. (1985). Supervenience Revisited. In I. Hacking (Ed.), *Exercises in Analysis: Essays by Students of Casimir Lewy*. Cambridge University Press.

Blau, P. M. (1977). A Macrosociological Theory of Social Structure. *American Journal of Sociology* 83:26-54.

Blau, P. M. (1981). Introduction: Diverse Views of Social Structure and Their Common Denominator. In Blau and Merton (Eds.) *Continuity in Structural Inquiry*, 1-23. Sage.

Block, N. (1995). Concepts of Consciousness. In Chalmers, D. (Ed.) (2002). *Philosophy of Mind: Classical and Contemporary Readings*, 206-218.

Block, N. and Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review*, 108, 1-46.

Bouwsma, O. K. (1954). The Mystery of Time, *Journal of Philosophy* 51 (12):341-363

Braddon-Mitchell, D. (2007). Against Ontologically Emergent Consciousness. In B. P. McLaughlin & J. D. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell.

Brandon, R. N. (2006). Teleology in Self-Organizing Systems, in B. Feltz, M. Crommelinck and P. Goujon, (Eds.), *Self-organization And Emergence In Life Sciences*, 267-281. Springer.

Bratman, Michael (1992). Shared Cooperative Activity, *Philosophical Review*, 101:1, 327–342.

Brigandt, I. and Love, A. (2012). Reductionism in Biology. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.

Broad, C. D. (1925). *The Mind and its Place in Nature*, Routledge & Kegan Paul.

Broad, C. D. (1930). *Five types of ethical theory.* New York: Harcourt, Brace and Co.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies In Philosophy* 4, 73–121.

Burge, T. (2009). Modest Dualism. In R. Koons & G. Bealer (Eds.), *The Waning of Materialism: New Essays*. Oxford University Press.

Byrne, Alex. (1999). Cosmic Hermeneutics. *Noûs.* 33, 347-383

Carnap, R. (1928). *Der logische Aufbau der Welt.* Leipzig: Felix Meiner Verlag.

Carnap, R. (1947). *Meaning and Necessity*. Chicago: University of Chicago Press.

Caston, V. (1997). Epiphenomenalisms, Ancient and Modern. *The Philosophical Review*, 106(3), 309-363

Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press.

Chalmers, D. & Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110, 315-61.

Chalmers, D. (2006). Phenomenal Concepts and the Explanatory Gap. In T. Alter & S. Walter, (Eds.) *Phenomenal Concepts and Phenomenal Knowledge*. Oxford University Press.

Chalmers, David. (2006). Strong and Weak Emergence. In P. Clayton and P. Davies, (Eds.) (2006) *The Re-emergence of Emergence*, Oxford University Press.

Chalmers, David. (2010). *The Character of Consciousness*. Oxford University Press.

Chomsky, N. (1959). Review of 'Verbal Behavior'. *Language* 35:26-58.

Chomsky, N. (1965). *Aspects Of The Theory Of Syntax*. MIT Press.

Colyvan, M. (2009). Naturalising Normativity. In D. Braddon-Mitchell and R. Nola (Eds.), *Conceptual Analysis and Philosophical Naturalism*, MIT Press, 303–313.

Corkum, P. (2008). Aristotle on Ontological Dependence *Phronesis: A Journal for Ancient Philosophy*. 53(1): 65-92

Corkum, P. (2012). Aristotle and Ontological Priority. Invited Symposium, American Philosophical Association, Pacific Division Meeting, April 6, 2012.

Correia, F. (2008). Ontological Dependence *Philosophy Compass* 3/5: 1013–1032.

Chalmers, D. (2011). Revisability and conceptual change in 'two dogmas of empiricism' *Journal of Philosophy* 108 (8) (2011)

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes, *Journal of Philosophy* 78: 67-90.

Crutchfield, J. (2008). Is Anything Ever new? Considering Emergence. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press.

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, 72.20: 741-765.

Damiano, L. (2012).Co-emergences in life and science: a double proposal for biological emergentism. *Synthese*, 185:273-294.

Dancy, J. (2006). Nonnaturalism in Copp, D. (Ed.), *The Oxford Handbook of Ethical Theory*. 122-145. Oxford University Press.

Davidson, D. (1973). The Material Mind. In Patrick Suppes (Ed.), *Logic, Methodology and the Philosophy of Science*. North-Holland.

Davidson, D. (1973). Radical Interpretation *Dialectica* 27 (1):314-328

Davidson, D. (1970).. Mental Events, in D. Davidson, *Essays on Actions and Events*, Oxford: Oxford University Press, 207–223.

Davies, J. C. (1962). Toward a Theory of Revolution. *American Sociological Review*, 27(1).

Fine, K. (1994). Essence and Modality: The Second Philosophical Perspectives Lecture, *Philosophical Perspectives*, 8: 1-16.

Derix, R., Van Hooff, J., De Vries, H., and Wensing, J. (1993). Male and Female Mating Competition in Wolves: Female Suppression vs. Male Intervention. *Behaviour*, 127:1, 141-174.

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Dennett, D. (1991). Real Patterns. *The Journal of Philosophy*, 88(1), 27-51.

Dretske, F. (1995). *Naturalizing the Mind.* MIT Press.

Durkheim, E. (1964). *The Rules of Sociological Method*. New York: The Free Press. (Original work published 1895).

Durkheim, E. (1951). *Suicide*. New York: The Free Press. (Original work published 1897).

Durkheim, E. (1953). Individual and Collective Representations. *Sociology and Philosophy*, 1-34. New York: The Free Press. (Original work published 1898).

Elder-Vass, D. (2007). For Emergence: Refining Archer's Account of Social Structure. *Journal for the Theory of Social Behavior*. 37.1: 25-44.

Evans, G. (1982). *The Varieties of Reference.* Oxford University Press.

Fine, K. (1995). Ontological Dependence. *Proceedings of the Aristotelian Society*, New Series, 95, 269-290

Fine, K. (1995): Senses of Essence. In Walter Sinnott-Armstrong, editor: *Modality, Morality, and Belief. Essays in Honour of Ruth Barcan Marcus*. Cambridge: Cambridge University Press, 53-73.

Fine, K. (1994). Essence and Modality. *Philosophical Perspectives*, 8, 1-16.

Fine, K. (2001). The question of realism. *Philosophers' Imprint* 1 (2):1-30.

Fine, K. (2003). The Non-Identity of a Material Thing and its Matter, *Mind* 112, 195–234

Fine, K. (2009). The Question of Ontology. In D. Chalmers, D. Manley, & Wasserman (Eds.) *Metametaphysics: new essays on the foundations of ontology*.

Fine, K. (2010) Towards a Theory of Part. *Journal of Philosophy* 107 (11):559-589

Fodor, J., (1974). Special sciences, or the disunity of science as a working hypothesis, Synthese 28: 77-115. reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (Ed.) (2002) . 126-134.

Foot, P. (2003). *Natural Goodness*. Clarendon Press.

Francescotti, R. M. (2007). Emergence *Erkenntnis*, 67:47-63

French, Peter, (1984). *Collective and Corporate Responsibility*. Columbia University Press.

Gould, S.J. and Vrba, E.S. (1982). Exaptation – a missing term in the science of form. *Paleobiology* 8: 4-15.

Gould, S.J. & Lewontin, R.C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme *Proceedings of the Royal Society of London*, Series B, 205(1161), 581-598.

Granovetter, M. (1973).The Strength of Weak Ties. *American Journal of Sociology* 78:6, 1360-1380.

Hare, R. M. (1952). *The Language of Morals*. Clarendon Press.

Hasker, W. (1999). *The Emergent Self.* Cornell University Press.

Hawthorne, J. (2002). Blocking Definitions of Materialism, *Philosophical Studies*, 110: 103–13.

Hempel, C. & Oppenheim, P. (1948). Studies in the Logic of Explanation., *Philosophy of Science* 15: 135-175.

Henrich & Henrich (2007). *Why Humans Cooperate*. Oxford University Press.

Heudin, J. (2006). Artificial Life and the Sciences of Complexity: History and Future. In B. Feltz, M. Crommelinck and P. Goujon, (Eds.), *Self-organization And Emergence In Life Sciences*, 227-247. Springer.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 11, 587-612.

Hollich, et. al. (2000). Emergentist Thinking in Language. In P.B. Anderson (Ed.) *Downward Causation*. Aahus University Press.

Horgan, T. (1983). Supervenience and Cosmic Hermeneutics. *Southern Journal of Philosophy* 22, Supplement, 19-38.

Horgan, T. (1993). From Supervenience to Superdupervenience: Meeting the Demands of a Material World, *Mind* 102, 555-86.

Horgan, T. & Tienson, J. (2001). Deconstructing New Wave Materialism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its Discontents*. Cambridge University Press.

Horgan, T. (2001). Causal compatibilism and the exclusion problem. *Theoria*, 16(40), 95-116.

Horgan, T. (2002). Replies. *Grazer Philosophische Studien*: *Essays on the Philosophy of Terence Horgan* 63, 306-307.

Horgan, T. (2009). Materialism, Minimal Emergentism, and the Hard Problem of Consciousness. In R. Koons & G. Bealer (Eds.), *The Waning of Materialism*. Oxford University Press.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32. 127-136.

Jackson, F. (1994). Finding the Mind in the Natural World. In D. Chalmers (Ed.) (2002) *Philosophy of Mind: Classical and Contemporary Readings*. 162-169.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.

Kallestrup, J. (2006) The causal exclusion argument *Philosophical Studies* 131 (2):459-85.

Kant, I. (1781). *The Critique of Pure Reason*.

Kitcher, P. (1984). 1953 and All That: A Tale of Two Sciences. *Philosophical Review* 93: 335-373.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research* 52 (1):1-26.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies* 95: 3-36.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kim, J. (2010). *Essays in the Metaphysics of Mind*. Oxford University Press.

Kim, J. (2011). *Philosophy of Mind, 3rd Edition.* Westview Press.

Kim, J. (2011b). From Naturalism to Physicalism: Supervenience Redux. Romanell Lecture at the Central Division Meeting of the American Philosophical Association, Minneapolis, March 31, 2011.

Kim, J. (1998). The Many Problems of Mental Causation, excerpt from Mind in a Physical World, 29-47. reprinted in Philosophy of Mind: Classical and Contemporary Readings, D. Chalmers (Ed.) (2002) . 170-178.

Knobe, J. and Nichols, S. (2008). An Experimental Philosophy Manifesto. In Knobe & Nichols, Eds., *Experimental Philosophy.* New York: Oxford University Press, 3-16.

Kobes, B. (2009). Burge's Dualism, in R. Koons & G. Bealer(Eds.), *The Waning of Materialism: New Essays.* Oxford University Press.

Kripke, S. (1980). *Naming and Necessity.* Basil Blackwell, Oxford.

Laplace, P. S. (1902). *A Philosophical Essay on Probabilities, 6th Edition.* (F. W. Truscott & F. L. Emory, trans.) New York: John Wiley & Sons. (Original work published 1840)

Leuenberger, S. (2008). Ceteris Absentibus Physicalism, *Oxford Studies in Metaphysics*, Volume 4, D. Zimmerman (Ed.), Oxford University Press, 145–170.

Levine, J. (1993). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, 64(4), 354-361.

Lewes, G. H. (1874). *Problems of Life and Mind, Vol I.* Trubner & Co., London.

Lewis, D. (1973). Causation, *Journal of Philosophy*, 70: 556–67.

Lewis, D. (1974). Radical Interpretation. *Synthese* 27(3), 331-344.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Nous*, 13, 4, 455-476

Lewis, D. (1983). New Work for a Theory of Universals. *Australasian Journal of Philosophy*, 61:343-77.

Lewis, D. (1984). Putnam's Paradox. *Australasian Journal of Philosophy* 62 (3):221 – 236.

Lewis, D. (2000): Causation as Influence, *Journal of Philosophy*, 97: 182–97.

Loar, B. (1997). Phenomenal States, in Block, Flanagan & Guzeldere, Eds., *The Nature of Consciousness*, Cambridge, Mass.: MIT Press, 597-616.

Lowe, E. J. (2006). Non-Cartesian substance dualism and the problem of mental causation. *Erkenntnis*, 65(1), 5-23.

Lowe, E. J. (2005). Ontological Dependence. *Stanford Encyclopedia of Philosophy*.

Lowe, E. J. (1998). *The Possibility of Metaphysics*. Oxford: Clarendon Press.

Lycan, W.G. (1998). In Defense of the Representational Theory of Qualia: Replies to Neander, Rey and Tye. In Tomberlin (Ed.) *Language, Mind, and Ontology*, *Philosophical Perspectives*, 12. Ridgeview Publishing.

MacFarlane, J. (2004). In what sense (if any) is logic normative for thought? (draft April 21, 2004). [Retrieved: Feb 2., 2013]. Available from: <http://johnmacfarlane.net/normativity_of_logic.pdf>.

MacWhinney, B. (2001). Emergentist Approaches to Language, in Bybee, J. & Hopper, P. (Eds.) *Frequency and the emergence of linguistic structure.* Benjamins.

MacWhinney, B. (2006). Emergentism - Use Often and With Care. *Applied Linguistics* 27.4: 729–740.

Maslen, C., Horgan, T., and Daly, H. (2009). Mental Causation. Ch. 24 in *The Oxford Handbook of Causation*, Helen Beebee et. al. Eds. Oxford University Press.

Mayr, E. (1959). Typological versus Population Thinking, *Evolution and Anthropology: A Centennial Appraisal*, 409-412. Washington: Anthropological Society of Washington.

Mayr, E. (1996). The Autonomy of Biology: The Position of Biology Among the Sciences. *The Quarterly Review of Biology*, 71: 97-106.

Mayr, E. (1996b). What Is a Species, and What Is Not? *Philosophy of Science*, 63.2: 262-277.

Mayr, E. (1997). *This is Biology: The Science of the Living World.* Harvard University Press.

McLaughlin, B. (1992). The Rise and Fall of British Emergentism. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press.

McLaughlin, B. (2008). Emergence and Supervenience. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*. MIT Press.

Merricks, T. (2003). *Objects and Persons*. Oxford: Oxford University Press.

Millikan, R. (1989a). An ambiguity in the notion of function. *Biology and Philosophy* 4: 172-176.

Millikan, R. (1989b). In defense of proper functions. *Philosophy of Science* 56: 288-302.

Mill, J.S. (1974). A System of Logic Ratiocinative and Inductive. In John M. Robson. (Ed.) *The Collected Works of John Stuart Mill, Volume VII.* Toronto: University of Toronto Press. (Original work published 1843)

Moore, G. E. (1922). *Philosophical Studies*. Routledge & Kegan Paul.

Morgan, C. L. (1923). Emergents and Resultants, in *Emergent Evolution*. London: Williams & Norgate.

Morrison, M. (2006). Emergence, Reduction, and Theoretical Principles: Rethinking Fundamentalism *Philosophy of Science*, 73.

Nadelhoffer, T., & Nahmias, E. (2007). The Past and Future of Experimental Philosophy. *Philosophical Explorations* 10(2): 123-149.

Nagel, E., (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Hackett.

Nagel, T. (1979). Panpsychism. *Mortal Questions*. Cambridge: Cambridge University Press.

Nagel, T. (2012). *Mind and Cosmos*. Oxford University Press.

Neander, K. (1991). The teleological notion of 'function'. *Australasian Journal of Philosophy*, 69(4), 454-468

Neuber, M. Realism as a Problem of Language. in Creath, R. (2012). *Rudolf Carnap and the Legacy of Logical Empiricism.* Springer.

Ney, A. (2008). Physicalism as an Attitude. *Philosophical Studies*, 138.

O'Connor, T. & Wong, H. (2012) Emergent Properties. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.*

O'Grady, W. (2011). Emergentism, in Hogan, P. (Ed.) *Cambridge Encyclopedia of Language Sciences*, Cambridge University Press.

O'Grady, W. (2010). An Emergentist Approach to Syntax. in Narrog & Heine (Eds.) *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

Papineau, D. (2009). Naturalism. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.*

Putnam, H. (1962). It Ain't Necessarily So. *Journal of Philosophy*, 59: 658-671

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, *Seventieth Annual Meeting of the American Philosophical Association Eastern Division.* 70(19), 699-711

Putnam, H. (1975). The Meaning of 'Meaning', in *Language, Mind, and Reality*. Cambridge University Press.

Quine, W.V.O. (1960). *Word and Object*. MIT Press.

Quine, W.V.O. (1992). *Pursuit of Truth*. Revised Edition. Harvard University Press.

Quine, W.V.O. (1975). Two dogmas of empiricism (pp. 41-64). Springer Netherlands.

Robertson, Teresa, (2008). Essential vs. Accidental Properties. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.).

Ronald, E. M. A., Sipper, M., and Capcarrere, M. S. (2008). Design, Observation, Surprise! A Test of Emergence. in *Emergence: Contemporary Readings in Philosophy and Science*, MIT Press.

Rosenberg, A. (1995). *Philosophy of Social Science, Second Edition*. Westview Press.

Rothschild, L. (2006). The Role of Emergence in Biology. Ch. 6 in *The re-emergence of emergence: the emergentist hypothesis from science to religion.* Edited by Philip Clayton and Paul Davies. Oxford University Press.

Parfit, D. (1986). *Reasons and Persons*. Oxford University Press.

Polger, T. W. (2011). Are Sensations Still Brain Processes? *Philosophical Psychology* 24 (1):1-21.

Putnam, H. (1981). *A Problem About Reference*. Cambridge University Press.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193.

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy* 70 (19):699-711.

Salmon, W. (1971). 'Statistical Explanation', in *Statistical Explanation and Statistical Relevance*, W. Salmon, (Ed.), 29–87, University of Pittsburgh Press.

Salmon, W. (1994). 'Causality Without Counterfactuals.', *Philosophy of Science*, 61: 297–312.

Sawyer, R. K. (2001). Emergence in Sociology. *American Journal of Sociology*, 107:3, 551-585.

Sawyer, R. K. (2002). Durkheim's Dilemma: Towards a Sociology of Emergence. *Sociological Theory*. 20:2, 227-247.

Sawyer, R. K. (2005). *Social Emergence: Societies As Complex Systems.* Cambridge University Press.

Searle, J. Reductionism and the Irreducibility of Consciousness. In M. Bedau and P. Humphreys (Eds.) (2008). *Emergence: Contemporary Readings in Philosophy and Science*, 69-80. MIT Press.

Scerri, E. (2006). Reduction and Emergence in Chemistry—Two Recent Approaches. *Philosophy of Science.* 74:5, 920-931

Schaffer, J. (2008). On What Grounds What. In: David Chalmers, David Manley, Ryan Wasserman (Hrsg.): *Metametaphysics: New Essays on the Foundations of Ontology.* Oxford University Press, 357–383.

Schaffer, J. (2010). Monism: The Priority of the Whole *Philosophical Review*, January 2010 119(1): 31-76.

Schelling, T. (1978). *Micromotives and Macrobehavior*. W. W. Norton and Company.

Skiles, A. (2013). Getting Grounded. Colloquium Paper delivered at the Central Division meeting of the American Philosophical Association, 2013.

Sider, T. (2011). *Writing the Book of the World.* Oxford: Clarendon Press.

Smart, J. J. C. (1959). Sensations and Brain Processes. *The Philosophical Review*, 68:2, 141-156.

Smiley, M. (2011). Collective Responsibility. *The Stanford Encyclopedia of Philosophy* Edward N. Zalta, ed.

Soames, S. (2005). *Reference and Description: The Case against Two-Dimensionalism*. Princeton University Press.

Soames, S. (2011). Kripke on Epistemic and Metaphysical Possibility, in Saul Kripke, Edited by Alan Berger, Cambridge University Press.

Sober, E. (1980). Evolution, Population Thinking, and Essentialism. *Philosophy of Science* 47: 350-383.

Stalnaker, R. (1978). Assertion. Syntax and Semantics. 315-332

Stalnaker, R. (2001). Metaphysics without Conceptual Analysis. *Philosophy and Phenomenological Research*, 62, 631-636.

Stalnaker, R. (2004). Assertion Revisited: On the interpretation of two-dimensional modal semantics. Philosophical Studies.

Strawson, P.F. (1959). *Individuals: An Essay in Descriptive Metaphysics.* London: Methuen.

Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind & language*, 20(5), 469-494.

Stoljar, D. (2009). Physicalism. In E. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.

Stoljar, D. (2010). *Physicalism*. Taylor & Francis.

Stoljar, D. (forthcoming). Distinctions in Distinction, in Jesper Kallestrup and Jakob Hohwy (Eds.) *Being Reduced: New Essays on Causation and Explanation in the Special Sciences*, Oxford University Press.

Stuart-Fox, D., and Moussalli, A. (2009). Camouflage, communication and thermoregulation: lessons from colour changing organisms. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 364: 463-470.

Stich, S. (2000). Jackson's Empirical Assumptions. *Philosophy and Phenomenological Research*, 62(3) 637–643.

Thompson, M. (1995). The Representation of Life in *Virtues and Reasons*, Hursthouse, Lawrence, and Quinn (eds). Oxford: Clarendon Press.

Thornton, S. (2011). Karl Popper. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed.

Tye, M. (1995). *Ten Problems of Consciousness*. MIT Press.

Tye, M. (1999). Phenomenal consciousness: the explanatory gap as cognitive illusion. *Mind*. 108 (432): 705-725.

Van Cleve, J. (1990). Panpsychism Versus Emergence. *Philosophical Perspectives*, 4, 215-226

Van Fraassen, B. (2002). *The Empirical Stance*. Yale University Press.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford University Press.

Van Inwagen, P (1990). *Material Beings*. Ithaca, NY: Cornell University Press.

Varela, F. J. (1992). Autopoesis and a Biology of Intentionality in *Autopoesis and Perception*, McMullin, B. and Murphy, N. eds., Dublin City University.

Waters, C. K. (1990). Why the Antireductionist Consensus Won't Survive the Case of Classical Mendelian Genetics. *PSA 1990*, Philosophy of Science Association, 1: 125-139.

Weisbuch, G. (2006). The Complex Adaptive Systems Approach to Biology, in B. Feltz, M. Crommelinck and P. Goujon, (Eds.), *Self-organization And Emergence In Life Sciences*, 7-28. Springer.

Wilson, J. (1999). How superduper does a physicalist supervenience need to be? *Philosophical Quarterly* 50 (194):33-52

Wisdom, J. O. (1970). Situational Individualism and the Emergent Group-properties in *Explanation in the Behavioural Sciences*, Borger and Cioffi eds., 271-311. Cambridge University Press.

Wittgenstein, L. (1922). *Tractactus Logico-Philosophicus*. (1974). Trans. Pears, D. F. & McGuinness, B. F., London: Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical Investigations*, 3rd. Edition. (1958). Trans. by Anscombe, G. E. M., Macmillan:New York.

Woodward, J. (2010). Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation. *Biology & Philosophy*, 25:3, 287-318.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press.

Woodward, J. (2011). Scientific Explanation. In E. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*.

Wright, J. & Bregman, A. (1987). Auditory Stream Segregation and the control of dissonance in polyphonic music. *Contemporary Music Review*. 2:1.