INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

EXPLAINING PSYCHOLOGY: PSYCHOPHYSICAL REDUCTIONISM, EXPLANATION, AND THE UNITY OF SCIENCE

John Herbert Bolender

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
1996

ABSTRACT

EXPLAINING PSYCHOLOGY:
PSYCHOPHYSICAL REDUCTIONISM, EXPLANATION,
AND THE UNITY OF SCIENCE

JOHN HERBERT BOLENDER

Since functionalism implies that mental categories cross classify physical categories, it has classically been construed as precluding both the reduction of psychological theory to physical theory as well as the replacement of psychological by physical theory. However, many recent arguments for psychophysical reductionism and eliminative materialism also presuppose that mental categories cross classify physical categories. This raises questions as to the true significance of the cross classification of mental and physical categories.

I argue that viewing psychophysical reductionism or eliminative materialism as compatible with the cross classification of mental and physical categories presupposes a flawed view of explanation. More specifically, it disregards the fact that explanation essentially involves an audience. One cannot judge that a certain quantity of information does in fact explain a given explanandum without taking into consideration certain features of the relevant audience such as its interests and cognitive ability. I

present reasons for believing explanation to be cognitively constrained and interest-relative. Given such reasons, arguments for psychophysical reductionism and eliminative materialism lose their plausibility. I conclude that functionalist metaphysics does indeed preclude both eliminative materialism and psychophysical reductionism.

While most of my discussion concerns the physical irreducibility of psychological theory, I dispute Jerry Fodor's claim that this irreducibility implies that psychology must remain an autonomous science. I attempt to show that, while functionalist psychology does have a certain prima facie plausibility, it is only likely to be confirmed by being explained in terms of some other field of science. I do so by appealing to Michael Friedman's argument to the effect that the unification of the sciences plays a crucial role in the confirmation of theories. Since a physical reductive explanation of psychology has been ruled out, this raises the question as to what field one should look to in explaining psychological theory.

In the Appendix, I attempt to show the plausibility of seeking an explanation of functionalist psychology in terms of evolutionary biology. I show that while the cross classification of mental and physical categories foils a physical reduction of psychological theory, it does not rule out an adaptationist explanation of psychological features.

TABLE OF CONTENTS

Dedicated to Robert Woo

Introduction

In what relation will a completed theory of psychology stand to other sciences? More specifically, will the truth of a completed psychology be explained by appeal to some other field of science? If so, which science will that be? The dissertation is an attempt to work toward answering these questions. The arguments are not meant to be incontrovertible, but I hope to raise considerations which result in a greater appreciation of the plausibility of the conclusions.

In order to structure the discussion, a functionalist view of psychology is assumed. It is important to take some time to dwell on this assumption, for issues in the dissertation largely concern functionalism's implications for the relation of psychology to other sciences. Here, functionalism is taken to be the claim that, with sufficient time, scientists should be able to complete a true computational theory of behavior. It is also the claim that folk psychology, or some refinement of it, will be reductively explained in terms of this computational theory via the identification of psychological properties with the relevant computational properties.

It is important to consider the controversy surrounding this assumption. There are very many philosophers who consider functionalism to be untenable. Hence, the

dissertation's conclusion should be understood as conditional, viz. if functionalism is true then psychology cannot be physically reduced. It must remain conditional in form as long as the antecedent remains controversial. However, there is enough of a consensus in contemporary philosophy of mind in favor of functionalism to make the dissertation's conditional conclusion interesting. Furthermore, even many philosophers who contest functionalism often only do so vis-à-vis some limited aspect of psychology rather than for psychology as a whole. It is not uncommon, for example, to find a philosopher objecting to a functionalist construal of (e.g.) intentional properties while leaving open the possibility of a functionalist construal of sensory properties. Hence, even some philosophers who are sceptical of an unbridled functionalism may find the conclusion of the dissertation plausible vis-à-vis some limited aspect of psychology. For the sake of keeping the argument simple, however, I will assume an unqualified functionalism.

Functionalism makes both a metaphysical claim and an epistemic one. The former is the claim that psychological states are type identical to certain computational states. The latter is the claim that, in virtue of this type identity, a refined version of folk psychology will be reductively explained in terms of a computational theory.

Chapters One through Four concern functionalism's

metaphysical claim, while Five concerns its epistemic claim. Two important theses follow from functionalism's metaphysical claim: the multiple realizability of the mental and strong psychophysical supervenience. Given the metaphysical claim that psychological properties are computational, any given psychological property is realizable by a wide array of distinct physical properties, i.e., any psychological property is multiply realizable. Moreover, given functionalist metaphysics, any instantiation of a psychological property is nothing but an instantiation of some physical property with the requisite causal powers. Hence, functionalism implies that mental properties depend upon physical properties in a way known as "strong supervenience."

Classically, multiple realizability has been taken to foil psychophysical reductionism by disallowing the formation of the psychophysical bridge laws which this kind of reduction evidently requires. However, there have been recent attempts by some philosophers to argue that either strong psychophysical supervenience or multiple realizability actually implies some form or other of psychophysical reductionism. I attempt to show that these arguments are implausible.

In the first four chapters, the aim is to show that functionalist metaphysics does indeed preclude the reduction of psychology to physical science. The strategy is to

discredit recent attempts to argue from functionalist metaphysics to some form or other of psychophysical reductionism. By discrediting these arguments, I trust that plausibility is restored to the older view that functionalist metaphysics does indeed preclude the formation of the psychophysical bridge laws needed for such a reduction.

Chapter One addresses Jaegwon Kim's argument to the effect that strong psychophysical supervenience actually implies Nagelian psychophysical reductionism. Chapter Two concerns Kim's argument that the multiple realizability of the mental implies the truth of local psychophysical reductionism. Chapter Three is concerned with eliminative materialism. Eliminative materialists are usually not thought of as assuming any aspect of functionalist metaphysics in reaching their conclusion, but I argue that the most influential eliminativist arguments assume that mental properties are explanatorily empty in virtue of being multiply realizable.

I argue that all of these arguments depend crucially upon implausibly nonpragmatic views of explanation. Chapter Four is an elaboration upon the implausibility of the nonpragmatic views of explanation which are assumed in these arguments. I conclude that the attempts to infer some form of psychophysical reductionism from functionalist metaphysics are implausible, thus lending plausibility once

again to the original view that such metaphysics is physicalistically antireductionist.

Chapter Five, however, is not as exclusively concerned with the metaphysical aspects of functionalism. In that chapter, functionalism's epistemic component is shown to commit one to believing that computational psychology is reducible to some other field of science. Since a physicalist reduction has already been ruled out, in the Appendix I suggest a nonphysicalist reduction, viz. the reduction of psychology to evolutionary biology.

This is meant as an answer to the questions raised in the first paragraph, but I do not pretend to have presented conclusive arguments in favor of these views. I hope to have achieved the more modest aim of revealing some of their plausibility.

CHAPTER ONE

NAGELIAN PSYCHOPHYSICAL REDUCTIONISM

The aim of the present chapter is to defend the claim that functionalist metaphysics precludes Nagelian or global psychophysical reductionism. Given functionalist metaphysics, mental properties are second-order properties. Their being second-order properties implies both that they are multiply realizable by and strongly supervene upon physical properties. Their multiple realizability is often thought to preclude physicalist reductionism. Their strong supervenience on the physical, however, is thought by some to entail physicalist reductionism. In this chapter, I argue that the former view only is correct.

1.1. The Multiple Realizability of the Mental

According to functionalist metaphysics, pain[1] is a functional property, which is to say that a mental state is one of being in pain just in case it stands in the appropriate causal relations to sensory inputs, other mental states, and behavioral outputs. This view, in turn, implies

---

[1] I focus on the example of pain for the sake of simplicity. The points made about pain are meant to be perfectly general, applying to all mental properties including intentional ones.

that pain is a second-order property,[2] which is to say that it is the having of some property or other with a given causal specification. More specifically, the functionalist would say that being in pain is the property of having a (first-order) property whose instantiation is caused by .... and which causes .... .[3] The first series of dots are to be filled in by a description of the sorts of sensory inputs characteristic of pain, such as pin pricks or by other mental states of certain specified sorts. The latter series of dots are to be filled in by descriptions of the typical effects of pain, both behavioral and mental, such as groaning and the production of anxiety. The reference to inputs and behavioral outputs is characteristic of functionalism. Second-order properties, however, are quite common in the special sciences, and, hence, the causal specifications of the first-order properties which realize them need not include reference to inputs or outputs. Being fragile, for example, is the having of some property or other with the appropriate causal specification, even though that causal specification does not involve sensory inputs or

---

[2] For a discussion of functionalism and its relation to second-order properties, see Ned Block, "Can the Mind Change the World?", in Meaning and Method: Essays in Honor of Hilary Putnam, ed. George Boolos (Cambridge: Cambridge University Press, 1990), 137-70.

[3] In other words, a second-order property is the having of any (first-order) property figuring in some specified set of causal laws. Hence, a second-order property defines some specific set of causal properties which all the realizing properties must share.

behavioral outputs. Second-order properties are precisely the dispositional properties.

Second-order properties are generally multiply realizable in the sense that a wide array of distinct physical properties can possess the causal powers suitable for a given second-order property. (Moreover, we may assume that they can only be realized by physical properties or properties which are themselves ultimately realized by physical properties.[4]) Fragility, for example, is realizable by many different types of physical structures. Dormitivity, the property of having some property which causes sleep, is realized by many different chemical properties. Whether or not all second-order properties are multiply realizable is uncertain. But computational properties, which are indeed second-order, are clearly enormously multiply realizable given that computationally equivalent devices can be constructed out of radically different physical materials.

Given that pain is a computational property characterizable in terms of inputs, outputs, and state-to-state transitions, one could, at least in principle, build a

---

[4] This assumption is not actually entailed by the claim that pain is a second-order property, for it is at least imaginable (even if not nomically possible) that the causal specification characteristic of pain is satisfied by a nonphysical property which does not supervene on physical properties. However, the assumption is part of standard functionalist metaphysics in that functionalists usually grant the metaphysical priority of the physical.

computer capable of being in pain. The computer's being in pain just is the computer's running through the appropriate algorithm. However, computers of radically different construction and substance could run the same algorithm and hence equally well experience pain. Hence, there is strong prima facie reason to believe that pain, as a second-order property, is realizable by an enormous, perhaps infinite, number of physical properties, and no apparent reason to believe otherwise.[5] The enormous number of the various possible physical realizers of pain plays an important role in the later development of the argument of this chapter. It is by virtue of that enormousness that global or Nagelian psychophysical reductionism is false.

But what is it to say that a property realizes a property? To say that pain is realizable by a set of physical properties is to say that any exemplification of pain just is the exemplification of one or the other of those properties. The token instance of pain and the token instance of the physical property are identical. The instantiation of the physical property is all that is actually there.[6] This follows from the claim that pain is

---

[5] The reader must bear in mind that functionalist metaphysics is assumed throughout. One could, of course, deny the multiple realizability of pain simply by denying that it is a second-order property.

[6] Australian philosophers evidently use the term 'supervene' in the sense that I am here using the term 'realize'. For an example, note Keith Campbell's use of the term 'supervenience' in his Abstract Particulars (Oxford:

a second-order property. For, given that pain just is the having of some physical property with the right causal specification, any instance of pain just is an instance of such a physical property.

Despite the fact that functionalist metaphysics is assumed, it is worthwhile to make some remarks in defense of the multiple realizability of the mental. This should at least help render the antireductionist conclusion more plausible to those who are unwilling to accept the presupposition of functionalist metaphysics. Now, some might argue against the claim that pain is multiply realizable by claiming that there is little or no evidence for its actually being multiply realized. But this would be a non sequitur. For the claim that mental properties are multiply realizable is distinct from, and weaker than, the claim that mental properties are multiply realized. To say that a mental property is multiply realized is to say that it is realized by more than one property in the actual world. For example, pain is multiply realized if it is realized by one physical property in (e.g.) an actual lizard and some distinct physical property in an actual human. It follows that mental properties can be multiply realizable without being multiply realized. More specifically, a

---

Basil Blackwell, 1990). I adopt the use of the terms 'supervene' and 'realize' which is more common in the U.K. and the U.S. where they are assigned distinct meanings. For more on supervenience, see section 1.4.

mental property can be multiply realizable even though it may be realized by only one physical property in the actual world. Nonetheless, the property is multiply realizable if there are other physically possible worlds in which it is realized by different physical properties. Accordingly, the second-order nature of pain provides strong grounds for its being multiply realizable, but it provides less strong grounds for its being multiply realized. Hence, a failure to produce evidence for the claim that pain is multiply realized does not impugn the claim that it is multiply realizable. This is fortunate for the argument of this chapter, which only requires the weaker claim that pain is multiply realizable.

There is, however, some neuroscientific evidence for pain's being multiply realized in humans. For example, which areas of the brain are responsible for language vary in different people. Two people can be indistinguishable in terms of their linguistic performance (and hence fall under the same intentional generalizations and possess the same intentional properties) and yet process linguistic information in neurally distinct ways. The only means of discovering this difference is to examine the brain directly. Further, on the most credible neuroscientific theories, type identical semantic memories (e.g., one's understanding of the word 'chair') are realized by different synaptic weightings in different humans, and perhaps even by

different synaptic weightings by the same human at different points in its life.[7]

Multiple realizability is an important issue upon which much debate over psychophysical reductionism turns. It is not, however, the only such issue. Another concerns Donald Davidson's claim that psychological predicates are not nomic and hence that our commonsense view of the mind cannot be reduced to any sort of theory physical or otherwise.[8] Davidson's objection to reductionist ambitions is profound, urgent, and will not be addressed in this dissertation. (In Chapter Two, I will criticize Kim's claim that pain is not a nomic property, but Kim's reasons for making this claim are not Davidson's.) What will be addressed is the dispute over whether the multiple realizability of pain precludes reduction of psychology to the physical. I will defend the position that indeed it does and that any temptation to believe the contrary derives from overlooking essential pragmatic aspects of explanation.

1.2. Metaphysical Assumptions

In discussing the debate between psychophysical reductionists and functionalist antireductionists, it is

---

[7] See Owen Flanagan, Consciousness Reconsidered (Cambridge, MA: MIT Press, 1992), p. 20. For more on synaptic weightings, see Chapter Three, Section 3.3.

[8] Donald Davidson, "Mental Events," in Lawrence Foster and J. W. Swanson, eds., Experience and Theory (Amherst: University of Massachusetts Press, 1970), 79-101.

important to bear in mind that there is a substantial metaphysical view which both camps accept, viz. materialism.* Discussing materialism will not only serve to clarify the common agreement but will also elucidate the meanings of some important terminology, such as 'physical' and 'higher- and lower-levels'. Materialism is inspired by the great success of contemporary science in predicting large-scale phenomena on the basis of facts pertaining to small-scale constituent phenomena. This commonly held metaphysical view is often described as depicting a "layered view of the world." On this view, there is a lowest level characterized by entities (or processes) which are not decomposable into simpler entities. These most basic entities compose all other objects and processes. The basic entities themselves can be described using predicates appearing in the laws of the complete and true theory of basic physics. Those entities, predicates, and properties corresponding to the predicates are physical by definition. Mereological aggregates of the physical entities are also physical by definition, and Boolean operations on physical predicates yield predicates which refer to properties which are physical by definition. According to materialism, all particulars are physical. The view leaves open the question

---

* I understand materialism classically to be roughly equivalent to the claim that all particulars are physical, (and physicalism to be the claim that all properties are physical). Here, I am adding claims about supervenience to this classical view in defining 'materialism'.

as to whether all properties are physical.

Materialism allows for the fact that there are some predicates which, though satisfiable by compound individuals, cannot be satisfied by those of their parts which are too small and too simple. For example, the predicate 'is liquid' can be satisfied by some instances of water, but it cannot be satisfied by the constituent hydrogen and oxygen atoms themselves. Such a predicate is referred to as a "higher-level predicate." This highness must be understood as a matter of degree corresponding to the complexity of the object which is capable of satisfying it. For example, 'is liquid' is higher than 'is a molecule', since there are objects capable of satisfying the latter predicate but which are too small and simple to satisfy the former. However, 'is a molecule' is a higher predicate than 'is an electron' because there are some individuals which satisfy the last predicate but which are too small and simple to be molecules. The lowest-level predicates are physical predicates. Not until we have a fully developed basic physics can we say just what these predicates are.[10] Predicates at increasingly higher levels refer to properties which are only had by correspondingly more complex aggregates. Hence, one can properly speak of higher- and lower-level properties corresponding to higher-

---

[10] In a sense, the full set of physical predicates does not exist until basic physics is completed.

and lower-level predicates (in their atomic form).[11] Entities corresponding to any level of description are composed out of objects corresponding to lower levels of description, with the exception of the bottom-level entities which have no constituent parts.

Instantiations of any given higher-level property depend upon instantiations of specific lower-level properties. In the case of mental properties and their relation to lower-level properties, this is certified by the second-order nature of the former, as previously noted. In the case of higher-level properties in general, it may also be due to their having a second-order nature. For example, such special science properties as ductility and being a blue-eye gene are clearly second-order properties. Perhaps all properties above the microphysical level are second-order properties. But even if this is not the case, the plausibility of the dependence of the higher on the lower owes to the fact that science is so often successful in

---

[11] I do not want the definitions of 'physical', 'level', and so on to exclude the possibility of type materialism i.e., that a given higher-level property just is some lower-level property in virtue of being the referent of a lower-level predicate. Hence, I have deliberately left open the possibility that some higher-level properties, liquidity for example, can be referred to by disjunctive or conjunctive physical predicates. Accordingly, the metaphysical distinctions between property levels could collapse, but the descriptive distinctions involving predicates cannot. 'is liquid', for example, can only be a higher-level predicate, since it could never appear in basic physical theory. These points are important to my later discussion of why the possible truth of type materialism does not threaten antireductionism.

predicting a higher-level property instance by appealing to lower-level property instances. This dependence of the mental on the physical seems to imply, at a minimum, the "supervenience" of the mental upon the physical, a notion that will be explicated in Section 1.4.

The materialistic agreement among psychophysical reductionists and functionalist antireductionists may be summed up as follows: Mental properties are only had by individuals possessing physical properties, and mental property instances depend upon and co-occur with specific physical property instances. This implies that higher-level laws are true only in virtue of the truth of lower-level laws. This agreement, however, is purely metaphysical. It will be seen that the psychophysical reductionist believes that this metaphysical position is alone sufficient to show psychological science to be reducible to physical science, a claim which I shall argue to be false.

1.3. Nomic Reductionism

Psychophysical reductionists and their opponents, as discussed, agree on the truth of materialism, a purely metaphysical position. The disagreement, on the other hand, involves explanation, a notion which is not purely metaphysical but at least partly epistemic and so involves understanding. The disagreement concerns what might be called "nomic reductionism," viz. the view that laws couched

in terms of higher-level predicates can be explained by appealing to laws couched in terms of lower-level predicates. Psychophysical reductionism is simply nomic reductionism applied to psychological laws in particular instead of nonphysical laws in general.

Ernest Nagel gave the classic account of nomic reductionism.[12] A quote from Nagel will serve to emphasize a crucial point:[13]

> Whatever else may be said about reductions in science, it is safe to say that they are commonly taken to be explanations, and I will so regard them. In consequence, I will assume that, like scientific explanations in general, every reduction can be construed as a series of statements, one of which is the conclusion (or the statement which is being reduced), while the others are the premises or reducing statements.

Hence, Nagel claims that if one theory (what he calls the "secondary science") is reduced to another (the "primary science"), then the former is shown to be logically deducible from the latter. This condition of derivability must be satisfied in order for there to be a successful reduction.

I wish to emphasize that for Nagel this requirement is itself derivative from a more fundamental requirement, viz.

---

[12] See Ernest Nagel, The Structure of Science (Indianapolis: Hackett Publishing Company, 1979), Chapter 11; and "Issues in the Logic of Reductive Explanation", in Teleology Revisited (New York: Columbia University Press, 1979), 95-117.

[13] Teleology Revisited, p. 97.

that the primary science explain why the secondary science is true. Since a reduction must be an explanation, the derivability of the latter from the former is only of interest because the derivation is taken to be an explanation. If one could show that the primary science does not explain the secondary science, even if the latter is in principle derivable from the former, then the latter would not be reducible to the former. This is worth emphasizing because I will later argue that psychophysical reductionism is false even though the condition of derivability can in principle be met. It is false, I claim, because of the explanatory impotence of the relevant predicates appearing in the apparent primary science.

To repeat, the condition of derivability is met when the laws of the secondary science are deducible from primary science laws. However, if the secondary science contains concepts which are absent from the primary science, then one cannot derive the higher-level science from the lower-level science without the aid of some additional assumptions linking the conceptual apparatus of one theory to that of the other. Using Nagel's terminology, the condition of derivability cannot be met in such a case unless one satisfies the condition of connectability. Since higher-level predicates are not found in lower-level sciences, this latter condition is satisfied by showing there to be nomic relations (expressible in "bridge laws") between the

extensions of higher- and lower-level predicates. The most commonly discussed sort of bridge law quantifies over a biconditional open sentence. For example, '(x)(Ex <=> Tx)' could express such a bridge law enabling the reduction of thermodynamics to statistical mechanics, given that 'E' refers to the property of having a certain mean molecular kinetic energy and 'T' refers to having the relevant temperature. A sufficient quantity of such biconditional bridge laws, if available, could be used to derive thermodynamic laws from mechanical laws. According to Nagel, bridge laws can also be one-way entailments. For example, the secondary science law $(x)(M_1x \Rightarrow M_2x)$ can be derived from the primary science law $(x)(P_1x \Rightarrow P_2x)$ with the supplementary bridge laws $(x)(M_1x \Rightarrow P_1x)$ and $(x)(P_2x \Rightarrow M_2x)$.

Multiple realizability provides at least a prima facie reason for rejecting psychophysical reductionism. Since psychological predicates do not appear in physical or biological theory, any reductionist hope must rest upon the possibility of formulating psychophysical or psychobiological bridge laws. However, given that pain has more than one basal condition,[14] the necessary nomic

---

[14] 'Basal condition' is an old term from the emergentist literature which I find convenient. In this dissertation, it is used to refer to any physical property which realizes pain in some physically possible world. Hence, the claim that pain is multiply realizable is tantamount to the claim that it has more than one basal condition.

coextensions between mental and physical properties are apparently lacking.

Robert Richardson has argued that multiple realizability poses no threat to psychophysical reductionism, but I believe I can show Richardson's argument to be unsound. Richardson claims that the antireductionist argument falsely assumes that psychophysical reduction requires nomic coextensions between any mental property and some physical property. However, Richardson claims that Nagel's allowing for one-way entailments undermines the antireductionist argument.[15] According to Richardson,

> The [bridge laws] demanded by the condition of connectability need not be biconditional. Derivability... [] is adequately provided for, in turn, if only we find sufficient conditions at a lower level of organization capable of accounting for phenomena initially dealt with at a higher level; and this too requires no more than a mapping <u>from</u> lower <u>to</u> higher level types and <u>not</u> a mapping from higher to lower level types.

Richardson is correct in noting that Nagel allowed for one-way bridge laws, but he is mistaken in claiming that <u>all</u> such laws in a successful reduction can be mappings from lower- to higher-level properties. Some mapping from higher- to lower-level properties is essential. For consider any psychological law of the form '$(x)(M_1x \Rightarrow M_2x)$'. Does a physical law of the form '$(x)(P_1x \Rightarrow P_2x)$'

---

[15] Robert C. Richardson, "Functionalism and Reductionism", <u>Philosophy of Science</u> 46 (1979), 533-58. See p. 548. Emphases in the original.

entail this psychological law? Richardson believes that physical-to-mental bridge laws are sufficient to enable this entailment. But this isn't so, for given the physical law and the one-way entailments $(x)(P_1x \Rightarrow M_1x)$ and $(x)(P_2x \Rightarrow M_2x)$, the psychological law does not deductively follow.[16] (Devising modal reformulations of the laws does not alter this point.) In order to derive the psychological law from the physical law, one needs at least one mental-to-physical entailment.[17] For example, given the bridge laws $(x)(P_2x \Rightarrow M_2x)$ and $(x)(M_1x \Rightarrow P_1x)$, the psychological law is derivable from the physical law. Therefore, since multiple realizability appears to preclude mental-to-physical entailments, there is still an apparent obstacle for the psychophysical reductionist to overcome.

1.4. The Disjunction Strategy

Kim has suggested a more interesting reductionist response to multiple realizability, one which attempts to secure the existence of the necessary mental-to-physical entailments. I will present Kim's position and then attempt to show that it violates a plausible pragmatic constraint on explanation. Kim points out that the second-order nature of the mental implies that mental properties are strongly supervenient on

---

[16] Please imagine that each of these one-way entailments is vacuously true.

[17] I wish to thank Isaac Levi for clarifying some points on this matter in conversation.

physical properties. He then proceeds to propose an ingenious argument to the effect that strong psychophysical supervenience implies psychophysical reductionism. The present chapter is an attempt to show that the disjunction strategy does not work.

Strong supervenience can be defined as follows:

M-properties strongly supervene on P-properties just in case, necessarily, for each x and each M-property M^*, if x has M^*, then there is a P-property P^* such that x has P^*, and necessarily if any y has P^*, it has M^*.

Hence, the strong supervenience of M-properties on P-properties implies that, for any M-property, say M^*, there is a set of P-properties $(P_1, P_2, P_3, \ldots P_n)$ such that $M^* <=> (P_1 \text{ v } P_2 \text{ v } P_3 \text{ v } \ldots P_n)$ is necessarily true. (The modality here must be taken to refer to all possible worlds which have the same physical laws as the actual world.)

Take mental properties to be the M-properties and physical properties to be the P-properties. To say that the mental strongly supervenes on the physical just is to say that for any mental property there is a set of physical properties with the following characteristics: an individual's having any property in that set is sufficient for its having the mental property, and its having the mental property is sufficient for its having some property

or other belonging to that set. (Note that this conception of supervenience does not rule out type materialism, for any set of properties supervenes upon itself. I mention this point in anticipation of my later claim that even psychophysical antireductionism does not rule out type materialism.)

What makes this kind of supervenience strong is expressed in the definition's reference to physical necessity and sufficiency. That M-properties strongly supervene on P-properties implies modal links between M-property instantiations and P-property instantiations. It is customary in the literature to recognize weak supervenience as well which only implies a correspondence between M-property instantiations and P-property instantiations in the actual world.[18] If M-properties weakly supervene on P-properties without strongly supervening on them, then it is just a coincidence, an accidental generalization, that no two individuals are P-identical without also being M-identical. (Merely weak supervenience is of no great concern in this dissertation.)

A commitment to metaphysical functionalism brings with it a commitment to strong psychophysical supervenience. For, as noted in the preceding section, the second-order

---

[18] See Kim "Supervenience as a Philosophical Concept", Metaphilosophy 21 (1990), 1-27; reprinted in Kim Supervenience and Mind (Cambridge: Cambridge University Press, 1993), 131-60.

nature of pain implies that it is realizable by a set of physical properties with the right causal credentials.[19] Further, to say that pain is realized by a set of physical properties is to say that an individual cannot have the property of being in pain unless it has one of those physical properties. Hence, the second-order nature of pain implies its supervenience upon the physical. This supervenience, moreover, must be strong. For it is not simply an accidental generalization that the having of pain is coextensive with the having of some physical property or other with the appropriate causal credentials. It is, rather, of the very nature of pain that this coextension holds. Moreover, as Kim notes, weak supervenience alone does not do justice to the dependence of the mental on the physical. That dependence must at least imply that there are <u>necessary</u> links between mental and physical property instantiations, not simply accidental ones. Hence, the coextension must hold across all physically possible worlds.

The reader will have noted that I have already spoken of the physically necessary <u>coextensivity</u> of pain with the having of some property or other of a given type. That is to say, pain is coextensive with the disjunction of physical properties having the appropriate causal powers. Kim believes that <u>this</u> coextensivity is all that one needs for

---

[19] See footnote 3.

psychophysical reduction.[20] According to Kim, we have mental-to-physical entailments after all, but the physical property is disjunctive in nature. This is Kim's "disjunction strategy" for saving psychophysical reductionism from the apparent threat posed by multiple realizability.

As Kim points out, one cannot dismiss the disjunction strategy by claiming that there is something unsatisfactory, unnatural or nonnomic about all disjunctive properties. For a disjunctive property, as the notion is here understood, is simply a property which can be expressed using a disjunctive predicate. Hence, the property of being helium is disjunctive in that the predicate 'is helium in the sun or is helium elsewhere' expresses it. Hence, the problem with the disjunction strategy, and I do find it problematic, does not lie simply in the disjunctivity of the pertinent physical properties. Nor is it apparent that a disjunctive predicate cannot appear in a reductive bridge law. The opponent of psychophysical reductionism must show that there is something unacceptable about these properties or their corresponding predicates aside from mere disjunctivity.

1.5. Some Critics of the Disjunction Strategy

A typical antireductionist response is that the sort of

---

[20] Kim, "Concepts of Supervenience", Philosophy and Phenomenological Research 45 (1984) 153-76; reprinted in Supervenience and Mind, 53-78.

disjunctive property which is coextensive with pain fails to be nomic. Given that our model of reduction is Nagelian, it is essential that the reducing properties be nomic, for psychological laws must be reduced to physical or biological laws. Jerry Fodor and Ausonio Marras have made attempts to show that disjoining the basal conditions of pain fails to yield a nomic property. I agree with and will defend their conclusion, but in the present section I question their arguments.

Fodor anticipated a reductionist move similar to the disjunction strategy even before Kim publicly espoused it.[21] Fodor assumes that the reductionist, in order to play the disjunction card, must assume that nomicity is truth-functional. That is to say, the reductionist must assume that, generally, if (x)(Fx => Gx) and (x)(Hx => Ix) are both laws, then (x)[(Fx v Hx) => (Gx v Ix)] must also be a law. Fodor then proceeds to warn that if we allow nomicity to be truth-functional we will no longer have a criterion for distinguishing natural kinds from merely conventional kinds. For example, if being gold and being a tiger are natural kinds, then supposedly the property of being either gold or a tiger would also have to be a natural kind.

---

[21] Jerry A. Fodor, "Special Sciences, or The Disunity of Science as a Working Hypothesis", Synthese 28 (1974) 97-115; reprinted in Block Readings in the Philosophy of Psychology, vol. 1, 120-33.

Fodor, however, is attacking a straw man. Clearly, the disjunctive reductionist is not committed to the claim that nomicity is truth-functional. Rather, he is committed to the more modest claim that the disjunctive property coextensive with pain is nomic. To say that the disjunctive physical properties coextensive with mental properties are nomic is not to say that all disjunctions of nomic properties are nomic. For in order to say that some disjunctive properties are nomic, one does not need to say that all of them are. This becomes even more apparent when one considers the fact, noted above, that all properties are disjunctive. Therefore, it is patent that one can accept some disjunctive properties as natural kinds without forfeiting the distinction between natural and conventional kinds.[22]

Marras also has an objection to the claim that supervenience alone implies physical reductionism.[23] Let 'N' refer to the disjunctive physical property necessarily coextensive with pain, i.e., the property referred to by the predicate formed by disjoining the physical predicates referring to the basal conditions of pain. Given that pain is a second-order property, it is likely that a physical

---

[22] One should admit a disjunctive property as a natural kind if it guarantees sufficient similarity in causal powers for the individuals satisfying it. This is discussed in the following chapter.

[23] Ausonio Marras, "Psychophysical Supervenience and Nonreductive Materialism", Synthese 95 (1993), 275-304.

predicate expressing N would be enormously or even infinitely disjunctive. This is obviously the case for other clear examples of second-order properties, e.g., being fragile, being a market, being toxic to humans. Given that reduction is a relation between theories, this raises some serious doubts about the possibility of formulating the primary science laws which are meant to reduce pain-laws. Assuming, as may indeed be the case, that the physical predicate referring to N is infinitely disjunctive, then it is highly dubious that it could appear in a theory or anywhere else. However, let us assume, for the sake of granting Kim as much ground as possible, that the predicate would not be infinitely disjunctive. Nonetheless, it would be enormously disjunctive. Marras claims that the generalizations incorporating such finite but enormous predicates would fail to exhibit "representational economy." According to Marras, "[a]n infinite mind that could 'read off' the [enormous] disjunctions would enjoy no deeper <u>understanding</u> of the nature of the psychophysical coextensions, or of the physical basis for mentality....."[24]

As I interpret Marras, the problem with such enormously disjunctive predicates and the generalizations which feature them is that they lack the kind of unifying simplicity which a theory must possess in order to be explanatory. Hence,

---

[24] Ibid.

even an omniscient mind would gain no insight from the physical "theory" which reduces psychological theory. Marras' criticism, however, is unfair. Indeed, the reducing theory could possess a great deal of unifying simplicity despite its incorporating such tremendously disjunctive predicates. One must bear in mind that the generalizations in a theory are not necessarily all axiomatic. The long, complex generalizations which Marras finds objectionable can themselves be derived from other generalizations in the theory which are simpler in form. For example, assuming that the complex generalization (x)[(Fx v Hx) => (Gx v Ix)] appears in a theory, there is no reason to exclude the possibility that the laws (x)(Fx => Gx) and (x)(Hx => Ix) also appear in that theory. That is to say, the complex generalization in question could be a truth-function of less complex laws within the same theory. One cannot rule out a priori that the highly disjunctive generalizations contained in the physical theory in question are not derivable from nondisjunctive laws contained in the same theory. Therefore, an infinite mind could indeed gain understanding from this physical theory. The mind would simply have to be able to grasp the logical connections between the complex disjunctive generalizations and the simpler axiomatic ones. Being an infinite mind, it would have no difficulty doing so.

1.6. A Cognitive Constraint on Explanation

I do not criticize the arguments of Fodor and Marras as a means of defending psychophysical reductionism. On the contrary, I agree with Fodor in claiming that the disjunctive predicates in question are nonnomic, and I agree with Marras in that they are nonnomic in virtue of being nonexplanatory. My position is superficially similar to that of Marras, and so it is worthwhile to spend some time distinguishing the two.

Marras evidently believes that since the reducing physical laws would be unwieldy, highly disjunctive rewrites of psychological laws, they would fail to be illuminating, even to an infinite mind. I reject this claim, since an infinite mind would be able to continue the reduction, resolving the unwieldy physical generalizations into the simpler basic laws. At the very least, Marras has provided no grounds for rejecting this possibility. Hence, it remains open that such unwieldy generalizations could play a genuine role in the explanation of psychological laws relative to the cognitive capacities of an infinite mind.

I claim that the disjunctive predicates in question are nonexplanatory but for a different reason. In explaining my reason, I shall make remarks on the nature of explanation which may appear somewhat cursory and undefended. However, my views on explanation will be presented in greater detail and supported by argument in Chapter Four. For now, I trust

that it is sufficient to present some prima facie plausible theses on explanation and to show that they preclude psychophysical reductionism.

One such thesis is that whether or not a certain quantity of information explains an explanandum is partly a contextual matter. The interests and cognitive powers of those seeking an explanation place contextual constraints on whether a candidate explanans actually does explain a given explanandum. Any such context-sensitive view of explanation is often referred to as a "pragmatic" view of explanation. The explanandum of particular interest here is the complete and true theory of psychology, whatever that may turn out to be. The candidate explanans would be the physical "theory"[25] to which it is supposedly reducible along with the necessary psychophysical bridge generalizations. Given the pragmatic view of explanation here presupposed, the fact that the physical "theory" in conjunction with psychophysical bridge generalizations explains the psychological theory for an infinite mind does not imply that it explains the psychological theory simpliciter. The possibility is left open that, relative to any human explanatory context, the physical "theory" would fail to explain the psychological theory.

---

[25] Since a theory is necessarily explanatory and since I argue that the collection of physical generalizations in question is not explanatory, I place the word 'theory' in quotes.

Let us assume that this is the case, that, relative to humans, there is no physical theory which reduces psychological theory. My reasons for believing this to be the case will emerge shortly. One might insist, nonetheless, that the reducibility of psychology relative to (e.g.) an infinite mind is all that reductionism requires. Psychophysical reductionism is true, so one might claim, provided that an infinite mind is able to discern a physical explanation for each psychological law. One might insist that it is anthropocentric to say that reductionism is false simply because there is no physical explanation of psychological laws relative to human contexts. I reply by saying that if this is what one means by 'psychophysical reductionism', then so be it; psychophysical reductionism is true. But this is not the sort of reductionism which reductionists have been trying to defend. Most reductionists take their doctrine to have consequences for scientific methodology. These matters will be further addressed in Chapter Four, but for now it is enough to note that one of these consequences is that any psychological theory should be developed so that it is possible to rewrite its laws in physical terminology as physical laws. Hence, the sort of reduction anticipated by psychophysical reductionists must enable scientists to put the corresponding physical generalizations <u>to the same explanatory uses</u> which psychological laws serve. Therefore,

psychophysical reductionists believe that physicalist reduction guarantees to scientists the ability to use the corresponding physical generalizations in explanations. However, the physical irreducibility of psychology relative to humans precludes this sort of reduction.

All of this is, of course, contingent upon the claim that psychology is indeed physically irreducible for humans. In defending this claim, I propose another thesis on the nature of explanation, viz. that a candidate explanans E explains explanandum e relative to audience A only if A comes to understand e in virtue of grasping the semantic content of E. This thesis is intended to rule out such cases as the following: A comes to understand e by virtue of hearing E but only because the sound of the explainer's voice (and not the content of E) resulted in A's producing another explanans for e. The point is that it is a necessary condition for E explaining e relative to A that A understand e by reason of grasping E. (This constraint on explanatory efficacy will be further elaborated in Chapter Four.)

I claim that the candidate physical "theory" to which psychology is supposedly reducible is not explanatory relative to any context in which the audience is composed of human beings. Since it is not explanatory, it not only fails to be a theory but cannot reduce anything since reductions are necessarily explanations. As noted

previously, a psychophysical bridge generalization would link a functionally characterized psychological predicate to an enormously disjunctive physical predicate. The latter, far from being functionally characterized, would be tantamount to a list of all the basal conditions of pain. This list, we have every reason to believe, would be enormous if not infinite. The important point for this discussion is that no human could grasp the meaning of the physical predicate. To grasp the semantic content of that predicate would require abilities exceeding human limits on memory, if nothing else.

Consider an analogous case: The predicate 'is poison' is second-order in that the substances satisfying it are collectively characterized in terms of the effects which they produce when ingested by certain organisms. What if one were to attempt to formulate a bridge generalization linking 'is poison' to the disjunction of all of its physical (i.e., first-order) realizing properties? Since there is a virtually endless number of distinct toxic substances, one would produce a first-order predicate amounting to a mind-boggling list of distinct properties. No human being would be able to grasp the entire semantic content of such a first-order predicate, and so the predicate could have no explanatory power relative to humans. The same obstacle would face anyone attempting to devise a bridge generalization linking the second-order

predicate 'is in pain' with a first-order predicate. Since the first-order predicate would lack explanatory power, it could not play any role in a reduction. Therefore, given the second thesis concerning explanation, a generalization incorporating such a predicate could not explain anything to a human audience.

So as to avert a possible misunderstanding, I emphasize that it is not the complexity of the predicate's inscription which foils reduction but the complexity of its meaning. One could, of course, use a short physical predicate, an abbreviation, to represent the physical property and thus produce a succinct bridge generalization. But even if one were to substitute for the enormous predicate a short one, that short predicate would have to play the same role in physical theory. It would, in other words, have to be synonymous with the long predicate. Their shared meaning would consist in the same unwieldy enumeration of microphysical properties, and so the short predicate would be equally incomprehensible to humans.

I trust that it is evident that my antireductionism is not a metaphysical thesis. More specifically, it is not the denial of type materialism. Type materialism could be true, and this would not impugn the antireductionism here defended. If type materialism is true, then, given multiple realizability, pain is identical to the disjunction of its basal conditions. I am willing to allow that this might

actually be the case. Hence, in denying the existence of a bridge law linking pain with this disjunctive property, I do not deny the possibility of their identity. All I deny is that the nonfunctional characterization of the physical property could be humanly comprehensible (or, for that matter, formulatable).

This is important to note, since the question of whether psychophysical bridge laws can be formulated is often conflated with the question of whether type materialism is true. However, it is not the metaphysical identity or nonidentity of pain with the disjunctive physical property that foils reductionism. Rather, it is the complexity of the nondispositional (i.e., first-order) characterization of the physical property which does so.[26] If pain and the disjunction of its basal conditions really are one and the same property, then an attempted psychophysical bridge law would be relevantly analogous to the following claim:

C: (x)(x is iron <=> x is identical to terrestrial Fe atoms or lunar Fe atoms or Martian Fe atoms or ....),

where the dots continue the disjunction of all nomically possible locations of iron. The division of the universe

---

[26] I.e., it is the sense of the physical predicate which renders it incomprehensible and hence nonexplanatory, not its reference.

into places, one may suppose, is so fine-grained that the actual list represented by the dots would be too lengthy to be humanly grasped. (Note that the right-hand predicate is not synonymous with 'is identical to Fe atoms anywhere in the universe' but is simply a long list having the same extension as that simpler predicate.) What prevents C from being explanatory (and hence from being nomic) is not the nonidentity of the properties on either side of the biconditional sign, for they are indeed the same property. What renders it nonexplanatory is the complexity of the meaning expressed by the right-hand predicate. For an explanation must provide understanding in virtue of a grasp of its meaning, and the meaning of the right-hand predicate cannot be humanly grasped. Analogously, even if a bridge generalization expressing the necessary coextensivity of 'pain' and a physical predicate were to be formulated (by an archangel, say), it would lack explanatory power and hence would not be a bridge <u>law</u>. The statement would not be able to produce understanding in virtue of its content, which any law, being explanatory, must be able to do.

Now, one might attempt to defend Kim by claiming that I have misconstrued his position. I have been speaking of bridge laws as though they are linguistic entities doing explanatory work within theories. Accordingly, I have viewed the disjunction strategy as applying to predicates,

not properties. Kim is, in fact, careful to state that his disjunction strategy is only meant to apply to properties, and hence that the bridge laws with which he is concerned must be viewed as relations between properties, not as linguistic entities. Indeed, in discussing his disjunction strategy, he states that he is concerned with laws only insofar as they are construed as "nonlinguistic, nonconceptual, objective connections between properties."[27] One can plausibly point out that, while my cognitive constraint might apply to laws when construed as linguistic structures in theories, it does not apply to laws when construed as objective features of the universe.

However, Kim cannot be allowed this move, for it conflicts with the very model of reduction which he consistently presupposes in discussing the disjunction strategy, viz. the Nagelian model.[28] On the Nagelian model, as is evidenced in the earlier quote from Nagel, reduction is a derivational relation between statements. Hence, in speaking of bridge laws, one is indeed speaking of predicates. Furthermore, and perhaps more importantly, the kind of reduction with which I am here concerned is the sort of reduction which has the most direct relevance to cognitive science. I am concerned with understanding whether or not the disjunction strategy works for

---

[27] Supervenience and Mind p. 73.

[28] Ibid., p. 150. See also p. 317.

intertheoretic reduction in the sciences. I am less concerned in this paper with whether or not psychophysical reductionism is true in the sense of there being some strong metaphysical link between mental and physical properties - except insofar as this has some bearing on scientific practice. I believe that I have raised considerations sufficient to show that psychological science is not physically explainable relative to a human audience, even though it might be physically explainable in principle.

1.7. Conjunctive Predicates

The candidate lower-level generalizations not only violate the cognitive constraint on explanation in virtue of their disjunctivity but also in virtue of their conjunctivity. Given strong supervenience, the having of any one of the disjunct properties implies the having of pain. That being so, any disjunct property must itself be a complex conjunction of physical properties. If one takes 'physical' to mean microphysical, as per my stipulations in Section 1.2, then the point is obvious. The having of any atomic microphysical property, such as being a quark, is patently not sufficient for being in pain. A physical property which is sufficient for being in pain would have to be a complex conjunction of microphysical properties including microphysical relational properties.

By way of analogy, consider the property of being

money. It would be a mistake to believe that the property of being a small silver disc with George Washington's profile impressed on it is in the supervenience base of the property of being money. This is because the having of a subvenient property must be sufficient for the having of the supervenient property, and it is clear that the property of being a small silver disc with George Washington's profile impressed on it is not sufficient for the property of being money. The subvenient physical property must be highly relational. Specifying that property in physical terms must involve reference to the complex physical environment in which the small silver disc is located. Moreover, it would most likely have to refer to quite a large portion of that physical environment, enough of that environment to correspond to what economists call a "market."[29]

It has been argued that semantic mental properties, such as the property of having the belief that water is wet, are at least partially externalistically constituted.[30] If

---

[29] Paul Teller uses a similar example to illustrate the complexity of pain's subvenient physical properties. Teller's point, however, is that overly complex properties are not natural kinds. Since it is not obvious that a property per se can be complex or simple, I am suspicious of Teller's point. See Teller, "Comments on Kim's Paper", in Terence Horgan, ed., The Spindel Conference 1983: Supervenience, The Southern Journal of Philosophy 22 (1984), Supplement, 57-61.

[30] Hilary Putnam, "The Meaning of 'Meaning'," in Putnam, Mind, Language, and Reality, Philosophical Papers, Vol. 2, (Cambridge: Cambridge University Press, 1975), 215-71.

that is indeed the case, then any physiological property in the supervenience base of a semantic mental property would have to be highly relational, not a purely physiological property per se but a physiologico-environmental property. My point is that there are many reasons for believing that each basal condition of a mental property would have an enormously complex expression in physical theory. Hence, even a single instance of pain cannot be given a physical explanation.[31] And, in order to explain a psychological law, one must add the element of disjunctivity, which just makes matters worse for the physicalist reductionist.

1.8. Conclusion

Given the metaphysics of functionalism, mental properties are second-order properties. As such, they are both multiply realizable and strongly supervenient upon physical properties. Since this supervenience entails a physically necessary coextensivity between any mental property and the disjunction of its subvenient physical properties, Kim has

---

[31] Since pain is not an intentional property, and hence presumably not externalistically constituted, it is conceivable that one could specify some conditions in its physiological supervenience base using, for example, exclusively biological terminology. But psychophysical reductionism, as an interpretation of the unity of science doctrine, requires that pain-theory be reducible to a physical (i.e., microphysical) theory. Any single condition in pain's physical supervenience base would have to have an enormously complex expression couched in terms of the predicates proper to basic physics and hence would violate the cognitive constraint on explanation.

argued that bridge laws can be formed linking the two and enabling the reduction of psychological science to physical science. If this is meant as a purely metaphysical thesis, a statement of property identity, then it might possibly be true. I have submitted no considerations militating against the identity of pain and its coextensive disjunctive physical property. (Moreover, if properties are individuated on the basis of their causal powers, then pain and the disjunctive physical property should be identified. For, unless one believes in downward causation, the causal powers of pain are derived precisely from its physical realizing properties.[32])

What foils psychophysical reductionism is not the truth or falsity of type materialism but the nondispositional characterization of the physical predicate which would have to appear in any genuinely reductive psychophysical bridge law. That predicate would either be a brute enumeration of all of pain's realizing physical properties or, if not, it would be synonymous with a predicate which is just such an enumeration. As such, its meaning would be too complex to be humanly comprehensible, for the basal conditions of pain must be enormous if not infinite. It is because of the human incomprehensibility of such an attempted bridge generalization that psychology cannot be reduced to physical

---

[32] This point is discussed in the following chapter under the theme of the "Causal Inheritance Principle."

theory. For bridge generalizations and the underlying physical theory are construed in terms of laws, and laws must be comprehensible in virtue of their essentially possessing explanatory power. This is especially salient when one considers the fact that reductions are, by definition, explanations. The bridge statements and the relevant generalizations of the physical theory, in order to be reductive, must serve to explain psychological laws. If they are incomprehensible, moreover, they cannot do so, for anything with explanatory power must have the potential to bring about understanding in virtue of the comprehension of its content.

The critique of psychophysical reductionism might seem too anthropocentric in that human cognitive limitations are taken to constrain which statements count as explanatory. One could claim that reductionism is true in some idealized sense, since the relevant generalizations really are explanatory relative to the cognitive powers of some creature of suprahuman intelligence. However, psychophysical reductionism is taken by its advocates to imply that psychological laws can be rewritten as physical laws, i.e., physical generalizations containing the same explanatory virtues as the psychological laws. So the conventional psychophysical reductionist is committed to the view that if psychological laws are explanatory relative to humans, so must be their corresponding physical rewrites.

So, in order for this methodological guideline to be legitimate, the physical rewrites would have to be explanatory for humans. Hence, the sort of reductionism which the psychophysical reductionist needs is one in which the physical generalizations have explanatory power for humans. Multiple realizability, however, prevents this from being the case.

CHAPTER TWO

LOCAL PSYCHOPHYSICAL REDUCTIONISM

In this chapter, I critique Kim's claim that functionalist metaphysics implies local reductionism. I first refute Kim's claim that belief in strong psychophysical supervenience carries a commitment to local reductionism. Subsequently, I refute his claim that multiple realizability implies local reductionism. The goal is to show that Kim's argument for local reductionism (the only one in the literature) presupposes an implausibly nonpragmatic view of explanation.

2.1. Does Supervenience Carry a Commitment to Local Reductionism?

Local reductionism is a more liberalized version of the Nagelian reductionism discussed in the preceding chapter. Whereas Nagelian reductionism is global in sanctioning a single reduction of psychology to physical science, local reductionism instead sanctions several distinct reductions of psychology, one reduction for each biological species or structural type. Local reductionism is superficially similar to Richardson's suggestion that bridge laws need not be required to state necessary and sufficient conditions but can in some cases state sufficient conditions only. That is, one might falsely presume that Richardson's claim that

physical-to-mental entailments suffice for reduction is tantamount to the claim that reductions can be local. But it is important to note that local reductionism and Richardson's view differ. Like Nagelian biconditional reductionism, Richardson's reductionism was shown to be at least prima facie incompatible with multiple realizability, owing to the impossibility of deriving psychological laws from physical laws without the aid of laws stating mental-to-physical entailments. The apparent difficulty posed by multiple realizability is that it precludes mental-to-physical entailments (other than ones involving enormously disjunctive physical predicates).

Local reductionism, however, is compatible with multiple realizability. Whereas a Richardsonian psychophysical reduction would use bridge laws which are simple one-way conditionals, and whereas Nagelian psychophysical bridge laws typically[1] are simple biconditionals, local reductionism's bridge laws are of the more complex form "S => (P <=> M)", where 'S' refers to some structural type which the bearer of psychology can possess such as membership in a given species, 'P' to a physical property, and 'M' to a mental property. Multiple realizability is compatible with the existence of such species- or structure-restricted correlations. More

---

[1] I say "typically" because Nagel's view, of course, also allows for one-way entailments.

specifically, even though multiple realizability might foil the attempt to formulate unrestricted biconditional and mental-to-physical entailments incorporating nondisjunctive physical predicates, it allows that such biconditionals can perhaps be formulated relative to such various restricted domains as the domain of reptiles, of mammals, of molluscs. Pain in mammals, for example, might be nomologically coextensive with physical property P_1, while pain in reptiles is nomologically coextensive with physical property P_2, even though P_1 and P_2 are distinct properties.

Granted that the existence of domain-restricted correlations is at least compatible with multiple realizability, is there any reason actually to believe in the former? The local reductionist must convince the functionalist that a law of form "S => (M <=> P_1)" holds in which 'P_1' does not represent a predicate which is too disjunctive to be humanly comprehensible.[2] The existence of such a law, however, does not deductively follow from any mere supervenience claim, e.g., that M <=> (P_1 v P_2 v .... P_n), so one might be tempted to conclude that the metaphysical functionalist's commitment to supervenience is not a commitment to domain-restricted correlationism. But, in fact, it is. For the functionalist belief in supervenience must be based on empirical evidence, and it is

---

[2] As established in the preceding chapter, only humanly comprehensible predicates can appear in bridge laws.

difficult to conceive of any evidence for supervenience other than the observation of psychophysical correlations. These correlations, moreover, must be found, at the very least, relative to restricted domains. Consider, once again, the claim that the mental supervenes on the physical, i.e., that M <=> (P_1 v P_2 v .... P_n). Just to be difficult, let us suppose that the series of P_1 is infinite. The correlation expressed in this claim is not humanly observable, since the infinite series exceeds human capabilities. So how could we possibly have grounds for believing it to be true? These grounds must consist in numerous observations of domain-restricted correlations which provide evidence for such laws as S => (M <=> P) or S => (M <=> (P_1 v P_2)). For example, we only believe that pain supervenes on the physical because we have grounds for believing that pain is coextensive with C-fibers firing (at least) relative to the domain of humans, and so forth for other domains. Therefore, whatever empirical evidence supports the functionalist belief in supervenience also supports belief in domain-restricted psychophysical correlations. (I believe that Kim's espousal of local reductionism partly stems from a sensitivity to the above considerations even though they do not explicitly figure in his work to date.)

Having established that the functionalist is committed to domain-restricted correlations, is it also established

that the functionalist is committed to local reductionism? In fact, it is not. One must bear in mind that local reductionism is meant to be a kind of nomic reductionism, identical to the reductionism of Nagel except that bridge laws now take a restricted form. Accordingly, the domain-restricted correlation statements can only serve as reductive bridge laws if they enable the logical derivation of psychological laws from physical laws. These domain-restricted correlation statements will only serve that end if psychological laws are themselves domain-restricted, and that they are so would require strong argument. On the contrary, the more reason one has for taking genuine psychological laws to be akin to the platitudinous generalizations of folk psychology, the more one doubts that such laws will turn out to be restricted in form.

Consider the following folk generalization,

(1) (x)[(x fears that p) => (x desires that -p)].

Is (1) derivable from a physical law via the local reductionist strategy of utilizing domain-restricted bridge laws? It is not, for consider what an attempted derivation would look like. A candidate physical law would have the form

(2) $(x)(P_1x \Rightarrow P_2x)$,

where the P_i refer to physical properties; and the domain-restricted bridge laws would have the forms,

(3) (x)[Sx => (P_1x <=> x fears that p)], and

(4) (x)[Sx => (P_2x <=> x desires that -p)], respectively.

Due to the unrestricted form of (1); (2), (3), and (4) do not entail (1). One thing which they do entail, of course, is

(1') (x)[Sx => [(x fears that p) => (x desires that -p)]].

the domain-restricted analogue of (1).

One could perhaps insist that the folk psychological generalization actually does have the restricted form of (1') and not the unrestricted form of (1). Unless, however, 'S' is construed to define a very broad class of actual and possible entities,[3] this is unlikely. Use of the generalization that fear that p causes desire that -p in both the prediction and attempted explanation of behavior commonsensically pertains to many nonhuman species.[4]

---

[3] And this would defeat the purpose of accomodating multiple realizability.

[4] If one is uncomfortable with the ascription of intentional states to nonhuman animals (which I am not), one may consider a law involving sensations in place of the one involving fear and belief, e.g., sensations of anxiety typically follow random sensations of pain.

Moreover, given our intuitive responses to much of science fiction, it is also plausibly seen as pertaining to such possibilia as creatures of a physical composition radically different from our own. That is to say, our folk theory, as gauged by our intuitive responses to such stories, shows no signs of being domain-restricted or of being divided up into separate domain-restricted theories. Hence, domain-restricted bridge laws do not enable the derivation of folk psychological laws. The folk psychological laws are too broad in scope to be derived from domain-specific generalizations. The upshot is that, despite the fact that the evidential grounds for supervenience also support domain-restricted correlationism, supervenience alone does not carry a commitment to local reductionism. What is needed for local reductionism is an additional premise to the effect that domain-restricted psychological generalizations are the only true laws in cognitive science.

2.2. Does Multiple Realizability Imply Local Reductionism?

The prospects for local reductionism need not appear so gloomy, however, if one questions the scientific worth of such unrestricted folk psychological generalizations as (1). One could perhaps reject (1) and replace it with various domain-restricted analogues having the form (1′). Returning to the example of pain, the local reductionist requires some reason for rejecting the scientific validity of the general

folk concept of pain and for replacing it with finer-grained pain concepts, each locally coextensive with a physical property. Kim has argued that the very multiple realizability of pain provides just such a reason. It is this second argument of Kim's which I now consider. (My response to the argument begins in the following section.)

Kim's argument is itself brilliantly ironic, if successful, in that he attempts to show that multiple realizability, the very thing which is so often thought to foil psychophysical reductionism, actually supports local reductionism. In order to understand Kim's argument fully, it is crucial to consider the antireductionist argument from multiple realizability to which he is responding. For Kim takes premises from that argument and uses them for his own reductionist purposes.

The antireductionist argument in question is in some ways similar to my antireductionist argument presented in the previous chapter. Both arguments are in response to the disjunction strategy. However, whereas my argument capitalized on the incomprehensibility of the physical predicates used in a putative reduction, the argument of current interest capitalizes on their nonprojectibility. Moreover, whereas the argument of Chapter One hinged upon the enormous disjunctivity of the relevant physical predicates, the current argument does not depend upon their enormousness. Even if pain were to have only two physical

basal conditions, the argument would be unaffected. This argument is perhaps implicit in Fodor and in David Owens,[5] but it finds its most explicit expression in William Seager.[6]

According to Kim's disjunction strategy, there is a disjunctive physical predicate which can be biconditionally linked to the predicate 'is in pain' to form a bridge law. Seager's antireductionist response to that strategy is to question the nomicity of the disjunctive physical predicate by casting doubt on its projectibility. Given that laws must be general claims which support counterfactuals, are explanatory and projectible, then the nonprojectibility of the relevant disjunctive predicate would render both the bridge generalizations and the relevant physical generalizations nonnomic and thus nonreductive.

A projectible predicate, of course, is one capable of playing a nontrivial role in a generalization capable of being supported by inductive evidence, i.e., a generalization which is confirmed by its instances. One classic example of a nonprojectible generalization is the claim that all nonblack things are nonravens.[7] Initially,

---

[5] David Owens, "Disjunctive Laws", Analysis 49 (1989) 197-202.

[6] William Seager, "Disjunctive Laws and Supervenience", Analysis 51 (1991) 93-8.

[7] The necessity of projectible predicates for projectible generalizations was first noted by Nelson Goodman. See his Fact, Fiction, and Forecast (Cambridge,

one might expect this generalization to be projectible since it is simply the contrapositive of the (apparently projectible) claim that all ravens are black. However, if it were projectible, then any green leaf or white tablecloth would tend toward confirming the claim that all ravens are black, for any green leaf or white tablecloth is both a nonblack thing and a nonraven. Evidently, what prevents the claim that all nonblack things are nonravens from being projectible is the appearance of such suspicious predicates as 'nonblack' and 'nonraven', predicates which are, accordingly, categorized as nonprojectible. (The criterion for distinguishing projectible from nonprojectible predicates will be discussed presently.) According to Seager, the disjunction of pain's basal conditions constitutes a nonprojectible and hence nonnomic property.

Let us suppose, for the sake of simplicity, that (throughout all physically possible worlds), pain has only two physical basal conditions, viz. F and G. The disjunction strategy would sanction (x)(x is in pain <=> (Fx v Gx)) as a reductive bridge law. Seager's claim, however, is that this is not a genuine law due to the nonprojectibility of the predicate 'FvG'. Hence, even though the generalization has the form of a law and sustains counterfactuals, its alleged failure to be projectible disqualifies it as a law.

MA: Harvard University Press, 1955), Chapters III and IV.

But why would the predicate 'FvG' be nonprojectible? More specifically, given that the physical generalizations (x)(Fx => Hx) and (x)(Gx => Hx) are projectible, how is it possible for (x)[(Fx v Gx) => Hx] to be nonprojectible? This is similar to asking how, given that (x)(x is a raven => x is black) is projectible, that (x)(x is not black => x is not a raven) could possibly fail to be so. The answer in both cases is that Boolean operations on projectible predicates do not always yield projectible predicates. The latter case shows this to be true in the case of negation. 'black' is projectible, but 'nonblack' is not. Owens and Seager claim that the same is true of disjunction. Suppose, by way of example, that 'emerald' and 'jadeite' are each projectible and that the claims 'All emeralds are green' and 'All jadeite is green' are each projectible. It does not follow that 'Anything which is either an emerald or jadeite is green' is projectible. The reason this is so, to put it roughly, is that observation of samples of green emeralds would not be relevant to the question of whether or not all jadeite is green. Nor would observations of green jadeite confirm the claim that all emeralds are green.[*] (This point will be expanded upon shortly.)

This is not, however, to claim that all disjunctive

---

[*] Kim makes this point using a slightly different example. See Kim, "Multiple Realization and the Metaphysics of Reduction", Philosophy and Phenomenological Research 52 (1992), 1-26; reprinted in Supervenience and Mind, 309-35. See pp. 319-22.

predicates are nonprojectible. (It is only to claim that predicate forming operations involving disjunction do not guarantee that projectibility will be preserved.) As Owens and Kim both point out, some disjunctive predicates are projectible and others are not. By way of illustration, consider the following claims:

(5) Anything which is either jadeite or an emerald is green.

(6) Anything which is either an African emerald or a nonAfrican emerald is green.

It is at least intuitively obvious that (6) is confirmed by its instances while (5) is not. Hence, it is not mere disjunctivity of form which renders certain predicates nonprojectible. The task of Owens and Seager is to discern what nonformal feature of 'is either jadeite or an emerald' renders it nonprojectible and to show that the physical predicate 'FvG' also bears that feature.

Classically, a projectible predicate is construed as one which guarantees a sufficient degree of similarity among the individuals which satisfy it. One can see how this would distinguish 'is either jadeite or an emerald' from 'is either an African emerald or a nonAfrican emerald'. Two individuals' satisfaction of the former guarantees less

similarity between the two than would their satisfaction of the latter. Presumably, the former predicate does not guarantee enough similarity for projectibility while the latter does. Moreover, the relevance of similarity to projection can be seen by constrasting (5) and (6). It is because 'is either jadeite or an emerald' does not guarantee sufficient similarity that evidence for one individual which satisfies the predicate being green is not necessarily relevant to whether another such individual is green. Instances of green jadeite do not constitute evidence for the greenness of emeralds and vice versa. By way of contrast, consider 'is either an African emerald or a nonAfrican emerald'. In this case, the degree of similarity which the predicate ensures does render the greenness of one individual evidence for the greenness of another. In order for the claim (x)(Fx => Gx) to be confirmed by its instances, the G-ness of one F must constitute evidence for the G-ness of any other (as yet unobserved) F, and this is the case only if 'is F' guarantees a certain sufficient degree of similarity among the individuals satisfying it.

The notion of similarity, however, is notoriously vague. If, for example, one were to stipulate that two particulars are similar precisely to the extent that they share properties, then any two particulars might very well turn out to be indefinitely similar. For, given a sufficiently liberal construal of what it is to be a

property, any two particulars would share an infinite number of properties. Kim has suggested, however, that the properties of relevance to the issue of projection are causal properties.[9] Kim's position probably reflects the established metaphysical view that the only real properties are those which play a role in the world's causal relations.[10] In any case, I see nothing wrong with adopting this position. Accordingly, two particulars' satisfaction of the same projectible predicate guarantees more similarity in their causal powers than would their satisfaction of a nonprojectible predicate, where similarity in causal powers is understood in terms of falling under the same causal laws. For example, to be correctly informed that two individuals satisfy the predicate 'is either an African emerald or a nonAfrican emerald' brings with it greater assurance that the two jointly fall under a certain substantial number of causal laws than would the information that they satisfy the predicate 'is either jadeite or an emerald'.

It is now clear why Boolean operations on projectible predicates do not necessarily yield projectible predicates. This is clear, for example, in contrasting 'raven' with 'nonraven'. We may assume, at least for the sake of

---

[9] Ibid.

[10] His assumption may also reflect the fact that a projectible property must be "well-behaved" or predictively reliable, and this depends upon its causal properties.

discussion, that 'is a raven' is projectible. Its projectibility owes to the fact that the mutual satisfaction of that predicate by two individuals guarantees sufficient similarity in their causal powers. But one now sees why the predicate 'is a nonraven' is not projectible. That two things both satisfy this predicate guarantees little (I would even say it guarantees absolutely no) similarity in causal powers. Any quasar, Fodor's most recent typing of the word 'Granny', and the contemporary malaise are all nonravens, and yet they have no causal powers in common. Similarly, to form a predicate disjunctively out of two projectible predicates need not imply that the resultant predicate guarantees sufficient causal similarity. The predicates 'is a neutrino' and 'is a market' are each, we may assume, projectible. Satisfaction of the former guarantees subsumption under certain microphysical causal laws. Satisfaction of the latter guarantees subsumption under many economic laws. What causal laws necessarily subsume all individuals satisfying the predicate 'is either a neutrino or a market'? Evidently, very few or none. Clearly, projectibility is not closed under disjunction.

Owens and Seager's task is clear: to show that the physical predicate 'FvG' fails to ensure the degree of causal homogeneity necessary for projectibility; i.e., to show that, even though 'FvG' might ensure some overlap in causal laws among the individuals which satisfy it, it does

not ensure enough for projectibility. This is the task which Seager appears to set for himself, and it is precisely the task which Kim sees the antireductionist as attempting to perform.

Moreover, given functionalist metaphysics, the causal inhomogeneity of FvG is assured. Recall that the second-order nature of pain implies that it is realizable by distinct physical kinds just as the second-order nature of fragility implies that it is realizable by distinct physical kinds. What does it mean to say that these physical kinds are distinct? Kim cogently points out that kinds should be construed as causal kinds and, hence, that distinctness of kinds implies distinctness in causal powers. A principle of kind individuation recommended by Fodor supports this point:[11]

Fodor's Principle: Individuals fall under a kind insofar as they have similar causal powers.

Fodor's Principle itself reflects the view that scientists are concerned precisely with projectible predicates and hence require a method of taxonomization which groups things together according to causal similarity; hence the rationale for viewing a scientific kind as guaranteeing some degree of

---

[11] Jerry Fodor, Psychosemantics (Cambridge, MA: MIT Press, 1987), pp. 33-4.

causal homogeneity.

The opponent of physical reductionism has precisely what he wants: given that the basal conditions of pain are distinct kinds as recognized by physical science (or, in our pretended example, that F and G are distinct kinds), they are distinct vis-à-vis their causal powers, just as being jadeite and being an emerald are distinct vis-à-vis their causal powers. The physical realizers of pain, in fact, are often spoken of as being "wildly heterogeneous" as kinds, thus implying that they are wildly heterogeneous as causal powers. The upshot is that 'FvG' is more like 'is either jadeite or an emerald' than 'is either an African emerald or a nonAfrican emerald' and, hence, is nonprojectible. With the nonnomicity of 'FvG' exposed, the project of formulating psychophysical bridge laws must be abandoned. For 'FvG' is not the sort of predicate which can appear in a law.

The foregoing is an argument against the disjunction strategy which is perhaps implicit in Fodor, more fully articulated in Seager, and explicitly presented in Kim. Now, something must be said about Kim's adopting an argument <u>against</u> the disjunction strategy. Given that Kim is the principal defender of the disjunction strategy, this may appear to be an inconsistency on his part. I believe, however, that he can be interpreted more charitably. What Kim is doing, I suggest, is presenting the antireductionist with a dilemma. Anyone believing in the physical multiple

realizability of pain has, I take Kim to be saying, two options, viz. one can either accept or reject the disjunction strategy. Accepting it obviously makes one a psychophysical reductionst. However, rejecting it makes one a reductionist too, a local reductionist specifically. That, at any rate, is what Kim means to show; and that is why Kim is here willing to assume that the disjunction of pain's basal conditions fails to form a projectible property. The aim of the preceding chapter was to show the disjunction horn of the dilemma to be false. The current chapter's aim is to debunk the remaining horn.

In a nutshell, Kim's strategy is to show that if FvG is nonprojectible, then so is pain. This strategy, if successful, fills the lacuna in the local reductionist argument noted earlier. Recall that it was not sufficient for the local reductionist to argue for domain-restricted psychophysical correlations. He must also show that pain-generalizations in the unrestricted folk style are nonnomic and hence replaceable by domain-restricted analogues. This is precisely what Kim takes multiple realizability to accomplish. The multiple realizability of pain implies its nonnomicity. Hence, the current folk notion should be replaced with finer-grained pain concepts, each corresponding to one of the basal conditions of folk pain. (These remarks are intended only to orient the reader toward the tenor of Kim's conclusion. Kim's argument is as

follows.)

Kim's counterargument depends upon the following principle which he explicitly endorses:[12]

The Causal Inheritance Principle (CIP): If higher-level property M is realized at time t in virtue of physical property P, the causal powers of this instance of M are identical with the causal powers of P.

Rejection of CIP commits one to believing that higher-level properties have causal powers which are not derived from physical properties. Kim is correct in noting that the functionalist should not reject CIP. Rejecting it limits one to unattractive options. One could embrace downward causation and hence deny the causal closure of the physical, or one could accept a systematic causal overdetermination. CIP does indeed seem plausible.

Now, one might indeed dispute CIP by suggesting that the causal powers of this instance of M are a proper subset of the causal powers of P.[13] In raising this suggestion, what one evidently has in mind is that it is not P simpliciter which realizes M, but P in causal context C. Accordingly, one would say that this instance of P is

---

[12] Supervenience and Mind, p. 326.

[13] Prof. Shaughan Lavine pointed out this possibility to me.

identical in causal powers to P-in-C. (For example, the firing of C-fibers simpliciter does not realize pain. The C-fiber firing must be situated in a functioning nervous system, etc.) Given that P might possess different causal powers in other contexts, one would conclude that the causal powers of this instance of M should not be identified with the causal powers of P in general. This M might not have all the powers of P in a different context.

But I disagree with this line of reasoning. It results from too narrowly construing the property which realizes M on the occasion in question. In raising the above objection, what one is calling "P-in-C" is what I am calling "P" in the formulation of CIP. For if P realizes M, then M strongly supervenes upon a set of properties which includes P. This being so, it follows that P => M is necessarily true. Given that P => M is necessarily true, P does not require any additional context in order to realize M. Hence, one can dispense with the notion of a causal context by viewing what is referred to as "context" as being part of the realizing property itself. This construal renders the powers of this instance of M identical to those of P. I submit that CIP in its current formulation should be accepted.

Given CIP, the causal powers of any pain-instance are identical to those of the physical property which is realizing it. Consequently, the causal powers of pain as

such are identical to those of its basal conditions in general. As Kim points out, this implies that pain is as causally inhomogeneous as the disjunction of its basal conditions. Kim concludes that since FvG is too disunified as a causal power to be projectible, then so is pain.

Kim provides what he takes to be a concrete example of how the causal inhomogeneity of pain precludes a pain-generalization from being confirmed by its instances. For example, even though humans provide positive instances of the generalization

(7) Sharp pains administered at random intervals cause anxiety reactions,

they do not confirm it, at least not according to Kim. Since pain is realized by a radically different physical property in (e.g.) Alpha Centaurians than it is in humans, (7) is more like (5) than (6). I.e., two individuals being in pain guarantees too little shared causal power for pain to be a projectible property, just as being either jadeite or an emerald guarantees too little. Therefore, claims Kim, evidence that pain causes anxiety in humans does not confirm (7) any more than the greenness of emeralds confirms (5).

Therefore, according to Kim, multiple realizability shows folk pain to cross-classify the true causal kinds. But a true scientific kind must not cross-classify the

causal kinds. It must be one of them. Hence, multiple realizability forces scientists to fracture the current pain concept into many finer-grained concepts. This is tantamount to replacing general, folk pain laws with finer-grained laws. I.e., laws of form (1) should be replaced by domain-restricted analogues such as (1'). This, in turn, shows the domain-specific correlations to ensure reductive bridge laws after all and thus guarantees the truth of local reductionism. If successful, Kim has turned the force of one antireductionist argument against itself. Multiple realizability is shown to imply local reductionism, which is far from what the antireductionist had intended.

Kim's argument can be summed up as follows: A commitment to supervenience carries with it a commitment to belief in domain-restricted correlations between mental and physical properties. As I have pointed out, these domain-restricted correlations are not genuinely reductive bridge laws unless it can be shown that psychological properties are just as fine-grained as the relevant physical properties. Our current folk laws of pain, however, posit much wider-grained psychological properties. However, Kim believes that multiple realizability shows these supposed "laws" to be nonprojectible and thus nonnomic. For, according to functionalist metaphysics, pain is realizable by a number of physical properties which are diverse as physical kinds. Their diversity as kinds implies that they

have diverse causal powers. Given that the disjunction of pain's basal conditions is too causally inhomogeneous to form a kind, then pain itself must be too causally inhomogeneous to form a kind. For pain derives all of its causal powers from its physical realizing properties, and the latter, evidently, are a causal motley. One must, according to Kim's argument, fracture the concept of pain into finer-grained concepts, each corresponding to one of the physical kinds realizing folk pain. In doing so, one introduces a much larger number of psychological laws, each one relativized to a physical structural type. The statements of domain-restricted psychophysical correlations can be used to derive these restricted psychological laws. Hence, multiple realizability, <u>if Kim is right</u>, when conjoined with supervenience, entails local reductionism.[14]

2.3. <u>A Relativized View of Kind Individuation</u>

Kim is incorrect in saying that if FvG is nonprojectible as a physical kind then pain is nonprojectible as a psychological kind. The criteria of projectibility and hence of kind individuation could vary among the different

---

[14] I wish to remind the reader that Kim only reaches this conclusion by assuming that FvG is too causally inhomogeneous to constitute an explanatory property. Thus, even Kim would have to admit that the conclusion is conditional. However, Kim seems to believe that the only alternative is to grant that FvG is an explanatory property thus implying global psychophysical reductionism. So, Kim believes that functionalist metaphysics implies psychophysical reductionism either way.

branches of science so that the nonprojectibility of FvG need not impugn the projectibility of pain. Further, as I will show, to think otherwise is to assume a nonpragmatic view of explanation. More specifically, Kim's argument depends upon the view that the antireductionist is inconsistent in saying that pain is projectible and that FvG is not (even though they have the same causal powers). I will argue that this appearance of inconsistency rests upon a nonpragmatic view of explanation.

Kim's challenge to the antireductionist can be made even starker by making a bold metaphysical assumption, viz. that type materialism is true; more specifically, that pain is FvG. I find this assumption plausible, and it will serve to simplify the argument.[15] In making this assumption, I am clearly not begging the question in the antireductionist's favor. For now, the challenge to the antireductionist is quite stark indeed. He must show that the very same property can be both projectible and nonprojectible, both an explanatory kind and not an

---

[15] It may be disconcerting to find someone presupposing type materialism in the process of defending an antireductionist position. But type materialism and psychophysical reductionism should be distinguished. Since type materialism concerns metaphysics only, there is no cognitive constraint to be considered in evaluating whether or not it is true. By contrast, since psychophysical global reductionism involves explanation, there is a cognitive constraint. Hence, reductionism can fail without impugning type materialism. This shows that the assumption of disjunctive type materialism does not conflict with the antireductionist stance of the first chapter.

explanatory kind.

That Seager's antireductionist argument appears to imply a contradiction shows that Kim has assumed a purely metaphysical view of kind individuation. If, on the other hand, one adopts a less purely metaphysical view, the appearance of contradiction disappears. On Kim's implicit view, the amount of causal homogeneity which satisfaction of a kind predicate must ensure remains invariant throughout all the various branches of science. To put the matter bluntly, Kim evidently assumes that for any property to be a scientific kind it must pass a certain test, viz. it must guarantee enough causal homogeneity, and how much is enough remains fixed from one scientific field to any other. On such a view, one can take a count of all the world's kinds just by checking to see whether any candidate property ensures the minimally required degree of causal homogeneity. That is to say, one could ascertain all the scientific kinds relative only to the world itself without having to consult the various sciences for their own standards of kind individuation. Accordingly, if a property were too causally disunified to be a kind in one science, it could not qualify as a kind in any science.

There is something attractive, falsely attractive, in the view that kind individuation must have an exclusively metaphysical basis. This attractiveness, however, derives from the assumption that the only alternative is a

thoroughly arbitrary or conventionalist view of kind individuation. For, given that Fodor's Principle sanctions taxonomization on a metaphysical basis, that of similarity in causal powers, one might conclude that a less than purely metaphysical view would be incompatible with Fodor's Principle. Without Fodor's Principle or some similar principle appealing to objective features of the world, the alternative methods of taxonomization remaining would be arbitrary, or so one might be tempted to conclude.

However, the rejection of any purely metaphysical method does not commit one to an entirely arbitrary or nonobjective method. In fact, it does not commit one to rejecting Fodor's Principle. The principle leaves open the possibility that how much causal homogeneity a scientific taxon must ensure may partly depend upon contextual factors such as which scientific field is at issue. This is, in fact, the view of taxonomization which I espouse. Taxonomization must be based on metaphysical considerations involving similarity of causal powers, but different scientific fields have different standards for how much causal homogeneity is minimally sufficient for kindhood, some adopting more lenient methods which include a wider range of properties as causal kinds, and others adopting more stringently exclusive methods. Hence, I do not advocate the rejection of Fodor's Principle but the adoption of a contextualized or relativized interpretation of it.

What I am suggesting, assuming that it is not already obvious, is that the property of being in pain (i.e., FvG) ensures enough causal homogeneity to be a kind according to the standards of psychology but not enough to be such according to the standards of any physical science. In Quinian language, this is to say that physical scientists use a more fine-grained similarity measure than do psychologists, a similarity measure which would not group pained individuals together even though the psychologist's similarity measure would.

Willard V. O. Quine anticipated this sort of relativized view of taxonomization some time ago.[16] According to Quine,

> Different similarity measures, or relative similarity notions best suit [classification in] different branches of science, for there are wasteful complications in providing finer gradations of relative similarity than matter for the phenomena with which the particular science is concerned. Perhaps the branches of science could be revealingly classified by looking to the relative similarity notion that is appropriate to each.

The plausibility of this position is supported by a consideration of the various branches of science. It is unlikely that the satisfaction of a predicate typical of geological theory guarantees as much causal homogeneity as would the satisfaction of a predicate typical of physical

---

[16] W. V. O. Quine, "Natural Kinds", in Ontological Relativity and Other Essays (New York: Columbia University Press, 1969) 114-38. See p. 137.

theory. For example, individuals sharing the property of having a fault line are less homogeneous as causal powers (and hence less predictable) than individuals sharing the property of being an electron. Moreover, it is evidently higher-level sciences, of which geology is an example, whose predicates guarantee less causal homogeneity and lower-level sciences whose predicates guarantee more. Presumably, satisfaction of a predicate of basic physics ensures a maximal degree of causal homogeneity. Satisfaction of a predicate of higher-level physics ensures somewhat less. Satisfaction of a chemical predicate ensures even less, and similarly as one ascends the hierarchy, the methods of taxonomization become increasingly more inclusive. Hence, it should not be surprising to find that a property counts as a natural kind relative to the standards of one scientific level but not so relative to the standards of any lower level. Therefore, it is at least very plausible to say that pain is a scientific kind relative to the standards of psychology even though FvG is not a scientific kind relative to the standards of physical science, even given the assumption that pain is FvG.

On this view, two individuals which are microphysically identical are also identical in terms of causal properties (i.e., CIP is true). However, above the microphysical level, one does not have such a strong guarantee. Instead, as one moves increasingly higher, one has increasingly

unsystematic correlations between a taxon and a collection of causal properties. So, whereas microphysical identity guarantees causal identity, biological identity only guarantees a limited amount of similarity. Psychological identity, presumably, guarantees even less. Of course, this reflects the fact that ever higher properties are ever more greatly multiply realizable. The microphysical basal conditions of (e.g.) pain are much greater in number than those of the firing of C-fibers. The latter, in turn, are greater in number than the basal conditions of being a DNA molecule, which are evidently greater than those of being a hydrogen atom. This inverted cone of ever increasing basal conditions explains why higher-level predicate subsumption gives one less assurance as to the causal powers and hence the behavior of a particular.

Hence, the fact that mutual subsumption under a higher-level predicate ensures less similarity among particulars than would mutual subsumption under a lower-level predicate, means that generalizations couched in the former are less predictively reliable than those couched in the latter. More precisely, an exclusively higher-level generalization is less reliable in making predictions than an exclusively lower-level generalization. This should not, however, be confused with the (probably false) claim that higher-level generalizations are statistical in nature while bottom-level generalizations are not. For, as is well publicized, word

is not yet in as to whether the laws of basic physics are irreducibly statistical. What I claim is that, even if basic laws are statistical, they are precisely quantifiable nonetheless. At ever higher levels, the laws remain statistical but it becomes increasingly difficult to quantify the statistical relations with precision. The higher one goes, the more the statistical relations approach randomness.

That being so, one might wonder why scientists ever rise above the level of the most basic physics. Why do they ever forfeit the maximal predictive reliability of the most fundamental predicates? The answer, evidently, lies in the greater generality of higher-level predicates. As one moves toward the broader end of the inverted cone, not only are the predicates wilder and less reliable, they are also increasingly broader in scope, applying to an ever larger range of phenomena. In other words, there is a greater variety of possible[17] pains than there are possible C-fiber firings, and there is a greater variety of the latter than there is of possible hydrogen atoms.[18] This increase in generality bestows on higher-level generalizations an ever wider range of application. Hence, the various levels of

---

[17] The possibility here is physical.

[18] I.e., there are more physical types of pain than there are physical types of C-fiber firings. Indeed, the set of physical types of pain subsumes the set of physical types of C-fiber firings.

science complement each other. Each level represents a decision as to how much predictive reliability to sacrifice for the sake of a gain in generality. Moreover, this shows why it would be a mistake for psychologists to follow Kim's advice in rejecting generalizations like (7) for finer-grained generalizations. If they were to do so, they would lose the generality characteristic of their science in exchange for an increase in precision characteristic of biology. In doing so, they would actually be abandoning the project of creating a psychological science and would simply be doing biology instead. But it is not necessary to do biology if someone else is doing it.

This is to say that psychology has more lenient standards for nomicity than do lower-level sciences. But this is not, of course, to say that psychology has no standards at all. Satisfaction of the predicate 'is in pain' must guarantee <u>some</u> degree of predictive reliability and hence causal homogeneity in order to be nomic. Hence, Kim's case against the nomicity of pain would be vindicated if it could be shown that the second-order nature of pain does not ensure any similarity in the causal powers of its realizing properties. However, the second-order nature of pain must indeed guarantee some identity in the causal powers of its basal conditions. Recall that a second-order property is the possession of some property or other which stands in certain specified causal relations. What those

causal relations are determines the identity of the second-order property. For example, the second-order property of fragility is the possession of any (first-order) property such that the individual breaks when subject to a certain degree of stress, etc. What this means is that in order for a first-order property to be a realizer of fragility, it must fit the right causal specification. But the right causal specification consists in falling under the right causal laws, e.g., the law about breaking when subject to a certain degree of stress. Functionalist metaphysics has it that pain is a second-order property. As such, only certain physical properties, those falling under the requisite causal laws, realize pain. Given CIP, the causal powers of pain must have some degree of homogeneity. The antireductionist can still, however, speak of pain's realizing conditions as forming an "unsystematic and heterogeneous lot," as he is prone to do, for he is here speaking from the more demanding perspective of physical science.

At this point, I must be careful so as to avert a possible misunderstanding. I am not defending Seager's antireductionist argument to the effect that FvG is nonprojectible. It is not necessary for me to do so, since Seager is arguing against global reductionism, and I have already presented an argument in the preceding chapter against global reductionism. (Of course, I am not

attempting to refute Seager's argument either.) Instead, I am critiquing Kim's argument for local reductionism which takes as one of its assumptions that FvG is not a physical kind. I claim that Kim cannot argue from the physical nonkindhood of FvG to the psychological nonkindhood of pain, even if pain is FvG. He would need to rule out the possibility of a relativized view of taxonomization. It is a view which not only has Kim not ruled out but which is at least prima facie plausible.

2.4. The Pragmatics of Explanation

The purpose of the first three chapters is to show that arguments inferring psychophysical reductionism from functionalist metaphysics assume implausibly nonpragmatic views of explanation. That is to say, such arguments typically presuppose that whether or not a putative explanans stands in the explanatory relation to a given explanandum is not even partly determined by local factors. The aim of this section is to show that Kim's purely metaphysical view of kind individuation also presupposes a nonpragmatic view of explanation.

The link between kind individuation and explanation is illuminated by considering projectibility. Scientists, it will be recalled, are interested precisely in the projectible properties, so the projectible properties are the scientific kinds. For the sake of discovering the

connection between projectibility and explanation, consider once again the psychological generalization

(7) Sharp pains administered at random intervals cause anxiety reactions.

As already noted, Kim claims that the multiple realizability of pain renders (7) nonprojectible. Now in saying that (7) is not projectible, Kim is saying that it is not confirmed by its instances. Supposedly, sharp pain causing anxiety in (e.g.) mammals does not confirm (7) in its present unrestricted form since this guarantees nothing as to the effects of sharp pains in reptiles. There is, moreover, a close connection between confirmation and explanation. This connection has often been noted and sometimes explicitly defended.[19] It is beyond the scope of this dissertation to defend the link or to reiterate arguments for it. It should be noted, however, that the link between confirmation and explanation has strong intuitive appeal. When one moves inductively from evidence (observation of instances) to a conclusion, one is moving from explananda to explanans. When one observes a few black ravens without ever having observed any nonblack ones, one is justified in inferring that all ravens are black only if that generalization

---

[19] It has been defended by Gilbert Harman in his "Inference to the Best Explanation", Philosophical Review 74 (1965) 88-95.

explains the blackness of the ravens one has seen. A generalization is confirmed by its instances just in case it explains them. Confirmation is, so to speak, the converse of explanation. That being so, in claiming that (7) is not confirmed by its instances, Kim is claiming that (7) does not explain them. More specifically, on the Kimian view, if someone experiences anxiety, it could never be explanatory to point out that the person has just experienced sharp pains at random intervals.[20]

A minimal degree of causal homogeneity is required in order that a property be a scientific kind. Given the link between projectibility and explanatory efficacy, this means that there is a minimal degree of causal homogeneity which a property must possess in order to be explanatory. If one were to defend Seager's view that the nonprojectibility of FvG precludes psychophysical reduction, one would have to contend that FvG is explanatory relative to standards of psychology while being nonexplanatory relative to (e.g.) biological standards. Hence, this antireductionist position makes use of a partially pragmatic interpretation of the explanatory power of a property. Kim's reductionist argument, on the other hand, must establish that the

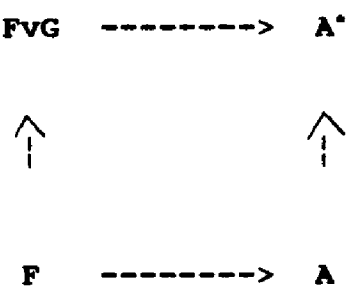
---

[20] The kind of pain of interest here is, of course, the comprehensive folk pain felt by humans and Alpha Centaurians alike. Kim obviously does believe that a finer-grained notion of pain could be appealed to in explaining some instances of anxiety, but that is not of current concern.

explanatory power of a property remains invariant across all sciences. Kim needs this nonpragmatic view of explanation in claiming that if FvG is nonexplanatory relative to a lower-level science, then it is nonexplanatory simpliciter. This implicitly Kimian view is, of course, quite dubious. Given recent work in the philosophy of science, it is more plausible to claim that a property's explanatory efficacy partly depends upon local factors, e.g., the particular interests which define a given scientific field.

It will help in discussing these matters to consider a specific example. According to the Kimian, in attempting to explain an instance of anxiety, one might be tempted to appeal to FvG or to some biological kind. Now if Seager is correct in saying that FvG is too inhomogeneous to figure in biological explanation, then (according to the Kimian) it is too inhomogeneous to figure in any kind of explanation at all. Therefore, as one might argue, FvG does not explain anxiety. Instead, it is to be explained by appealing to some biological property.

Consider the matter schematically:



The letters, including 'FvG', represent property instances. 'FvG' represents pain construed as a disjunctive property. (This reflects the type materialist assumption made earlier that pain just is the disjunctive property FvG.) 'F' represents a biological property realizing FvG and is the same property expressed by the left-hand disjunct of 'FvG'. 'A*' represents anxiety, and 'A' represents a biological property realizing anxiety. The arrows indicate relations of sufficiency such that 'P --------> Q' means that P is sufficient for Q. (The notion of sufficiency can be spelled out in terms of physically possible worlds.) The vertical arrows indicate the nomic sufficiency of the relevant biological properties for psychological properties.

The local reductionist claim is that FvG is too causally inhomogeneous to explain A* which can only be explained by either F or A. The Kimian position can, in fact, be posed as a dilemma to the antireductionist: Either FvG explains A* or it does not. If FvG does explain A*, then FvG belongs in the explanatory apparatus of biology, and so reductive bridge laws can be formed between psychology and biology after all. If FvG does not explain A*, then it must be replaced with properties which do, viz. F or A, and so local reductionism is true. My claim is that Kim has not considered the possibility that FvG might be able to explain A* relative to the interests of psychologists while failing to have this ability relative to

the interests of biologists.

Before developing this point, I would like to consider another response to Kim which might strike some antireductionists as appealing but which I am less inclined to endorse. Some antireductionists might claim that different sciences recognize different properties as explanatory because different sciences are concerned with different explananda. The gist of this antireductionist position is that the explanatory efficacy of a property must be judged relative to an explanandum, and since biology and psychology focus on different explananda, they will recognize different sorts of properties as possessing explanatory efficacy. Presumably, biologists will recognize more causally homogeneous properties as explanatory, for their explananda-properties are also more causally homogeneous. Psychologists, on the other hand, will recognize less causally homogeneous properties as explanatory, since their explananda-properties are less causally homogeneous.

I am not completely happy, however, with this approach, for I do not see that the inter-level sharing of explananda has been ruled out. It clearly has not been ruled out in the case of lower-level property instances explaining higher-level property instances. Since F and A are each nomically sufficient for A* (in the case of F, the sufficiency works via A), biologists could explain A* either

by appealing to F or to A.[21] (One would, at least, require some argument in saying otherwise, and I am not aware of any such argument.)

For the sake of simplifying the discussion, let us assume that biologists and psychologists are, in fact, both seeking explanations of A*. Kim's challenge to the antireductionist now becomes quite stark. The antireductionist must show that it is possible that FvG explains A* relative to the field of psychology while failing to do so relative to the field of biology.

There is, in fact, no difficulty in showing that it is coherent to say that FvG explains A* relative to the standards of psychology but not relative to the standards of biology. For one only needs to point out both the interest relativity of explanation and the fact that different

---

[21] One might object that my claim that different fields of science (corresponding to different levels of description) can share explananda is inconsistent with some of my claims in Chapter One. In the previous chapter, I claimed that the cognitive constraint on explanation prevents any one disjunct of the enormously disjunctive physical predicate from having explanatory efficacy. From this it would follow that such a single physical predicate could not be used in explaining a mental property instance. It might seem that I have now forgotten this. It might seem, that is, that I am now allowing that a higher-level property instance can be explained by a physical predicate. In fact, I am not, for one must bear in mind my use of the word 'physical' to mean 'pertaining to physics'. It is quite plausible that any attempt to explain a pain-instance in terms of basic physics would pass beyond human capacities. Too much complexity would be involved. But this is not at all to say that a pain-instance could not be explained by appealing to biological or neuroscientific predicates. Such an explanation of a pain-instance could indeed be within human abilities.

scientific fields are defined by different sorts of interests. There are two dimensions relevant to the current discussion along which biologists and psychologists have different interests. Both dimensions were noted in the preceding section. One is predictive precision. Lower-level scientists, in this case biologists, have greater standards for predictive precision than do higher-level scientists, in this case psychologists. Moreover, FvG is less predictively reliable than F due to the disjunctively heterogeneous nature of the former. FvG, we are free to assume, is sufficiently predictively reliable to meet the interests of psychologists while not being sufficiently reliable to meet the interests of biologists. The other dimension is generality. Psychologists seek greater generality than do biologists. That is why psychologists are willing to use generalizations such as (7) in explanations despite their diminished predictive reliability. This interest in generality inclines psychologists to accept FvG as having explanatory efficacy.

It is, in fact, not difficult to produce cases in which one's interest in a particular level of description has a direct bearing on which data one takes to explain a given explanandum. Someone ingests a pill and falls asleep shortly thereafter. How is one to explain this instance of falling asleep? In some quotidian contexts, it would normally be enough to point out that the person ingested a

sleeping pill, i.e., a pill with the property of dormitivity. Dormitivity is a higher-level, second-order property, the property of having some (first-order) property which causes sleep. It is not completely uninformative to cite dormitivity. As David Lewis has pointed out, doing so rules out the possibility that the cause of sleep was external to the pill's inner constitution.[22] However, in the context of the study of physiology, one would probably reject dormitivity as an explanatory property and appeal to chemical or medical predicates instead. For example, that the pill contained an antihistamine explains the instance of sleep. 'Antihistamine' is a lower-level predicate than 'dormitive'. The two predicates correspond to a realizing property and its realized property respectively. As a lower-level predicate, 'antihistamine' better satisfies the physiologist's or physician's interest in forming more predictively reliable generalizations than would the predicate 'dormitive'. Moreover, the physiologist or physician must sacrifice some generality in order to buy this increased precision, for 'anithistamine' subsumes fewer physically possible pills than does 'dormitive'. (This is not meant to contrast scientific explanations with pedestrian ones but to contrast the different descriptive levels from which people might seek explanantia for a given

---

[22] David Lewis, "Causal Explanation", Philosophical Papers, ii (Oxford and New York: Oxford University Press, 1986), 214-40.

explanandum.)

In sum, Kim has failed to take into account the possibility that different scientists working in different fields can have different standards for what sort of properties they consider explanatory even of the same explananda. It is, in fact, quite plausible that the explanatory power of a property partially depends upon what descriptive level the relevant scientists have chosen to investigate. One corollary of this position, which might seem initially surprising, is that projectibility itself must be partially interest relative. Given that projectibility is the converse of explanation and that explanatory efficacy is partially contextually determined, the very projectibility of a property must be partly contextually determined. Generalization (7), for example, might well be confirmed by its instances relative to psychological standards while failing to be confirmed by its instances relative to the standards of biology. If this is surprising, it is, I believe, because one has confused projectibility with predictive reliability. Predictive reliability is indeed something wholly noncontextual since it is grounded in the degree of causal homogeneity guaranteed by a property. The amount of predictive precision ensured by a property or generalization will not vary according to contextual factors. However, the degree of predictive reliability which is minimally necessary for

projectibility is indeed a variable matter. It reflects the amount of predictive precision which scientists working in a given field have come to expect in their generalizations.

2.5. Conclusion

Kim is correct in claiming that functionalists should acknowledge the existence of domain-restricted psychophysical correlations. For the functionalist belief that pain is a supervenient property could only be justified by the sort of evidence confirming such correlations. However, one cannot derive psychological laws with the aid of these domain-restricted correlation statements unless one takes the psychological laws themselves to be domain-restricted. Since the sorts of psychological laws in which we are here concerned at least approximate the age-old generalizations of folk psychology, they should be highly unrestricted, applying to any species or structural type as long as it exhibits the right sorts of behavior. Hence, domain-restricted correlations are not sufficient for reducing these laws.

Hence, the local reductionist must provide grounds for believing that psychological laws really are restricted according to various structural types. Kim takes multiple realizability, a consequence of functionalist metaphysics, to have this consequence. Kim borrows a page from antireductionists such as Seager to claim that the multiple

realizability of pain implies the causal heterogeneity of its physical realizing properties and hence the nonnomicity of those properties taken disjunctively. However, due to CIP, Kim takes this to mean that pain must be too causally inhomogeneous to be a natural kind. This is meant by Kim to show that folk pain is too wide-grained to be genuinely explanatory and so must be replaced by finer-grained pain concepts each coextensive with one of the physical realizing properties. This completes the argument for local reductionism by apparently invalidating the unrestricted folkish generalizations.

Kim, however, has not actually shown that the explanatory inefficacy of pain's physical realizers taken disjunctively implies the explanatory inefficacy of pain. For he has not ruled out the possibility that different fields of science have different standards for the explanatory efficacy of properties. Physicists (in their role as physicists) may indeed expect so great a degree of predictive precision in an explanatory generalization that they would reject such a disjunctive property as FvG. Psychologists (qua psychologists), however, certainly demand less predictive precision and so have more lenient criteria as to which properties are explanatory. A plausibly pragmatic view of explanation can account for this. Kim, however, must assume that the standards for determining the explanatorily efficacious properties must remain the same

throughout all scientific fields. This, however, is to overlook the importance of local interests in determining a property's explanatory power and so is to assume an implausibly nonpragmatic view of explanation.

CHAPTER THREE

ELIMINATIVE MATERIALISM

As is the case with local reductionism, eliminativism too attempts to argue from the multiple realizability of mental properties to their explanatory inferiority and does so only by disregarding pragmatic elements of explanation. I show that the strongest arguments for eliminative materialism presuppose what Frank Jackson and Philip Pettit have called the "fine-grain preference," i.e., the view that any explanation of a property-instance, e.g., that Jones is in pain at t, should be couched in terms of the lowest possible descriptive level thus placing the explanatory efficacy of multiply realizable properties in doubt.[1] Instead of calling this view the "fine-grain preference," I will refer to it as the "low-level preference" so that its import may be as perspicuous as possible. The low-level preference contrasts with the following pragmatic view of explanation: that which descriptive level is explanatorily relevant to an explanandum is partly a matter of the interests of those seeking an explanation. This pragmatic view will be further elaborated upon and defended in the following chapter. For now, it is enough to show that the case for eliminativism presupposes the falsehood of this pragmatic view.

---

[1] Frank Jackson and Philip Pettit, "In Defense of Explanatory Ecumenism", Economics and Philosophy 8 (1992), 1-21.

The low-level preference is itself ambiguous, and in some cases one interpretation and not the other motivates the eliminativist argument in question. The low-level preference could be taken to mean that the only genuine explanation of any property-instance is in terms of the very lowest level of description, i.e., basic physics. On the other hand, it could be understood as the view that whenever there is a genuine option between two or more levels of description, the lowest feasible level should be chosen in formulating explanations. So, for example, even if one does not have the resources to explain Jones' pain in terms of microphysics, if one can only explain the pain either biochemically or psychologically, one must choose the former option. The biochemical account, being couched in terms of a lower descriptive level than the psychological, is explanatorily superior. I shall not distinguish between these two possible interpretations of the low-level preference except when it is necessary to do so.

Eliminativism, as predicated on the low-level preference, is in striking contrast to the classic antireductionist argument from multiple realizability. According to that argument, it is the very multiple realizability of the mental which refutes psychophysical reductionism. However, according to the low-level preference, it is the very multiple realizability of mental properties which renders them explanatorily inadequate.

These radically divergent views concerning the significance of multiple realizability reflect fundamentally different views concerning explanatory efficacy, views which will be more fully examined in the chapter to follow. By raising considerations for the more pragmatic take on explanation, I wish to reveal the plausibility of the antireductionist position.

It is worth noting that the present chapter's concern with eliminative materialism may seem out of place to some. After all, eliminative materialists don't anticipate the reductive explanation of psychology in terms of physical science but the elimination of the former, or so one might claim. However, eliminativism is properly viewed as a form of psychophysical reductionism in that it is one way of furthering the unity of science project. In fact, I believe that a useful characterization of reduction is as follows: theory T reduces theory T′ just in case either T explains T′ or T replaces T′ as a means of achieving greater unification in the sciences. Hence, physical theory can conceivably reduce folk psychology in two distinct ways, viz. by explaining it or by replacing it. (I also point out in Section 3.5 that eliminativism and local reductionism are indistinguishable anyway. They are one and the same position, and so if the latter is reductionist, the former is as well.)

The strategy of the present chapter is as follows: in

Sections 3.2 through 3.4, the most cogent and influential eliminativist arguments are shown to reject folk psychological explanation simply because there are alternative ways of explaining behavior which are lower in descriptive level. (Section 3.1 summarizes and rejects those less influential eliminativist arguments which do not fit this pattern.) Accordingly, a corollary of functionalist metaphysics, viz. the multiple realizability of the mental, is taken by eliminative materialists to imply the truth of their doctrine. In Section 3.5, a pragmatic view of explanation, incompatible with the low-level preference, is broached. It is further elaborated upon in Chapter Four.

3.1. Other Eliminativist Arguments

My aim is to show that the dispute between psychophysical reductionists and their opponents turns upon whether or not one accepts a pragmatic view of explanation. Toward this end, in the current chapter, I attempt to show that the case for eliminativism rests upon the low-level preference which in turn rests upon a nonpragmatic view of explanation. However, not all eliminativist arguments exhibit the low-level preference. It is the task of the present section to consider those which do not and to show them to be less convincing than other eliminativist arguments. Accordingly, the best eliminativist case rests upon the remaining

eliminativist arguments, to be addressed in the following three sections, those which do presuppose the low-level preference. Hence, the current section is a kind of housecleaning, an attempt to dispense with those eliminativist arguments which do not fit the dialectic with which the dissertation is concerned.

One eliminativist argument to which the low-level preference is evidently irrelevant is to the effect that folk psychology fails to explain many phenomena which should fall within its domain of explananda, e.g., many phenomena involving human cognition and behavior. According to Paul Churchland, such phenomena include the psychological function of sleep, how memory is possible, how we construct a 3-D image from subtle differences in the array of 2-D retinal stimulations, mental illness, etc.[2] The apparent point is that if folk psychology fails to explain so many cognitive and behavioral phenomena, that must be because it does not reflect the true causal underpinnings of cognitive and behavioral phenomena in general. Supposedly, it is only a superficial misrepresentation of a deeper reality.

There are many viable responses, however, which this argument has not ruled out. For example, it might be the case that some but not all cognitive and behavioral

---

[2] Paul Churchland, "Eliminative Materialism and the Propositional Attitudes", The Journal of Philosophy 78 (1981); reprinted in Churchland, A Neurocomputational Perspective (Cambridge, MA: MIT Press, 1989), 1-22.

mechanisms are folk psychological in design. (Why assume that, for example, intentional behavior and sleep must be explained by the same theory?) In that case, folk psychology would be true, but it would not be the whole story. Another possibility is that folk psychology would explain these phenomena if the proper research were done. If the mechanisms of sleep and perception are unknown, they could turn out to be anything, including belief-desire mechanisms. What appears to be a failure of folk psychology might instead be our failure to attempt to apply it to these mysterious explananda.

Churchland also presents an argument to the effect that folk psychology fails according to a standard of scientific adequacy suggested by Imre Lakatos.[3] According to Lakatos, a good scientific theory should grow insofar as its applications expand and interesting new consequences are derived from it. But Churchland tells us that folk psychology has utterly failed on this score. The psychology of Thornton Wilder is largely the psychology of Sophocles. Moreover, Churchland claims, we are hardly better at explaining human behavior in folk psychological terms.[4]

This argument has been rightly criticized. Let us

---

[3] Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes", in Lakatos and Allan Musgrave (eds.) Criticism and the Growth of Knowledge (Cambridge: Cambridge University Press, 1970).

[4] Op. cit.

suppose, hypothetically, that folk psychology has failed to expand. According to Stephen Stich, it is still misplaced criticism to lay much emphasis upon folk psychology's remaining in a condition of stasis.[5] Lakatos' criterion is really only fairly applied to sciences in which researchers make a point of pushing their theories into new applications. In other words, it is only fairly applied to genuine research programs. Despite what Churchland implies, there have been many sciences whose basic categories remain respectable but which were stagnant for centuries simply because people saw no need to develop them further. Until the seventeenth century, for example, people did not attempt to extend empirical theories into new areas of application. The Aristotelian inspired medieval view of all knowledge was of a static system of logically linked concepts deductively grounded in unquestionable basic principles. Such a view did not encourage scientific progress, and, hence, one should not apply Lakatos' criterion so strictly to sciences during such a period of deliberate stasis. Hence, such a deliberate stasis need not reflect badly on the categories of a theory. For centuries, biology and chemistry remained pretty much unchanged, but this does not imply that they did not contain some truths or that none of their categories corresponded to natural kinds. That is to say, these fields

---

[5] Stephen P. Stich, From Folk Psychology to Cognitive Science: The Case Against Belief, (Cambridge, MA: MIT Press, 1983), p. 213.

were not yet research programs, and so it is unfair to critique them according to the standards of research programs.

When psychology finally emerged as a genuine research program in the early twentieth century, it did so in a non-folk psychological form, viz. eliminative behaviorism. As Stich writes,[6]

> It is only with the flourishing of the cognitive paradigm during the last decade or two [or three] that the idea of exploiting folk psychological notions in experimental psychology has [become respectable]. So those who would defend the conceptual apparatus of folk psychology might plausibly protest that the program of exploiting these notions in serious science has barely begun. The charge of stagnation is thus, perhaps, premature.

Andy Clark provides an interestingly different criticism of the same argument.[7] According to Clark, folk psychology has in fact shown signs of expansion into new areas of application. Our day-to-day explanations of each other's behavior are often enriched by new folk psychological categories suggesting new insights into the causes or reasons behind behavior. Freudian psychology is an obvious case in point. Clark also mentions some folk psychological concepts which appear to be enrichments over

---

[6] Ibid.

[7] Andy Clark, Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing, (Cambridge, MA: MIT Press, 1989), p. 40.

Sophoclean psychology, e.g., 'mauvaise foi', and 'Schadenfreude'. Perhaps it is better to compare Albert Camus rather than Thornton Wilder to Sophocles.

In the same passage, Churchland presents another argument for eliminativism which has received scarce attention in the literature. According to Churchland, folk psychology has been in a state of retreat, its domain of attempted application continually shrinking. In our paleolithic past, we were animists, ascribing intentional properties (and perhaps also such nonintentional mental properties as feeling cold) to the inanimate elements. The wind, moon, and sea were literally taken to have beliefs and desires. However, as Churchland notes, folk psychology is no longer so widely applied. Its domain of applicability has shrunk so as only to include higher animals. The point Churchland is evidently making is that there is a general tendency through history for the domain of folk psychology to contract due to explanatory failure and no reason to suppose that this contraction will ever stop short of complete elimination. Folk psychology finds ever fewer employments, and one day, as seems evident, it will have no application at all.

One might find Churchland's portrayal of the history of folk psychology to be somewhat persuasive. However, I submit that even if one does find it compelling, one should find the following account of folk psychology's history no

less compelling: people living in Stone Age cultures have attributed mental properties to the elements because they have taken the elements to be human beings with limbs, facial features, and so on. The earth, sky, and sun have each been construed as a human being of extraordinary capabilities. It is only in virtue of humanizing nature that such cultures have imbued nature with mentality. Churchland's story, by contrast, is that folk psychology was once applied to humans and nonhumans alike but has since come to be applied only to humans (more or less). My story is that people once took almost everything to be human and have since stopped applying mental predicates to the elements only because they no longer take them to be human. Notice that on my account it is misleading to say that the domain of folk psychology has shrunk. For it has not shrunk vis-ã-vis the sort of entity to which its predicates are applied, viz. humans and other similar beasts. Instead of saying that the domain of folk psychology has become more narrow, it would be preferable to say that our original estimate of how many things are human was too liberal. Perhaps one could reject my interpretation of the historical facts, but I suggest that it is no less plausible than Churchland's interpretation. Therefore, this eliminativist argument is far from conclusive.

There is another argument for eliminativism which was first proposed by Churchland and then further elaborated by

Stich. According to Churchland, intentional properties cannot be comfortably ascribed to very young infants despite the fact that they exhibit a great deal of behavior in need of explanation. Hence, the explanation of their behavior will have to be couched in nonintentional, probably neuroscientific, terms. However, adult humans, the paradigmatic sufferers of intentional properties, have approximately the same sorts of causal mechanisms underpinning their behavior as very young infants have underpinning theirs. That is to say, the structure of the central nervous systems of adults and infants are appreciably similar. But if the explanations of infant behavior are devoid of any appeal to intentional properties, then so should be the explanations of adult behavior. Ergo, at least one important aspect of folk psychology, its appeal to intentional properties, should not be found in an adequate theory of behavior.*

Stich has expanded upon this argument, or proposed a very similar one. Stich echoes Churchland's claim that intentional generalizations are less comprehensive in explaining behavior than some nonintentional (in Stich's view, syntactic) ones. He claims that not only infants but

---

* Paul Churchland, Scientific Realism and the Plasticity of Mind, (Cambridge: Cambridge University Press, 1979), pp. 129-33.

also the psychotic[9] and the severely brain damaged can be subsumed under nonintentional generalizations which also explain the behavior of normal humans even though only the latter can be comfortably ascribed intentional properties.[10] Stich's point is that opting for an intentional science of behavior would result in our losing sight of important generalizations which cover the infantile, deranged, and brain damaged as well as normal adult humans.

It is, of course, crucial to this argument that there be such inclusive, nonintentional generalizations which exceed the comprehensiveness of any intentional theory of behavior. Given the current state of philosophy of mind, however, I take this to be very much an open question. Given attempts by many philosophers to find sufficient physical conditions for the possession of any given intentional property, it may well turn out that we will one day be able to assign definite contents to the mental states of the infantile, deranged, and brain damaged.[11] A

---

[9] In the case of psychosis, Stich's point is somewhat strained, for psychosis is partially characterized in terms of the holding of highly unrealistic beliefs.

[10] Op. cit., Chapter 7.

[11] For the classic discussion of the attempt to naturalize content, see Fred Dretske, Knowledge and the Flow of Information, (Cambridge, MA: MIT Press, 1981).

successful naturalization[12] of intentional psychology could enable us to make the content attributions which now seem unavailable. The contents of the mental states of exotica could be awaiting discovery.

One could also question this premise on other grounds. That is, instead of positing that there are, in fact, intentional generalizations which include everybody, one could raise doubts as to whether there are nonintentional generalizations as inclusive as Churchland and Stich claim there to be. Are there indeed nonintentional generalizations sufficiently egalitarian to subsume the exotic and nonexotic alike?[13] Perhaps there are, but there are at least two reasons for doubting that this is so. One reason is that there are physiological differences between the normal and adult, on one hand, and the abnormal or infantile, on the other. This fact is especially salient in considering those suffering brain damage, but it is also a fact worth considering in regard to the infantile and deranged as well. These physiological differences could underpin different explanantia for normal adults as opposed to the abnormal or immature. Hence, there might not be any

---

[12] A naturalization, by reason of providing only sufficient conditions for the having of a semantic property, is not a reduction. As argued in Section 1.3, physical-to-mental entailments are not sufficient for reduction.

[13] At the level of physics, there would be. But Churchland and Stich are evidently looking at a biological or asemantic computational level respectively.

gain in generality by discarding intentional explanation. The other reason pertains to the explananda themselves, i.e., the sorts of behavior being explained. The behavior of a normal adult is appreciably different from that of an infant or a psychotic. Given that the very thing to be explained is different, as it is in these cases, this raises the possibility that the generalizations appealed to in the explanantia are themselves different. So there is reason to doubt that there really are these nonintentional and more comprehensive generalizations as Stich and Churchland believe there to be.

However, as Colin McGinn has pointed out, even if one grants this premise to the eliminativists, the eliminativist conclusion does not follow. That is, one can grant that there are nonintentional generalizations explaining behavior which are more inclusive than the intentional ones without thereby being forced to reject the latter as nonexplanatory.[14] For if Churchland and Stich are correct in their assumptions, what one has here is a case in which one group of organisms is a subset of another; a set of laws explains the behavior of the entire set, whereas there is another set of laws or generalizations which only purport to explain the behavior of organisms in the subset. The first question to ask: why is this explanatory overdetermination a

---

[14] Colin McGinn, Mental Content (Oxford: Basil Blackwell, 1989). See p. 128.

bad thing? Why can't normal adults fall under their own unique explanatory generalizations while also falling under the more comprehensive laws? McGinn suggests that the greater comprehensiveness of the nonintentional generalizations need not cancel out the explanatory force of the intentional ones. Acknowledging one explanatory taxonomization does not carry a commitment to renouncing the other.[15] "Logically speaking, what we have is a semantically [i.e., intentionally] specified species falling within a [nonsemantically] specified genus. Why give up the narrower groupings just because you want to recognize wider ones? You can have both."[16]

The preceding was intended as a housecleaning. The less influential eliminativist arguments have been shown to be wanting. The eliminativist arguments to be considered in the following sections are usually taken as more serious threats to the scientific status of folk psychological categories. I will attempt to show that each of them requires the assumption that psychological explanations (or

---

[15] That the eliminativist argument necessarily presupposes that two distinct taxononomizations of the same phenomena are incompatible possibly suggests a nonpragmatic view of explanation. For it excludes the possibility that one could move from one taxonomization to the other as a function of local interests. However, I will not pursue this possibility. It is enough, I believe, simply to point out how implausible this particular eliminativist argument is.

[16] Ibid., n.

what are alleged to be explanations) fail to be genuinely explanatory by virtue of there existing lower-level alternative accounts.

3.2. The Poorness-of-Fit Argument

According to Churchland, a theory may merit serious consideration simply because it shows promise of being integrated with other sciences of overlapping domain. This is so even if the theory fares poorly in regard to other criteria of theory evaluation. Folk psychology, according to Churchland, rates very poorly in terms of showing "integration" with related sciences. Churchland notes that many sciences which take humans as at least in the domain of their subject matter, e.g., organic chemistry, evolutionary theory, biology, neuroscience, etc., tell a "coherent story of the species' constitution." Churchland describes this cluster of sciences as presaging "the greatest synthesis in the history of the human race." However, folk psychology "is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus."[17]

What exactly is Churchland's argument? At least two different arguments could be gleaned from this passage. One argument is to the effect that the intentional predicates of

---

[17] "Eliminative Materialism and the Propositional Attitudes", pp. 8-9.

folk psychology do not appear in sciences which study humans. If this were an accurate assessment, it might carry some force. But if this is Churchland's argument (and indeed it might not be), it is unsound. Stich has made this point quite well.[18] The social and political sciences, for example, would be unrecognizable without the intentional idiom. This particular interpretation of the passage yields an argument based on an unfair assessment of folk psychology's relation to other sciences which are equally high-level.

There is, however, another argument which can be gleaned from the aforementioned passage of Churchland. This argument is, I believe, a better exegesis. According to it, folk psychology will not be reduced to any lower-level science and, for that reason alone, it should be rejected as pseudoscience.[19] Some of Churchland's own remarks suggest that this is how the Poorness-of-Fit Argument is meant to be taken. In presenting the argument, Churchland adds that "[a] successful reduction cannot be ruled out [], but" folk psychology's many other supposed problems inspire little confidence.[20] Moreover, he also states quite

---

[18] Op. cit., Chapter 10.

[19] Churchland might also be assuming, unreasonably I think, that evolutionary biology is necessarily lower in level than psychology. For a discussion of this issue, I refer the reader to the Appendix.

[20] Op. cit., p. 9.

straightforwardly in a different context that "[e]ither [folk psychology] must be successfully reduced (to a matured cognitive neurobiology, for example), perhaps undergoing some modification in the process. Or it must be displaced by a better theory, one that does cohere with the rest of our growing scientific corpus."[21] Churchland is presumably anticipating a theory which is at least reducible to neurobiology if not basic physics.

Hence, on this interpretation, the Poorness-of-Fit Argument is the claim that folk psychology should be rejected in virtue of its categories failing to map onto lower-level categories. Churchland's model of reduction is Nagelian in that he takes one theory to be reducible to another in virtue of the availability of bridge laws making it possible to derive the secondary science from the primary science. Given the classic formulation of Nagelian psychophysical reductionism, as presented in the first chapter, there must be a one-one mapping of any psychological kind onto a lower-level kind. Hence, given that Churchland is assuming this Nagelian view of reductionism, he is insisting that any psychological kind must be identical in nomic extension to some lower-level kind. Hence, Churchland is committed to rejecting the

---

[21] A Neurocomputational Perspective, p. xii.

explanatory power of any multiply realizable property.[22] That is to say, he must reject any explanation which cannot be translated into a lower-level explanation.

In the following sections, one will find this to be a persistent feature of eliminativist argumentation: the lower in level is always favored over the higher in level, the physical over the nonphysical.

3.3. The Connectionist Argument

Some philosophers have argued that connectionism has eliminativist ramifications. As I shall show, this view is also motivated by the low-level preference.

Connectionist networks are models for performing a large number of distinct but closely related computations simultaneously. They at least roughly resemble and were initially inspired by assemblages of interconnected neurons in the brain. Eliminativists who take connectionism as damning of folk psychology and its ontology typically take connectionist networks to be sufficiently faithful reflections of the microstructure of the brain and its activities. Their contention is that connectionist models show that a future completed neuroscience will not posit

---

[22] It is important to note that Churchland's discussion predates Kim's proposal of the disjunction strategy, so there is every reason to believe that Churchlandian reduction requires that explanatory kinds map onto maximally fine-grained (nondisjunctive) physical kinds. Psychological properties are, hence, rejected for being too wide-grained.

anything approximating folk psychological properties. In order for this conclusion to be even remotely plausible, one must assume that connectionist networks are at least approximately brainlike. For the sake of discussion, I will grant this assumption. Hence, in speaking of connectionist networks, I will also be speaking hypothetically of the nature of the brain itself and the (at least) approximate form which a completed neuroscience might take. Of course, if this should turn out to be false and the structure and function of the brain accordingly unlike connectionist networks, the eliminativist position will forfeit this argument.

Another important assumption of the eliminativist argument from connectionism is that such models be construed as models of cognition and not mere models of how to implement psychological programs in neural wetware. It has been an implicit assumption in the two previous chapters that the psychological and the neural are distinct levels of description. In speaking, for example, of the realization of pain in creatures which do not have brains (e.g., Alpha Centaurians), or which do have brains in some sense but their "brains" being structurally quite unlike ours, it was tacitly assumed that psychological properties are more abstract than neural properties. The neural is simply one of the basal conditions of the psychological. Therefore, since connectionist models are neurally inspired, it is more

in keeping with this assumption to view them as possible models for the implementation of psychological algorithms. Accordingly, the eliminativist assumption that such models are actually cognitive models and not simply models for implementing cognitive models assumes that the psychological does not constitute a separate level of description but instead resides at the neural level of description. Eliminativists simply assume that the psychological should be collapsed into the neural. This very assumption, as I will show at a later point, derives whatever plausibility it has from the low-level preference. However, for now, I am willing to grant eliminativists this assumption as well. Let us assume, then, that connectionist models really are models of cognition. My most immediate aim is to show that even with this assumption, the elimination of the ontology of folk psychology does not follow. The assumption itself will be questioned at a later point.[23]

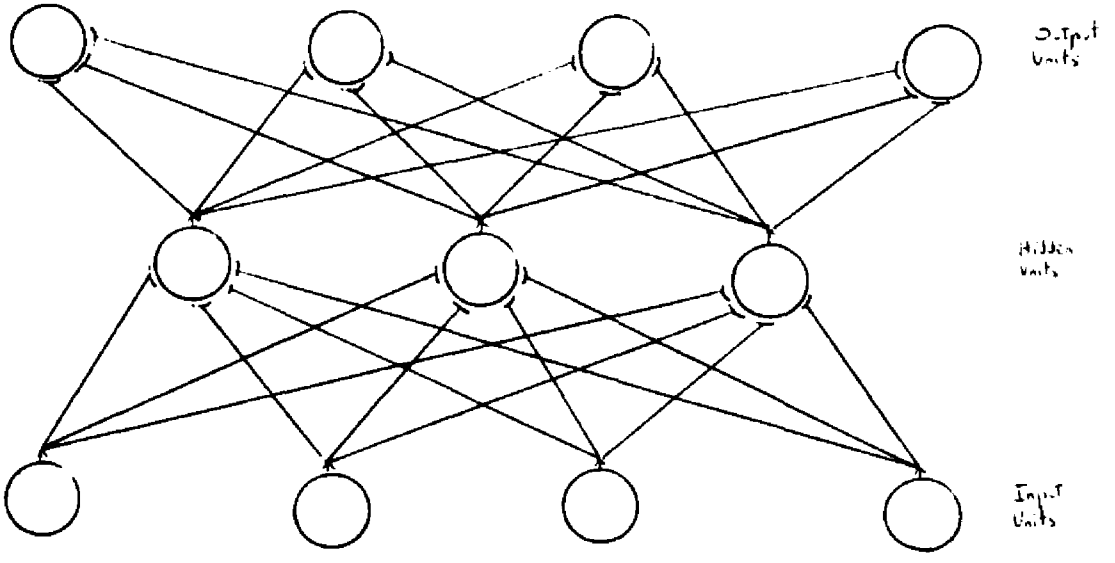
A connectionist network, although it can be extremely complex as a whole, is composed of relatively simple computing elements which operate "in parallel," i.e., simultaneously. One important type of simple computing element is the unit. In neuroscientific terms, it is meant to correspond to the cell body (the soma) of the neuron.

---

[23] Connectionist researchers take different positions on this matter, some viewing the models as cognitive and others viewing them as implementations of nonconnectionist cognitive programs.

Each unit produces an output, an activation value. This output is of a certain strength or intensity which can be characterized numerically. A unit computes its activation value at any given time partly from the inputs it receives from neighboring units. However, its activation value can also partly be a matter of a "bias" assigned to the unit itself. Units are arranged in layers or "populations," as they are sometimes called. At any given moment of activity, the units composing a particular layer have a number of different activation values. The total activation values of a given layer can be represented as a series of numbers, a vector. Connections link each layer to its neighboring layers. These connections are meant to correspond to the axons and dendrites which link neuronal cell bodies. Typically, there is one connection leading from a unit to each of the units in the adjacent layer. Hence, if a layer receives input from another layer, then the activation value of each of its units is a function of the values entering it via the connections to the other layer as well as its own particular bias. Each connection has a weight which changes the level of activation being fed into it. Hence, the activation value of a unit is not simply the result of its own bias in conjunction with the activation values of all the units feeding information into it, but it is also partly a function of the weights of the connections bringing it that information. For the connections themselves compute by

altering the degree of activation which they are fed according to the weight of each connection. Speculatively, connection weights correspond to actual features of synapses, i.e., of the points at which signals from the dendrites of one neuron are transmitted to the cell body of another. Hence, connection weights are sometimes referred to as "synaptic weightings." The layer of input units is specifically concerned with bringing information into the system. A vector is simply imposed upon the input units. This vector is a representation of the input, and each of its individual activation values represents some feature of the thing represented. The vector of the layer of output units represents the final result of the network's computations. In many networks, there is at least one layer of units between the input and output units. The addition of these hidden units increases the computational powers of a network. The following illustrates a possible network:



The connection weights and unit biases determine how the network transforms an input vector into an output vector. The combination of weights and biases encodes the knowledge of the network and corresponds to a program in more traditional artificial intelligence.

Organic brains learn to perform new computations. At the very least, a brain's repertoire of recognitional capacities is continually increasing. Accordingly, a neurally realistic connectionist network must have some means for learning how to perform new computations. Toward this end, many connectionist networks have procedures for adjusting their own weights in order to perform new computations. Rather than the connectionist researcher performing a hands-on manipulation of the weights to accomplish new computations, the networks are usually trained on examples. There are, that is to say, various automatic procedures for adjusting weights upon being fed the same set of input vectors repeatedly. Paul Churchland has provided an interesting example of the sort of learning task which can be achieved by such automatic procedures.[24] In Churchland's example, a network is fed a large number of input vectors one half of which correponds to sonar echoes from mines while the other half corresponds to sonar echoes from rocks. (Each input vector is derived from a frequency

---

[24] Paul Churchland, "On the Nature of Explanation: A PDP Approach", in A Neurocomputational Perspective, 197-230.

analysis of each echo.) The goal is for the network to produce a mapping of all mine echoes onto a single output vector and to produce a mapping of all rock echoes onto another output vector. In other words, the network's training should eventuate in its being able to discriminate mine from rock echoes. An automatic procedure achieves this end upon repeated exposure to the same input vectors by correcting the connection weights according to a mathematical rule in reponse to "mistaken" outputs. The network's margin of error is continually reduced until it consistently and correctly distinguishes rock echoes from mine echoes. Networks of the appropriate complexity can also be trained to make multiple discriminations. Sejnowski and Rosenberg's NETtalk model maps inputs corresponding to inscriptions of letters of the alphabet onto outputs corresponding to phonations.[25] In simpler terms, it computes functions from written text to spoken sounds. In even simpler terms, it reads aloud.

Connectionist networks come in different stripes only some of which are relevant to the eliminativist case. In localist networks, individual units or very small clusters of units are intended to represent specific elements of propositions. One could also make this point by saying that

---

[25] Terrence J. Sejnowski and C. R. Rosenberg, "NETtalk: A Parallel Network that Learns to Read Aloud", Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01, (Baltimore, MD: The John Hopkins University, 1986).

in such networks it is easy to impose a definite semantic interpretation on any given unit or small group of units. For example, a unit in one layer might represent dogs while a unit in another layer represents the possession of fur. A strong positive connection weighting leading from the former to the latter would mean that the network has encoded the information that dogs have fur, and the having by each unit of a sufficiently high activation value would indicate that the network is accessing its knowledge of the fact that dogs have fur. Eliminativists do not appeal to localist networks in their case against folk psychology, nor is it likely that such networks are faithful representations of the relation of semantics to neural structure. That is, it is not biologically realistic that a single pair of neurons would encode a propositional content corresponding to a sentence of ordinary language. If that were the case, then it would be possible to destroy specific memory traces, such as one's memory that George Washington once slept in the Morris-Jumel Mansion, by destroying highly specific parts of the brain. But, evidently, this is not the case.

Distributed networks, on the other hand, appear to be more neurally realistic and are also the sorts of networks of interest to eliminativists.[26] In these networks, one

---

[26] For the classic account of distributed networks, see David E. Rumelhart, James L. McClelland, and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations, (Cambridge, MA: MIT Press, 1986).

cannot ascribe specific propositional contents to individual weightings or activation values beyond the input layer. Instead, any given propositional representation is distributed across many units. Hence, it is not possible to assign to any given unit the sort of semantic interpretation which one would associate with the terms of a proposition. For example, we may know that a network has encoded the proposition that dogs have fur, and yet there is no single unit to which one could assign the meaning corresponding to the term 'dog' or 'fur'. This does not necessarily or apparently imply that an individual unit can receive no semantic interpretation. What it does imply, however, is that, even if there is such an interpretation, it would not correspond to any of the concepts we ordinarily ascribe to people in folk psychological explanation. Paul Smolensky, for example, believes that individual activation values in distributed networks[27] can be given semantic interpretations but that these interpretations can only be "subsymbolic" (i.e., subconceptual), not corresponding to the concepts recognized by the folk.[28] Smolensky cryptically suggests that each activation value of a hidden unit encodes a "microfeature" of the propositional content

---

[27] Whenever I refer to connectionist models in the following, it will be of the highly distributed sort.

[28] Paul Smolensky, "On the Proper Treatment of Connectionism", Behavioral and Brain Sciences 11 (1988), 1-74.

represented.

One might suppose that there is a higher and more abstract level of description of connectionist systems which would correspond to the propositional contents countenanced by folk psychology. Indeed there is, but for now I will focus on the lower, subconceptual level. It is this lower level upon which eliminativists focus, and it is because of their preoccupation with this lower level that connectionism appears to them to suggest the elimination of the folk psychological taxonomy. To date, the general form of all eliminativist arguments from connectionism is to the effect that folk psychological properties, such as the property of believing that dogs have fur or the property of being horrified that dogs have fur, do not correspond to the properties sanctioned by connectionist theory. And the failure to find a place for such folk psychological properties in connectionism results from an exclusive focus upon the most fine-grained level of description; that of individual weights, biases, and activation values.

Let us consider one such characteristic eliminativist argument. William Ramsey, Stich, and Joseph Garon have claimed that connectionism, if taken as reflecting a true psychoneural theory, shows intentional folk psychological properties not to be natural kinds.[29] More specifically,

---

[29] William Ramsey, Stephen Stich, and Joseph Garon, "Connectionism, Eliminativism, and the Future of Folk Psychology", in Ramsey, Stich, and Rumelhart (ed.),

they claim that folk psychology groups organisms together as belonging to the same kind in many cases in which connectionist neuroscience would not. They conclude that if connectionist models are taken as cognitive models, they pose a genuine threat to the folk psychological postulations of natural kinds. For example, two humans both sharing the belief that water is wet would be classified as belonging to the same folk psychological kind in virtue of that. However, that commonality alone, according to Ramsey et al., would not be enough to classify them as belonging to the same kind according to the methods of kind individuation found in connectionism.

Ramsey et al. attempt to illustrate this point by appealing to two actual connectionist networks. Admittedly, these networks are little more than toy models. The actual structure of a relevant portion of a human brain would be far more complex. Ramsey et al. hope, however, that the relevant features of these simple networks will also pertain to actual networks of neurons. We may, that is, be seeing here in a simpler form some important features found in the real neural networks someday to be described by a completed neuroscience. Ramsey et al. describe two networks which are initially (i.e., prior to training) indistinguishable. Each network is composed of sixteen input units, four hidden

---

Philosophy and Connectionist Theory, (Hillsdale, NJ: Lawrence Erlbaum Associates, 1991), 199-228.

units, and one output unit.[30] One such network is trained to respond appropriately to sixteen propositions, each one of which is encoded as an input vector. After a complete training, the one output unit produces either a very high activation value (interpreted as meaning 'true') or a very low activation value (interpreted as 'false') in reponse to each proposition. E.g., if fed the proposition 'Cats have fur', the fully trained network would respond 'true'. If fed the input 'Dogs have scales', it would respond 'false'. The other network is trained on seventeen propositions instead of sixteen. Moreover, the sixteen propositions of the first network are a subset of the seventeen propositions of the second. Hence, the second network is simply being trained on the same set of propositions but with one additional proposition. Since each proposition is encoded in a widely distributed manner, there is no particular subset of hidden layer units corresponding to a single proposition. That being so, the difference made by the addition of the seventeenth proposition in the second network amounts to a difference in the assignment of weightings throughout the network. Both networks believe that dogs have fur in the sense that they both produce a 'true' output if fed that proposition as input. However, there is no subsection of the assignment of weightings which

---

[30] This is not a hypothetical example. Ramsey et al. actually devised and ran the programs realizing both networks.

remains constant between the two networks to correspond to that commonality. According to Ramsey et al.,[31] the two networks, despite the fact that they are largely similar in folk psychological terms,

> have no projectible features in common that are describable in the language of connectionist theory. From the point of view of the connectionist model builder, the class of networks that might model a cognitive agent who believes that dogs have fur is not a genuine kind at all, but simply a chaotically disjunctive set.

NETtalk itself provides an example of the same sort of phenomenon which Ramsey et al. might find to be even more compelling. In training a NETtalk network, one begins with a random set of weightings which subsequently change during the process of training. Two NETtalk networks which begin with different weightings can, when trained upon the same samples, achieve the same discriminative powers. That is to say, each will produce the same output upon exposure to the same input. However, the actual connection weights in each network can be markedly different. NETtalk, as already noted, is a network for converting text to phonemes. The fact that two such networks could be indistinguishable in terms of discriminative abilities while being distinct in their connection weightings, shows that other networks which encode propositional information could also be distinct in

---

[31] Ibid., p. 213.

terms of weightings while being indistinguishable in terms of their stored propositional knowledge. Hence, as the eliminativist would evidently wish to interpret the matter, two networks which are folk psychologically indistinguishable can be significantly different from the viewpoint of connectionist psychology. Folk psychological natural kinds, so the eliminativist conclusion goes, are splintered or dissolved by connectionist psychological taxonomization.

Now, it is important to Ramsey et al.'s case that the science of connectionism itself, supposedly a prototype of a mature cognitive neuroscience, allows no room for folk psychological categories. Clark, however, has shown this to be false.[32] Connectionist researchers, in considering the information stored by networks, have, in fact, found it highly convenient, if not indispensable, to appeal to a more abstract level of analysis than simply recording all the connection weights and unit biases. For the latter level of description can quickly become highly complex and is often not a perspicuous means of understanding why two networks which are different in terms of weights are performing relevantly similarly. There is, accordingly, a more abstract level of description at which the folk psychological kinds could, evidently, be represented. As

---

[32] Andy Clark, "Connectionist Minds", Proceedings of the Aristotelian Society (1990), 83-102.

Clark indicates, cluster analysis is just such a more abstract means of taxonomizing connectionist networks. Cluster analysis was, in fact, developed by Sejnowski and Rosenberg as a means of co-classifying NETtalk networks whose performance skills are relevantly similar. Consider cluster analysis as it specifically pertains to NETtalk. In forming a cluster analysis, one begins by recording the hidden vector and output vector for each of the inputs. For all the inputs which yield the same output, one takes all of the hidden layer vectors and averages them so as to produce an average hidden layer vector. This is done for each distinct output. Subsequently, the most similar hidden layer vectors are grouped together into pairs, and then one finds the average of each such pair. One then takes these new vectors and groups them into the most similar pairs. This procedure is iterated until one has only a single pair of vectors. The result can be represented schematically as a structure of pairs each member of which branches off into a more similar pair.[33]

The point of representing the network's store of information in terms of such a branching structure is to show to what degree the network categorizes inputs as similar. In the case of Churchland's rock/mine detector,

---

[33] An excellent illustration of this branching structure for NETtalk can be found in Paul Churchland, A Neurocomputational Perspective (Cambridge, MA: MIT Press, 1989), p. 176.

the cluster analysis would be quite stark indeed. There would be slight differences in hidden layer vectors corresonding to various rock inputs or to various mine inputs. However, the rock inputs and the mine inputs would cluster around opposite ends of the branching structure. This is to say that any hidden layer vector for a mine input would be significantly more similar to the hidden layer vector for any other mine input than it would be to the hidden layer vector of any rock input. The cluster analysis would show that the network, as the result of being trained on rock and mine samples, has adjusted its weights so as to produce quite different hidden layer vectors for the two sorts of samples. In the case of NETtalk, on one end of the branching structure, highly similar inputs are paired. For example, 's' is treated as highly similar to 'z', 'e' as similar to 'y', 'W' as similar to 'w', and so on. As one moves to the most abstract level, the level at which only one pair remains, one finds a simple division of inputs into vowels and consonants. Of course, phoneme recognition and rock/mine discrimination are not the sorts of tasks with which Ramsey et al. are concerned. They are concerned with networks which store propositional information. However, there is no reason to believe that networks storing propositional information would not also be susceptible to cluster analyses. Simply changing the diet of samples upon which a network is trained is not enough to prevent a

cluster analysis.

Clark's point is that two networks which embody different weightings can still have the same cluster analysis. They will submit to the same analysis if, despite having different weights, they have settled upon some weighting or other which recognizes the same degrees of similarity between inputs. Hence, even in connectionist theory, there is a taxonomization according to which two networks which have encoded the same propositional information belong to the same kind. Connectionism does not invalidate folk psychological classification, or, at least, this eliminativist argument has not shown it to do so.

The upshot is that the argument of Ramsey et al. only works if one assumes that the only genuinely explanatory level of description within connectionist theory is its lowest level of description. Connectionist researchers do have the option of describing networks either at the lower units-and-weights level or at the higher cluster-analytic level. To reach the eliminativist conclusion, however, one must presuppose the low-level preference and thus dismiss the level of cluster analysis as not reflecting any genuinely explanatory taxonomization.

Earlier, I promised to discuss the very assumption made by the eliminativists that connectionist models should be construed as cognitive models. This discussion will illuminate one other aspect of the eliminativist's tendency

to discount higher-level predicates. The alternative is to construe connectionist models, assuming that they are approximately accurate representations of neural structure, as models of implementation. That is to say, connectionist models could be construed as corresponding to a lower level of description than the cognitive, a level on which a true cognitive model would supervene. The fact that eliminativists consistently fail to take this seriously as a possibility can only reflect a scepticism as to the scientific validity of higher-level properties. By way of illustrating this point, assume, as functionalist metaphysics requires, that any psychological property is a computational property.[34] As such, it is physically possible for any psychological property to be realized in a nonconnectionist machine of the right sort. For recall that according to the Church-Turing thesis, any computation can be performed by some Turing machine or other. Moreover, it is physically possible to build a univeral Turing machine, i.e., a Turing machine which can perform computations which are performable by all Turing machines.[35] Hence, given functionalist metaphysics, any psychological property is realizable by a universal Turing machine. Further, a

---

[34] A computational property is a second-order property of a specified type, viz. in which inputs and outputs are included in the property's causal profile.

[35] In fact, contemporary computers are universal Turing machines, or, at least, they would be if they had an indefinitely enlargeable memory capacity.

universal Turing machine need not be composed of connectionist networks. A variety of bizarre, nonneural contraptions can perform the state-to-state transformations definitive of a Turing machine. The conventional digital computer is one of them. Therefore, given functionalist metaphysics and a willingness to embrace higher-level properties as scientifically respectable, the most general psychological theory should be one including connectionist as well as nonconnectionist creatures within its purview. The cognitive models of this theory would not be parochial and, hence, would not be connectionist. The point I am trying to make is that presupposing that connectionist models are cognitive models also reflects a prejudice against admitting higher-level predicates as scientifically valid or nomic. Hence, eliminativist arguments from connectionism not only implicitly reject higher-level descriptions in their failure to take cluster analysis seriously but also in their initial assumption that cognitive models must be connectionist.[36]

3.4. Eliminativist Arguments from the Principle of Autonomy

---

[36] To put the matter somewhat more precisely, there are two levels of abstraction or greater generality which they reject. One is the level of cluster analysis. The other is the even more abstract level of cognitive modelling construed as expressing relations between computational properties. Instead, eliminativists opt for the units-and-weights level of description which is more fine-grained than either of the other two levels.

Stich's Autonomy Argument for eliminativism is meant to show that appealing to intentional properties in the explanation of behavior violates an allegedly plausible constraint on cognitive science. That constraint is what Stich calls the "Principle of Autonomy":

Autonomy Principle: All properties suitable for the laws of cognitive science are such that if any organism has one such property then all of its physically possible duplicates would also each have that property.[37]

(An organism's duplicate is an organism which is indistinguishable from it in terms of intrinsic microphysical properties at a given point in time. The environments and histories of the duplicates can vary to any degree whatsoever as long as that difference does not produce any difference in their intrinsic constitutions at the moment.) Stich claims that all intentional properties violate the Autonomy Principle. That is to say, no intentional property supervenes just on properties intrinsic to an agent but instead on a combination of intrinsic and environmentally or historically constituted properties. Stich concludes that intentional properties are of no concern to cognitive science. Instead, so Stich recommends, psychologists should develop purely syntactic theories of

---

[37] From Folk Psychology to Cognitive Science, p. 164.

mental representations without positing semantic interpretations for those representations.

I am not concerned with the alternative Stich endorses. For present purposes, it could just as well be a neuroscientific or biological theory as one concerned with syntactic properties. The pertinent point is that any alternative to folk or commonsense psychology must be asemantic, free of intentional predicates. That is the upshot of the Autonomy Argument with which I am presently concerned.

The thesis of the current section is that the Autonomy Principle, and hence the Autonomy Argument as well, derive whatever plausibility they have from the low-level preference. Hence, the Autonomy Argument is the third and final instance of an eliminativist argument which implicitly presupposes that a taxonomy should be rejected as nonexplanatory whenever a finer-grained taxonomy is an option. That the Autonomy Principle rests on such an assumption about explanation may not be initially obvious, but I trust that it will become so in the ensuing discussion.

For the sake of simplifying and focusing that discussion, I am willing to grant to Stich as many assumptions as possible, virtually everything except the Autonomy Principle itself and a nonpragmatic view of explanation. Specifically, I grant Stich the claim that all

intentional properties are broad, i.e., that all such properties fail to supervene on intrinsic properties alone but supervene instead on a combination of intrinsic and historical or environmental properties. In the cases of knowing that p (where p is a state of affairs external to the organism's present, internal constitution) and remembering that q, the intentional properties in question are uncontroversially seen as broad. For the having of either state depends partly upon the occurrence of some state of affairs either in the external world or in the past. In the cases of believing that p and desiring that q, however, the matter is controversial. But I will grant Stich everything, viz. that all intentional properties are broad and, hence, that appealing to any intentional property violates the Autonomy Principle.

Supposedly, a putative explanation of behavior using laws or generalizations violating the Autonomy Principle would either not be a genuine explanation of the behavior at all or would be inferior to at least one account which does not violate the principle. One important point here is redundancy: the account couched in terms of (broad) intentional properties is, according to Stich, superfluous, doing no more explanatory work than what is already done by some other account. Now, to say without further elaboration that two accounts are redundant is not necessarily to say that one is worse than the other; and even if it is to say

that, it is not to say that the intentional account must be the worse one. I will return to the issue of better and worse at a later point. For now I will address the matter of redundancy.

Stich considers (broad) intentional explanation to be dispensable at least in virtue of being redundant. Supposedly, any behavioral explanandum which an intentional explanans purports to explain can be explained by a strictly nonintentional explanans. The considerations Stich raises to support this point are of the following sort: Imagine that I am exhibiting a certain sort of behavior at a time, e.g., I am scratching my head. One can easily imagine the folk psychologist proposing an intentional explanation for my doing so, such as, that I desire that my head stop itching and believe that scratching it will achieve this end. Now it has already been granted that any intentional property partly supervenes on historical or environmental factors, so change those factors in the right way and the assignment of intentional properties must change. In this case, there must be counterfactual situations in which my intrinsic constitution would be the same but in which my history or environment would be different such that I would not have the same desire, viz. that my head stop itching, or the same belief, viz. that scratching my head will make it stop itching. Equivalently, not all of my possible duplicates would manifest these same intentional properties.

Let us grant that this is so. The eliminativist suggests, nonetheless, that the behavior of all my duplicates is identical. Viz., each of us scratches his head. Given that all of us exhibit the same behavior, then, according to Mill's method of agreement, there must be one causal explanation of the behavior holding for all of us. However, given that we do not all share the same intentional properties, that common explanation must be a nonintentional one. In other words, for any intentional property, if I have that property, I have at least one possible duplicate who (if actual) would lack it. Hence, the causal explanation which pertains to all of our cases must be nonintentional. Given that there is a nonintentional explanation of my scratching my head, so the eliminativist case continues, then any intentional explanation must be redundant.

Stich is well aware, however, that, as it stands, this argument is flawed. For there is a class of actions, <u>broad actions</u>, which an organism does not share with all of its possible duplicates. To put the matter more precisely, there are descriptions of at least some instances of behavior which an organism can satisfy even though some of its duplicates would not. For example, I reach for a glass of San Pellegrino while my duplicate on Twin Earth reaches, not for a glass of San Pellegrino, but for a glass of twin-San Pellegrino. If behavioral explananda can vary from

duplicate to duplicate, then perhaps intentional explanations are not redundant after all.

As noted, Stich is well aware of this objection. He suggests, however, that any intentional explanation of the broad behavioral explanandum really would be redundant. This is due to the fact that the broad explanandum is itself a logical consequence of two or more distinct explananda which can each be given a nonintentional explanation. One such explanandum is simply the behavior itself under a narrow description, i.e., a description which the organism and all of its duplicates would satisfy. In the case of my reaching for a glass of San Pellegrino, it would be a description of my movement the satisfaction of which in no way depends upon historical or environmental facts. The glass and the San Pellegrino would nowhere be mentioned. Behavior under such a narrow description can be given a nonintentional explanation as has already been established. The other explanandum would contain reference to those environmental or historical facts which are responsible for the narrow behavior satisfying the broad description. It is here that reference to glass and San Pellegrino would be found. Presumably, this second explanandum also contains information about the organism itself, viz. that it is ensconced in the environment in question, and that it has a certain sort of history. What this second explanandum does not contain is a description of the behavior itself. But

both explananda in conjunction entail the broad explanandum.

As already established, the narrow behavioral explanandum can be given a nonintentional explanation, and a nonintentional explanation can also be given of the second explanandum since it falls outside the domain of psychology. The conjunction of these nonintentional explanantia entails the broad explanandum, and so any intentional explanation of it turns out to be redundant after all.

In Stich's words,[38]

> [Broad] descriptions of behavioral events should be viewed as conceptually complex, resolvable into [a narrow] component and a potpourri of other factors which explain why the [narrowly] described event counts as satisfying the [broad] description.... [The external explanantia] may include the history of the individual in question, the history of the terms he uses, the linguistic, social, legal, and ritual practices that obtain in the society of which he is a part, and perhaps many other factors as well. [Therefore,] it is plausible to conclude that the descriptions of behavior that a psychological theory should use in its explananda will be [narrow] descriptions.

The following schema should help clarify why the eliminativist takes intentional explanations to be redundant:

(i) Nonintentional features possessed by all possible duplicates of the organism O.

---

[38] Ibid., p. 169.

(ii) Facts about Q's environment or history.

(iii) Q's behavior narrowly described.

(iv) Further facts about Q's environment or history.

(v) Q's behavior broadly described.

For example, (i) might consist in purely syntactic or neurophysiological information about Smith; (ii) in a nonintentional explanation of why there is a glass of water at space-time locale x,y,z,t; (iii) in Smith's reaching behavior narrowly described; (iv) in the information that a glass of water is at x,y,z,t; and (v) in a description of the fact that Smith reaches for a glass of water.

According to the Autonomy Argument; (i) explains (iii), (ii) explains (iv), and the conjunction of (iii) and (iv) entails (v). Therefore, (i) and (ii) also imply (v) and presumably explain it. Since we may assume that (ii), just like (i), is nonintentional, there is a nonintentional explanation of (v). The conclusion is that any intentional explanation of behavior, even when it is in broad terms, is redundant.

I am willing to accept this conclusion, at least for purposes of discussion. That is to say, I shall accept that for any intentional explanation of any instance of behavior

broadly described, there is also a nonintentional explanation of it. And I grant this on the basis of the reason given above. Now that we have it that any broad behavioral explanandum has two explanantia, one intentional and one nonintentional, we must ask why only the latter is thought to be genuinely explanatory. On the face of it, one could simply grant that there are two explanantia and leave it at that without bringing up the prospect of elimination. The eliminativist has yet to show that there is something superior about the nonintentional explanation. In what follows, I suggest the only plausible reason why one would take one explanation to be superior to the other. What is of utmost importance, as it will turn out, is that the nonintentional explanations in question are necessarily more fine-grained than their competing intentional explanations. Hence, the rejection of intentional explanations in favor of nonintentional ones reflects a preference for lower-level explanations over more general ones whenever there is such an option.

Jackson and Pettit have noted that the appeal to the broad properties of an object's behavior can buy one an increase in generality over the exclusive appeal to narrow properties.[39] The eliminativist, I contend, sees this increase in generality as a sign of explanatory inferiority.

---

[39] Jackson and Pettit, "Functionalism and Broad Content", Mind 97 (1988), 371-91.

Recall that explananda (iii) and (iv) must jointly entail (v). Otherwise, the eliminativist has not shown the intentional explanation to be redundant. My claim is that any plausible candidates for (iii) and (iv) will be more fine-grained than (v) would need to be. A (broadly) intentional explanation of (v) would abstract away from many of the finer-grained details found in (iii) and (iv). Hence, the eliminativist, by favoring the nonintentional explanations (i) and (iii), and (ii) and (iv) over the intentional explanation of (v) is favoring a finer-grained explanation over a wider-grained explanation.

By moving from the broad behavioral explanandum (v) to the explananda (iii) and (iv), one is moving from a relatively wide-grained description of a fact to more fine-grained descriptions. This is due to the fact that (v) contains information as to the relation of the behavior to its environment or history. When this relational information is disallowed from either of the two explananda, (iii) and (iv), this forces one to put more specific information into (iii) and (iv). Otherwise, the conjunction of (iii) and (iv) would not entail (v).

Here is an analogous example from outside the domain of psychology. It is similar to an example used by Jackson and Pettit to illustrate the virtues of higher-level explanations.[40] The thing to be explained is that electron

---

[40] Ibid., pp. 392-3.

A moves at the same velocity as another electron one micron distant from it. An apparently good explanation is that each electron is acted upon by a force of the same magnitude as that acting upon the other. However, in Stichian fashion, one could argue that this explanation is dispensable. For consider the following schema:

(i) Electron A is acted upon by a force of magnitude N.

(ii) One micron distant from A is an electron acted upon by a force of magnitude N.

(iii) A moves at velocity V.

(iv) One micron distant from A is an electron moving at velocity V.

(v) A moves at a velocity equal to that of an electron one micron distant from A.

Now, originally I proposed that (v) be explained by the following explanans: that there is an electron one micron distant from A and both electrons are acted upon by a force of the same magnitude. However, that is now supposedly shown to be redundant; for (v) is entailed by explananda (iii) and (iv), when taken in conjunction, which are in turn

explained by (i) and (ii). Hence, (i) and (ii) jointly explain (v) without our needing to appeal to the other explanans.

Notice that in all relevant respects, this case is analogous to Stich's argument against intentional psychological explanation. (v) is a broad explanandum. That is to say, it denotes a property which A would not share with all of its duplicates. However, (v) is a logical consequence of a narrow description of A's behavior, viz. (iii), in conjunction with information as to why A's satisfying (iii) implies (v), viz. (iv). Hence, the original explanation of (v) can be construed as redundant if we can explain (iii) and (iv) without appealing to that information. And indeed we can; for (i) explains (iii), and (ii) explains (iv).

The moral is that a broad explanandum contains relational information, and so in explaining a broad explanandum in relational terms, one is abstracting away from many possible alternatives which could realize the relation in question. However, if one replaces the broad explanandum with distinct explananda which entail the original broad explanandum, one is forced to spell out just which properties entail the relation. The relation itself could be held constant across a range of counterfactual situations. However, if one can no longer appeal to that relation, one must spell out the specifics in the actual

world which imply that the relation obtains.

Here is another example: Let us suppose that the explanandum of interest is

(v) Compass C is pointing north.

Note that this is a broad explanandum. A duplicate of C need not be pointing north. So, in Stichian fashion, one could replace (v) with a narrow description of its behavior, viz.

(iii) C's needle points toward 'N' on C's dial,

and a description of why (iii) counts as an instance of (v), viz.

(iv) The 'N' on C's dial points north.

But it is now clear that the mere appeal to the relational property cited in (v) abstracts from the details found in (iii) and (iv). (v) does not specify that C's dial is inscribed with an 'N' or that the 'N' points north, etc. The specific details cited in (iii) and (iv) could have been different, of course; but the inclusion of such additional information of some sort or other is unavoidable. If one takes away the citation of a relational property while

retaining information which entails that the relational property is instantiated, the information retained must be such as to the specific features of the situation which realize the relational property. Hence, one is moving from the more general and less specific to the less general and more specific.

To move from (broad) intentional explanation to nonintentional explanation is to make that same move. This can be illustrated by turning to an example of human behavior. Consider the following scenario: Smith is standing in an open field on a clear night facing due north. The moon is visible on the horizon. Smith desires to look at the moon and so turns in order to face it. Assume that the moon never appears on the horizon due north or due south but always either in the eastern sky or the western sky. Now, if Smith is at all normal, he will turn right to look at the moon if it is in the eastern sky and left if it is in the western sky.

I wish to consider the following possible explanandum:

(1) Smith turns toward the moon.

The predicate 'turns toward the moon' is a broad description of Smith's behavior, since not all of Smith's duplicates would satisfy it. Some would turn moonward, but others would turn twin-moonward. Even others would turn toward an

empty sky while yet having a moonlike visual experience.

Folk psychologists would profess to be able to explain (1) by appealing to an explanans which at least contains the following claim:

(2) Smith desires to look at the moon.

Now, Stich would claim that an explanans of the sort containing (2) is dispensable by virtue of the fact that (1) can be replaced with two distinct explananda, each of which is susceptible to an explanation which, unlike (2), is nonintentional. But what are the two explananda which could replace (1)? Clearly, they could not be the following:

(1a) Smith turns at time t.

(1b) The moon is visible on the horizon at t, and [appropriate information pertaining to Smith's spatiotemporal location at t but not including the information in (1a)].

Why not? Because the conjunction of (1a) and (1b) would fail to imply (1). (The bracketed information in (1b) does not change this.) For (1), by containing the word 'toward', implies the existence of some type of correlation between the location of the moon in the sky and the direction in

which Smith turns. The conjunction of (1a) and (1b) does not. That conjunction is compatible with Smith turning his back on the moon. (1), however, is not. The eliminativist might attempt to make up for this gap by suggesting that (1) be trifurcated so as to produce a third explanandum. Such a third explanandum would be something to the effect that the direction of Smith's turning is correlated with the location of the moon in the sky. But this, of course, would defeat the eliminativist's purpose, for this third explanandum would be a broad description of Smith's behavior, the very thing the eliminativist wishes to avoid.

I take it as established that the eliminativist cannot plausibly be advocating the bifurcation of (1) into (1a) and (1b). But what can the eliminativist suggest as a replacement of (1)? How can (1) be broken down into a narrow behavioral description and a "potpourri of other factors" which jointly imply (1)? Plausibly, this can only be done by replacing (1a) and (1b) with explananda that are more specific, which describe the phenomena to be explained in finer-grained detail. One such possibility is the following pair of explananda:

(1a') Smith turns left at t.

(1b') The moon is in the western sky at t, and [appropriate information pertaining to Smith's spatiotemporal location at

t but not including the information that Smith turns].

Another possible pair of explananda is

(1a'') Smith turns right at t.

(1b'') The moon is in the eastern sky at t, and [appropriate information pertaining to Smith's spatiotemporal location at t but not including the information that Smith turns].

Either pair would imply (1). More specifically, either pair does what the conjunction of (1a) and (1b) cannot do, viz. rule out Smith turning his back on the moon.

Since a broad description of behavior contains information as to the coordination of the behavior to the environment, the replacement of that description with a narrow description loses the information as to the coordination. This is not a problem for the eliminativist as long as the conjunction of both the narrow description and the collateral information implies the broad description. However, the new descriptions can only conjointly imply the broad description provided that they are more fine-grained than the broad description. They must be more fine-grained so that they conjointly imply the coordination without implying it when taken singly.

The relation between (1), on one hand, and either the conjunction of (1a') and (1b') or the conjunction of (1a'') and (1b''), on the other, is that of a multiply realizable property to its possible realizing properties. (1) captures something which remains constant for both [(1a') & (1b')] and [(1a'') & (1b'')], viz. the fact that Smith ends up facing the moon. In order to capture this fact while rejecting the broad behavioral description, the eliminativist is forced to move to a lower level of description. Hence, the folk psychologist in accepting (1) as a valid explanandum, is abstracting away from what distinguishes [(1a') & (1b')] from [(1a'') & (1b'')]. Explanations in terms of (broad) content highlight common features which are realizable by a plurality of different physical properties. To reject intentional explanations for taking broad descriptions of behavior as their explananda, as Stich recommends, is to reject the more general and multiply realizable in favor of the lower-level realizing properties.

Jackson and Pettit take the rejection of (broad) intentional explanation as a rejection of higher-level explanations, and I believe that the above considerations illustrate that they are correct. According to Jackson and Pettit:[41]

---

[41] Ibid., pp. 398-9.

> For a given broad content B there will be a number of ways of realizing that content by the appropriate combination of [narrow property-instance] N and environment E, say: N_1 & E_1, N_2 & E_2,.... One of these ways, say Na and Ea, will be the actual way B is realized. Now each of the N_i & E_i will explain and predict different behaviour in the subject, but it may be that there is a common thread T running through these different pieces of behaviour. In this case ascribing B explains [] T just as well as ascribing Na and Ea, and does something distinctive besides - it tells us that it did not matter as far as getting T goes that it was Na & Ea that was actual instead of, say, N_2 & E_2.

The "common thread T" is a property which stands in a one-many relation to the various possible (N_i & E_i). It concerns some relation between the behavior and the environment, e.g., that Smith turns <u>toward</u> the moon, that the compass points <u>toward</u> the north, etc. T remains constant across various possible (N_i & E_i) in that, for example, Smith would end up facing the moon no matter where it happened to be located, the compass would point north whether or not its dial is inscribed with an 'N'. Hence, the (broad) intentional explanation is pitched at a higher level than the nonintentional explanation groomed to be its replacement.

This would appear to provide the only explanation of why eliminativists prefer the nonintentional explanation of the broad behavioral explanandum to the intentional one. Recall that the possibility of devising an explanation according to the schema (i) through (v) shows that there is a nonintentional explanation for any broad behavioral

explanandum. However, prima facie, this is nothing more than to show that any broad behavioral explanandum is explanatorily overdetermined, having a nonintentional as well as an intentional explanation. What was needed in order to complete the eliminativist case was some necessary feature of any such nonintentional explanation distinguishing it from the intentional one and also showing it to be superior to the latter. As shown above in the examples involving the electron, the compass, and Smith's turning toward the moon; one thing which distinguishes any such nonintentional explanans from its intentional rival is the greater fineness of grain of the former. In replacing the appeal to a broad intentional property with appeal to intrinsic nonintentional properties, one has no choice but to move to a level of greater descriptive specificity. Evidently, it is this greater specificity which seems to recommend the nonintentional explanans over the intentional one.

Recapitulation: Given the assumption that all intentional properties are broad, there is no intentional property which an organism would share with all its possible duplicates at a given point in time. However, any organism would exhibit the same narrow behavior as any of its duplicates. Hence, there must be some nonintentional explanation of the narrow behavior. According to the eliminativist, this would render any intentional explanation

of the narrow behavior redundant. Intentional explanations of behavior broadly described are also shown to be redundant when one considers that any broad description of behavior is implied by (a) the corresponding narrow description, and (b) facts which show (a) to satisfy the broad description. Since both (a) and (b) can be given nonintentional explanations, then the behavior under its broad description can also be explained nonintentionally. Therefore, any behavioral explanandum admitting of an intentional explanans also admits of a nonintentional one. Stich claims that only the nonintentional explanation is genuinely explanatory, that the intentional one is counterfeit. This causes one to inquire into what distinguishes the two such that one is so clearly inferior to the other. I have claimed that the relevant difference here is that the nonintentional explanation is necessarily more specific and, hence, lower in level than the intentional. Hence, the eliminativist is rejecting intentional psychology in virtue of there being more fine-grained explanations to which one can appeal. The Autonomy Argument fits into the same pattern as do the Poorness-of-Fit and the Connectionist Arguments. In each case, there is a crucial appeal to the low-level preference.

3.5. Conclusion

Wherein lies the appeal of the low-level preference? The social scientist Jon Elster attempts to defend the

preference in the following terms:[42]

> The search for microfoundations, to use a fashionable term from recent controversies in economics, is in reality a pervasive and omnipresent feature of science. It corresponds to William Blake's insistence that 'Art and science cannot exist but in minutely organised Particulars.' To explain is to provide a mechanism, to open up the black box and show the nuts and bolts, the cogs and wheels of the internal machinery.

Elster's point, evidently, is that low-level accounts provide more specific information as to the explanandum's causal history than do higher-level accounts. This explains Elster's reference to "the cogs and wheels of the internal machinery."

A metaphysical justification of Elster's view would go as follows: higher-level causally relevant properties strongly supervene upon lower-level properties. That is to say, lower-level causal relations entirely determine higher-level causal relations. Moreover, given multiple realizability, higher-level properties stand in one-many relations to lower-level properties. So the appeal to the lower-level properties is more specific as regards causal information than is the appeal to higher-level properties. Therefore, a lower-level description provides more detail as to the actual causal history of a given explanandum than would a higher-level description. Presumably, scientific

---

[42] Jon Elster, Explaining Technical Change (Cambridge: Cambridge University Press, 1983), pp. 24-5.

explanations of behavior and cognition should be causal. So it is not difficult to see the appeal of the low-level preference. That appeal consists in the view that a lower-level description always provides a more informative account of the causal history of any property-instance than would any higher-level description.

So the eliminativist's way of thinking may be understood in the following terms: Nondisjunctive physical predicates provide maximally specific information in describing the explanandum's causal history. Any predicates providing less specific information would not count as nondisjunctive physical predicates. Hence, the standards for physical explanatory efficacy are maximal vis-à-vis specificity of information concerning causal history. Now, if one assumes that the standards of explanatory efficacy remain constant across all contexts, then one will also believe that if any account fails to meet standards of physical explanatory efficacy, then it is either not an explanation at all or is at least inferior to some physical explanation. By making this nonpragmatic assumption, the eliminativists conclude that lower-level accounts are always to be preferred over higher-level accounts of an explanandum.

That eliminativists prefer lower-level accounts because of their greater specificity vis-à-vis causal relations is discernable in the literature. For example, in his own

defense of the Connectionist Argument,[43] Paul Churchland rejects the cluster-analytic level of description in favor of the weights-and-units level on the grounds that

> the learning algorithm that drives the system [] does not care about [the cluster-analytic level]. All it cares about are the individual weights and how they relate to apprehended error. The laws of cognitive evolution, therefore, do not operate [at the cluster-analytic level. That level] certainly corresponds more closely to the "conceptual" level..., but the point is that this seems not to be the most important dynamical level.

It is also discernable in discussions of the appeal of Stich's Autonomy Principle. According to McGinn, the principle's persuasive charm consists in the recognition of the spatiotemporal contiguity of the causal relation. Accordingly, an explanation of a property-instance appealing to events that are spatially or temporally removed would provide less specific information as to the explanandum's causal antecedents and supposedly thus be inferior.[44]

What I especially wish to emphasize is that the premises of the eliminativist argument and Kim's argument for local reductionism are essentially the same and so are subject to the same refutation. In both, it is assumed that

---

[43] "On the Nature of Theories: A Neurocomputational Perspective", pp. 177-8.
Note Churchland's consistent use of causal terminology, such as 'drives', 'cares about', and 'dynamical' to show what recommends the lower level of description.

[44] Op. cit., pp. 132-9.

the maximal standards of specificity associated with physical explanation pertain to all explanation. (Kim couches this point in terms of the homogeneity of the causal powers of physical properties, but it is evidently the same point.) From this it is inferred that anything less specific than a physical explanation is either nonexplanatory or at least explanatorily inferior to some physical explanation. This would, of course, disqualify folk psychological accounts from being explanatorily adequate on the grounds that their posited properties, in virtue of being multiply realizable, are too causally inhomogeneous to meet the maximal physical requirements of specificity.

Moreover, not only do the arguments share the same premises, they share the same conclusion as well, for <u>eliminativism just is local reductionism</u>. Both conclude that lower-level or physical properties are the only scientifically explanatory kinds and hence that multiply realizable properties, such as folk psychological properties, should be banned from science. The only possible distinction between their conclusions is that Kim seems to think that it will still be appropriate to use such terms as 'pain-in-a-human', 'joy-in-an-Alpha-Centaurian', etc. to refer to those physical properties which explain behavior, while the eliminativists would prefer that such terms as 'pain' and 'joy' be discarded. The difference is

purely terminological and not at all significant. What matters is not terminology but taxonomization, and in that regard the positions are indistinguishable. For both, the older folk taxonomization is rejected in favor of one that is finer in grain by reason of being physical.

Given the identity of the eliminativist and local reductionist arguments, the same pragmatic view of explanation used to counter the latter in Chapter Two also applies to the former. Which level of description is suitable in explaining a given explanandum depends upon local interests. If, for example, one's greater interest is in understanding how the explanandum fits into general patterns of nature, one is more concerned with properties standing in a one-many relation to physical properties, and so one appeals to a higher-level description. However, if one works in a science in which a higher premium is placed on predictive precision, one chooses predicates which are finer in grain, and hence one appeals to a lower level of description. In short, the choice between opting for greater predictive precision and finding more general patterns in nature determines the level of description at which an explanation is to be couched. On this view, the low-level preference is indefensible, and so the argument for eliminativism is unsound.

However, one might object that it is not very interesting to argue from such a pragmatic view of

explanation against eliminative materialism. For in assuming the interest-relativity of explanation, one is assuming the existence of interests thus begging the question against eliminativism from the very start.[45] If one finds the pragmatic view of explanation so plausible in the first place, one might ask, then why go on to argue against eliminativism? Haven't you already presupposed its falsehood?

There are at least three reasons for believing that adopting a pragmatic view of explanation does not presuppose the falsehood of eliminativism. One such response is simply to point out, perhaps somewhat pedantically, that one can consistently maintain that explanation is interest-relative while also insisting that no one has any interests. What this means, of course, is that there have never been any explanations. The predicate 'is an explanation' would be characterized in terms of interests, but, since no one has ever had any interests, one is forced to conclude that the predicate has never been satisfied.

I hasten to point out, however, that I find this response extremely unappealing. I only include it for the sake of comprehensiveness. Part of what makes it so unappealing is the love of science which so clearly sets the tone for much of the eliminativist literature. There is, as

---

[45] Prof. Arthur Danto pointed out this possible objection to me.

I have already attempted to make evident, an implicit physics worship in eliminativist thought as manifested in the low-level preference. But this physics worship must presuppose, at a minimum, that there are physical explanations, that physical theories do indeed explain much of the phenomena we see around us. What perhaps contributes even more to this being an unattractive response is the fact that it defeats the purpose of arguing against eliminativism. That purpose is to defend the integrity of psychology as a science. This would be lost were one to maintain that there are no explanations. Hence, even though the sceptical denial of the existence of any explanation shows the pragmatic view not to beg the question against eliminativism, it is, all the same, not an attractive response.

A better response involves making a broad distinction between two different sorts of eliminative materialism. According to one sort, nothing possesses any mental properties. According to the other, mental properties are not fit for scientific explanation but may indeed be instantiated. On this latter view, the predicate 'believes that it is raining' would be like the predicate 'is a bust of Socrates', insofar as it is satisfied by some particulars but plays no role in scientific explanation. Even though the pragmatic view of explanation might beg the question against the former type of eliminativism, it need not do so

against the latter. Indeed, in the course of discussing the issue of eliminativism, I have not addressed the question of whether or not anyone feels pain or believes anything. My concern, rather, has been focused on whether or not these properties have explanatory power. Presupposing that people really do have interests does not beg the question against this type of eliminative materialism.

A third response is, I believe, the most interesting. It involves considering what is truly essential to any pragmatic view of explanation. What is essential to the pragmatic view is that no sentence of the form 'Statement (or theory) S explains phenomenon x' be taken at face value. Any such claim must be understood as containing an implicit reference to a possible audience. Hence, any such sentence must be understood as elliptical for one of the form 'If S explains phenomenon x relative to audience A, then A has property P'. 'P' is usually understood as denoting intentional properties such as A having certain interests. Indeed, the pragmatic view is most obviously attractive if one understands 'P' in this way. However, it is not essential to the audience-relativity of explanation that these properties be intentional. Why couldn't they be neurophysiological or physical properties of A? I believe that an eliminativist might actually find some view of this form attractive. Hence, presupposing the audience-relativity of explanation does not presuppose the falsehood

of eliminative materialism. If this nonintentional pragmatic view of explanation were adopted, the previous discussion of the interest-relativity of explanation would have to be understood in some new way. But the resulting view would be audience-relative just the same.

That explanation is audience-relative will be defended in the following chapter.

CHAPTER FOUR

THE PRAGMATICS OF EXPLANATION

In the preceding chapters, I have shown that the case for psychophysical reductionism rests upon certain nonpragmatic assumptions about explanation. In this chapter, I clarify the distinction between pragmatic and nonpragmatic views of explanation and also argue that it is plausible that there really are just such pragmatic elements to explanation. Toward this end, it is not necessary for me to spell out a complete theory of explanation. That is to say, I am not propounding a fully articulated theory of explanation but am instead arguing for there being certain necessary constraints on explanation. Revealing these necessary conditions is sufficient to undermine the case for psychophysical reductionism.

4.1. Why Nagelian Psychophysical Reductionism Is Inconsistent with a Plausible Constraint on Explanatory Efficacy

In the sciences, one finds many sentences appearing to be of the form

(1) **E** explains **X**,

in which **E** and **X** are each taken to be some set of sentences

or terms referring to phenomena. Roughly speaking, X is the thing to be explained, and E is what has explanatory power in relation to X. Examples of type-(1) sentences include 'Laws of statistical mechanics explain thermodynamical laws', and 'Relativity theory explains the precession of the perihelion of Mercury'.

According to any pragmatic conception of explanation, as I choose to characterize the notion, any type-(1) sentence is elliptical for some sentence of the following form:

(2) E explains X to A when cited by B.

In this type of sentence, B is a person or group of persons at some moment in time or some extended period of time who are potential explainers. A is a person or group of persons at some moment in time or extended period of time who are potentially an audience. A and B may also each be construed as a type or kind of audience or explainer respectively. A pragmatic view of explanation, then, is an audience-relativized view of explanation. On any such view, prima facie categorical claims (i.e., (1)-sentences) must be understood as containing veiled reference to possible audiences.[1] On such a view, E cannot be spoken of as

---

[1] It is also an explainer-relativized notion of explanation, but only audience-relativization interests us in the current discussion.

having explanatory power in relation to x simpliciter. Instead, its explanatory power vis-ā-vis the explanandum at least partially depends upon features of A, i.e., contextual factors.

The appeal of a pragmatic approach to explanation consists in its showing there to be a close link between (1)-sentences and sentences of the following form:

(3) B explained x to A by citing E.

Claims of this form refer to explaining episodes, particular acts of explaining something to someone. As Robert Matthews has noted, intuitively, (1)-claims and (3)-claims are closely linked.[2] The claim that, say, Newtonian mechanics explains the tides must have something to do with the claim that Jones explained the tides to Smith by citing Newtonian mechanical laws. The appeal of any pragmatic view of explanation is that it claims to show in a quite straightforward way how they are linked. Given that a pragmatic view simply states that every (1)-sentence just is a disguised (2)-claim, one can see that (1)-sentences and (3)-sentences are closely related. Since (2) is the claim that it is possible for (3) to be true, on a pragmatic view, the ostensible claim that E explains x just is the claim

---

[2] Robert J. Matthews, "Explaining and Explanation", American Philosophical Quarterly 18 (1981) 71-7.

that the corresponding (3)-sentence can possibly be true. For example, the claim that statistical mechanical laws explain thermodynamical laws must be understood as containing implicit relativization to an audience (A) and an explainer (B) such that it is possible for

(3′) B explained the laws of thermodynamics to A by citing laws of statistical mechanics,

to be true. By relativizing the explanatory efficacy of E vis-à-vis X to a given audience and explainer, a pragmatic approach posits a straightforward connection between (1)-claims and (3)-claims. Specifically, on a pragmatic view, an analysis of any (1)-claim is automatically an analysis of the corresponding (3)-claim.

In order for these points to be more vivid, let us consider concrete examples of these three types of claims.

(1′) Newtonian mechanics explains the tides.

(2′) Newtonian mechanics explains the tides to Smith when cited by Jones.

(3′) Jones explained the tides to Smith by citing Newtonian mechanics.

On a pragmatic view, to assert (1′) is to assert some contextually-relativized claim such as (2′). An analysis of (2′) automatically provides an analysis of (3′), since (2′) simply asserts the possibility of (3′) being true. Therefore, since a pragmatic view claims (1′) to be elliptical for, say, (2′), then such a view shows an analysis of (1′) automatically to carry with it an analysis of (3′). The appeal of a pragmatic take on explanation is that it provides a unified account of both (1)-claims and (3)-claims.

Contrast this with a nonpragmatic view of explanation such as Carl Hempel's deductive-nomological theory of explanation.[3] Hempel attempts to provide an analysis of (1)-claims which in no way appeals to possible audiences or explainers. He clearly does not consider (1)-claims to be disguised (2)-claims. As a result, his analysis of any (1)-claim does not provide a complete analysis of the corresponding (3)-claim. At best, it only provides a few necessary conditions for that (3)-claim. Hence, a Hempelian view of explanation requires two separate analyses of explanation: one pragmatic or context-relativized, the other nonpragmatic. If one finds this need for a dual analysis to be unattractive, as I do, then a pragmatic approach is

---

[3] Carl G. Hempel, Aspects of Scientific Explanation (New York: The Free Press, 1965). Hempel also suggested other models of explanation, but they are equally nonpragmatic.

recommended.

Matthews, however, has attempted a unified theory of explanation which is also nonpragmatic. That is to say, he has suggested a nonpragmatic analysis of (1)-claims which automatically carries with it a putative analysis of the corresponding (3)-claims. He has suggested that such an analysis can be arrived at first by devising an analysis of (2)-claims and then generalizing so that any reference to a particular explainer or particular audience is removed.[4] On this view, while any (2)-claim is about a specific audience and explainer, the corresponding (1)-claim makes the same claim about any possible audience or explainer. Hence, on such an approach, one must first find necessary and sufficient conditions for

(2) E explains x to A when cited by B.

One then goes on to say that

(1) E explains x,

is true just in case those conditions would be satisfied for any explainer who cites E to any audience. Let this attempt to show that (1)-claims are not disguised (2)-claims be called the "generalization strategy."

---

[4] Op. cit.

For example, one begins by analyzing a claim such as 'Newtonian mechanics explains the tides to Smith when cited by Jones'. This analysis will appeal to certain qualities of Smith and Jones, certain features which they must possess in order for the sentence to be true. Very plausibly, one feature which Smith (the audience) must possess is the capacity to understand Newtonian mechanics. On Matthews' proposed view, the corresponding (1)-claim 'Newtonian mechanics explains the tides' is true only if all possible explainers and audiences share those same features. Hence, all possible audiences must have the capacity to understand Newtonian mechanics in order for it to be true that Newtonian mechanics explains the tides. Evidently, the generalization strategy has extremely counterintuitive consequences. For there are certainly possible audiences which lack the cognitive power to grasp Newtonian mechanics. This being so, given the generalization strategy, it is not true to say that Newtonian mechanics explains the tides!

What this means is that the generalization strategy implies the falsehood of all (1)-claims which certainly seems to be a reductio ad absurdum of the generalization strategy.[5] Since the generalization strategy to date is the only nonpragmatic approach for providing a unified

---

[5] This would render Nagelian psychophysical reductionism trivially false anyway. If all (1)-claims are false, then it is false to claim that any physical theory reductively explains psychological theory.

analysis of both (1)-claims and (3)-claims, nonpragmatic approaches to explanation in general suffer a great loss of plausibility as a result. I conclude that some pragmatic approach or other must be more plausible. Prima facie categorical claims, therefore, such as 'Newtonian mechanics explains projectile motion' should be understood as elliptical for such claims as 'Newtonian mechanics explains projectile motion to a typical college freshman when cited by a typical physics professor', etc.

Let us return to the topic of global psychophysical reductionism. I claim that the pragmatic element of explanation is a major stumbling block for the disjunction strategy. Given an audience-relativized view of explanation, a given (1)-claim is true only relative to certain audiences. I claim that any psychophysical bridge generalization utilizing the enormously disjunctive physical predicates is not genuinely explanatory relative to a human audience. Hence, since any law is necessarily explanatory, it is not genuinely a bridge <u>law</u> relative to any human audience. I here rely on the point made earlier that a (2)-claim cannot be true unless the audience has the capacity to understand **E**. Given a pragmatic approach, one can say the same for any (1)-claim. Hence, a pragmatic approach quite plausibly commits one to accepting a <u>cognitive constraint</u> on any (1)-claim, i.e., a given (1)-claim can only be true if the implicit audience is able to understand **E**.

Now consider an instance of (2), e.g., one in which the complete and true psychological theory takes the place of '**X**', the set of disjunctive physical generalizations to which Kim appeals in his disjunction strategy takes the place of '**E**', and a term referring to the (human!) scientific community takes the place of '**A**'. Let this instance of (2) be called '(2*)'. The upshot: multiple realizability entails that (2*) is false. That is to say, multiple realizability entails that psychophysical reductionism is false relative to a human audience.

The reasons for this conclusion are as follows: The cognitive powers of **A** place a constraint on sentences of form (2). More specifically, if **A** is not capable of grasping the meaning of **E**, then no one can explain anything to **A** by citing **E**. Moreover, given that any (1)-claim just is an abbreviated (2)-claim, the cognitive powers of **A** also place a constraint on sentences of form (1).

The further premises supporting my antireductionist conclusion were defended earlier, and so I will review them rather briskly: Each of the highly disjunctive predicates which enters into the generalizations taken by Kim to reduce psychology is highly complex in its semantic content. This, of course, follows from the enormous multiple realizability of any mental property. This semantic content, in fact, is so great that it cannot be humanly grasped. The disjunctivity of any such predicate would, at the very

least, exceed the limits of human memory. Therefore, no generalization containing such a predicate can be explanatory relative to a human audience and so cannot be reductively explanatory.

Now that my objection to the disjunction strategy has been fully stated, one of my earlier points, which may have sounded paradoxical at the time, can be more forcefully defended, viz. that type materialism is compatible with the denial of psychophysical reductionism. That is to say, (e.g.) pain can actually be identical to the disjunction of its physical realizers even though psychological theory fails to be reducible to physical theory. Type materialism is simply the metaphysical claim that two predicates, one psychological and the other physical, refer to the same property. Since type materialism concerns metaphysics only and not explanation per se, there is no cognitive constraint to be considered in evaluating whether it is true. More specifically, whether or not a given audience is capable of grasping the sense or meaning of the physical predicate is irrelevant to the truth of type materialism. What matters is the reference of the predicate.

By contrast, in evaluating the claim that psychological theory can be reductively explained in terms of physical theory, there is a cognitive constraint. It is not enough that for any psychological predicate there is a physical predicate which refers to the same property. In order for

the putative bridge law to play a genuine role in explaining psychological theory, the audience in question must be able to comprehend the sense of the pertinent physical predicate. Due to the cognitive constraint on explanation, the question of type materialism comes apart from that of psychophysical reductionism.

Given multiple realizability, the disjunction strategy would be the only hope for Nagelian psychophysical reductionism. However, the above considerations show that global psychophysical reductionism is false relative to a human audience.

4.2. Possible Objections

A proponent of Nagelian psychophysical reductionism might wish to raise objections to the argument presented in the preceding section. I shall anticipate and respond to some of the possible objections.

One might claim that I have not done enough to dispel the notion that there is a separate analysis for (1)-claims and (2)-claims. If indeed there is, one might suggest that global psychophysical reductionism is true given the nonpragmatic conception of explanation expressed in (1)-claims. One might go on to claim that reduction in this nonpragmatic sense is reduction enough. Or one might object that even if there is a single, unified pragmatic analysis of explanation, that global psychophysical reductionism is

true relative to a possible superhuman audience endowed with superior cognitive powers. One might suggest that reducibility relative to this super-audience is reducibility enough.

Indeed, either possibility cannot wholly be ruled out. Perhaps a reasonable conception of nonpragmatic explanatory efficacy will be devised such that psychophysical reductionism is true on that conception. Or perhaps psychophysical reductionism is true relative to some superaudience. Now these might be good objections if the reductionism one were defending were purely metaphysical. For example, if one were defending the type-materialist claim that pain simply is the disjunction of all of its physical basal conditions, then the claim that psychology is reducible from some purely objective or suprahuman perspective might be a relevant point. However, the sort of reductionism at issue here is not metaphysical but intertheoretical, so the point is irrelevant.[6] That is to say, what is at issue is not metaphysical relations between properties but relations between theories.

But could these objections be used to defend intertheoretic reductionism? Perhaps, but one would pay a price for doing so. For if they were so used, it would not

---

[6] Note that in Chapter Two, I actually assume that type materialism is true for the sake of simplifying the discussion. Questions as to property identity are not of any real concern in this dissertation which instead addresses the issue of explanation.

be the sort of intertheoretic reductionism which is typically espoused in the literature. That type of reductionism requires that psychology be physically explainable relative to a human audience. For psychophysical reductionists, in the intertheoretic sense, take the supposed truth of their view to have implications for scientific practice. For example, according to Paul Churchland, the "bottom-up" research methodology "gives the most direct expression to the philosophical [theme] advanced by the reductive [materialist]."[7] Now, on the bottom-up research methodology, one's strategy is to begin with a physical theory and then derive the true psychological theory from it. However, the physical theory from which one begins must surely have explanatory efficacy relative to the cognitive capacities of human scientists. If the only physical "theory"* from which psychological theory is derivable is one which is nonexplanatory for a human audience, then human scientists would not have devised it in the first place. For humans would find it unilluminating and so it would not appear in their science. Hence, the sort of reductionism which encourages the bottom-up research

---

[7] Churchland, Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind Revised Edition, (Cambridge, MA: MIT Press, 1988), p. 97.

* I place the word 'theory' in quotes due to the fact that a theory must necessarily have explanatory power, and if a collection of generalizations lacks such power relative to a human audience, it is unclear whether it merits the title of 'theory'.

strategy is one in which psychology is physically explainable relative to a human audience.

Another research strategy taken by some reductionists to follow from their view is what Patricia Churchland calls the "co-evolutionary methodology."[9] On this approach, psychological and physical scientists cooperate in the formation of their respective theories. Specifically, scientists periodically attempt to devise bridge laws connecting generalizations in both domains. A failure to be able to devise bridge laws is taken to mean that at least one of the hypotheses must be revised. Revision is expected to take place within both fields until finally bridge laws are formulated thus showing the two fields to stand in the reductive relation. But since human beings are the ones expected to form the physical theory in question, it must be expected to have explanatory efficacy relative to a human audience. The "theory" would be of no use to human scientists if it were not illuminating to them.

The upshot is that the supposed truth of the sort of reductionism which intertheoretic psychophysical reductionists espouse is incompatible with the view that reductionism is false relative to a human audience. Reductionism relative to a suprahuman audience or reductionism in some categorical sense alone could not have

---

[9] Patricia Smith Churchland, Neurophilosophy: Toward a Unified Science of the Mind-Brain, (Cambridge, MA: MIT Press, 1986), pp. 373-6.

the methodological implications which reductionists take their position to have. Hence, the falsehood of reductionism relative to a human audience is sufficient to refute the sort of intertheoretic psychophysical reductionism espoused in the literature.

Another possible defense of psychophysical reductionism appeals to the distinction between the perlocutionary and the illocutionary senses of (3). Without attempting a complete analysis of either interpretation of (3), the following sufficiently delineates that distinction for present purposes: if B explained x to audience A by citing E in the illocutionary sense, then B attempted to produce an understanding of x in A by citing E; and if B explained x to audience A by citing E in the perlocutionary sense, then B attempted to produce an understanding of x in A by citing E and B actually succeeded in producing that understanding in A.[10] A defender of psychophysical reductionism might reason as follows: Not only is there an exclusively illocutionary analysis of (3), but there is an exclusively illocutionary analysis of (2) as well. That is to say, a generalization can explain an explanandum relative to an audience in the merely illocutionary sense provided that the

---

[10] Please note that I am here only stating necessary conditions for explanation. So I am not committing myself to the view that if someone presents me with a beer with the intention of my coming to understand equilibrium thermodynamics thereby that the beer counts as an explanation in the illocutionary sense. There are presumably other constraints as well.

generalization can be cited to the audience with the intention of producing an understanding of the explanandum even if that understanding is not produced.

Turning to Kim's disjunction strategy, let us consider the set of physical generalizations containing highly disjunctive predicates which Kim takes to explain psychological theory. One might wish to argue that this set explains psychological theory relative to a human audience in the merely illocutionary sense simply by virtue of the fact that someone (an archangel, presumably) could cite the set to a human audience with the intention of producing human understanding of why psychological theory is true. This understanding would indeed fail to be produced, but that is irrelevant to explanation in the merely illocutionary sense.

In response, it is not even clear that one should permit an exclusively illocutionary analysis of sentences of type (2). Intuitively, such sentences are always used in the stronger perlocutionary sense. But let us suppose, for the sake of discussion, that there really is a merely illocutionary sense of audience-relativized explanatory efficacy. Let us also suppose that psychophysical reductionism is true relative to a human audience in this weak illocutionary sense. But even with this much granted, this cannot be the reductionism taken to have the methodological implications described by the Churchlands.

To say that reductionism is humanly true in the illocutionary sense is simply to say that it is conceptually possible that some suprahuman intelligence capable of grasping the content of the disjunctive predicates could attempt to produce an understanding of psychology in physical terms in a human audience by citing the disjunctive generalizations to that audience. However, the truth of reductionism in this illocutionary sense leaves open the possibility that the human audience will still fail to understand why psychological theory is true in physical terms, and so this kind of reductionism clearly does not have the methodological implications of which the Churchlands spoke. If a set of generalizations fails to produce understanding in a human audience, then it will not appear in human science. Hence, neither the bottom-up nor the co-evolutionary research method would be feasible given that reductionism were true in the weak illocutionary sense while not being true in the strong perlocutionary sense.

Some might be tempted to raise the following, albeit somewhat pedantic, objection, viz. that since Nagel's own theory of reduction is nonpragmatic, and, since I have chosen to define global reductionism in Nagelian terms (in Section 1.3), I should bite the bullet and admit that psychophysical reductionism is true according to the very nonpragmatic view of explanation I have explicitly adopted. I mention this as a possible objection only because I sense

a tendency in the literature to regard Nagel as a Hempelian. In fact, he should not be viewed as a Hempelian, for Nagel's theory of explanation is pragmatic or audience-relative according to the very criteria presented in this chapter. Hence, by defining nomic reductionism in Nagelian terms, I have not committed myself to the view that the only type of reductionism at issue is reductionism in some nonpragmatic sense.

This calls for a bit of an exegetical foray into Nagel's theory of explanation. Like Hempel, Nagel recognizes certain formal, semantic, and metaphysical constraints on explanation. In order for a set of statements to count as an explanation, they must stand in the appropriate logical relations, the statements must have the proper form, some of the statements must sustain counterfactuals, and they must be true. However, Nagel also recognizes what he calls "epistemic" conditions on explanation which are indeed pragmatic. For example, Nagel claims that in order for E to explain X there must be good evidence for the truth of E. The following passage shows that Nagel clearly takes the notion of good evidence to be audience-relative:[11]

> The objection may nevertheless be raised against this condition that, since the evidence for a supposed universal law does not remain constant in time, an explanation that includes the law in its premises and

---

[11] The Structure of Science, p. 44.

> that is satisfactory at one time may cease to be satisfactory when unfavorable evidence for the law is discovered. But the objection is not a disturbing one, unless the dubious assumption is made that in judging an explanation to be satisfactory a timeless property is being predicated of the explanation.

I take it to be uncontroversial that what is being proposed here is a context-sensitive notion of good evidence which entails a context-sensitive notion of explanation. Furthermore, Nagel has at least implicitly committed himself to the view that the explanatory power of a generalization is partly determined by the cognitive powers of the audience. For an audience would not be able to accept anything as good evidence for E if that audience were unable to grasp E.

Let us consider this in the context of the disjunction strategy. According to Kim, strong psychophysical supervenience entails that psychological laws are derivable from physical generalizations which meet formal, semantic, and metaphysical criteria for having explanatory efficacy. However, due to the multiple realizability of the mental, these physical generalizations would be so immensely disjunctive as to be humanly incomprehensible. Manifestly, these physical generalizations could not play a part in a reduction in the Nagelian sense. For no human being could hope to have good evidence for the truth of a generalization whose predicates are so enormously disjunctive as to exceed the limits of human memory.

4.3. Local Reductionism and Eliminativism

In the case of both local reductionism and eliminativism, no effort is made to reduce psychological generalizations in their full generality, applying to human and nonhuman creatures alike. Instead, such wide-grained or high-level generalizations are rejected and replaced with more fine-grained, structure-specific generalizations which can readily be rewritten as physical generalizations. The ur-argument for eliminativism, as was shown in the preceding chapter, is to the effect that lower-level explanations of property-instances are to be preferred to higher-level explanations. The argument for local reductionism presented by Kim, the only argument in the literature for this position, is to the effect that the heterogeneity in causal powers of multiply realizable properties renders them explanatorily inefficacious, and so scientists should appeal to the lower-level physical properties instead by reason of their greater causal homogeneity.

Both positions are to the effect that a property-instance is better explained by lower-level properties than by higher-level properties. A single property-instance can indeed stand in counterfactual supporting relations to both higher- and lower-level properties, since the former strongly supervene on the latter. The localist/eliminativist position is that appeal to the subvenient properties is always explanatorily superior.

This position, however, is ambiguous. It could be taken to mean that the only explanation of a property-instance is in terms of the lowest-level properties. That is to say, that a microphysical explanation is the only genuine explanation of any property-instance. In some passages, Paul Churchland can easily be interpreted as advocating this view.[12] Or this interpretation could be taken to mean that there is some cutoff point in the hierarchy of levels, below which the properties are indeed explanatory but above which, they are not. This cutoff point, however, need not correspond to basic physics. The cutoff point could lie between biology and psychology. Kim is most easily interpreted as advocating this position. However, Stich could also be interpreted as advocating this view in that he never advocates reducibility to basic physics as a criterion of a theory's explanatory power, but does advocate the more fine-grained syntactic theory of mind even at the expense of losing the greater generalities of a semantic theory of mind. Or, finally, this position could be taken to mean that whenever it is feasible to do so, a lower-level explanation should be preferred over any higher-level one. On this interpretation, the

---

[12] See A Neurocomputational Perspective, pp. 287-92, in which Churchland suggests that the categories of basic physics are the only explanatory kinds and that putative explanations in terms of higher-level properties are only ersatz explanations which we pretend are explanatory merely for practical purposes.

localist/eliminativist is not claiming that basic physics is the only science but that the closer one is able to get to basic physics the better is one's ability to explain.

On the first interpretation, the low-level preference clearly violates the cognitive constraint on explanation. For it was already argued in Section 1.7 that, relative to a human audience, there is no microphysical explanation of any psychological property-instance. As noted in that section, any microphysical property sufficient for a mental property could only correspond to a predicate which is highly conjunctively complex, so complex that it would defy human comprehension. So if Kim or the eliminativists mean to say that the best explanation of (e.g.) human behavior must be couched in terms of basic physics, their position is surely incorrect relative to a human audience.

Let us turn to the other interpretations of the localist/eliminativist position. Both interpretations are subject to the same rebuttal, viz. that the audience's interests have a role to play in determining which level of description one must appeal to in truly explaining a given property instance. For example, Kim (as discussed in Chapter Two) points out cogently that similarity in causal powers must be guaranteed by a scientific taxon. Hence, if the physical realizers of pain are diverse as physical kinds, they must be diverse as causal kinds. Moreover, if they are too causally diverse to count as constituting a

physical kind (and, by hypothesis, they are), then they are too causally diverse to count as constituting any scientific kind. Kim concludes that pain, in its original unrestricted folk characterization, is too diverse in its causal powers to constitute a causal kind. It must be replaced by finer-grained (i.e., physical) properties within scientific ontology. As pointed out in Chapter Two, however, Kim's argument rests on the assumption that scientific kind individuation is determined solely by causal powers and not to any extent by the interests of people working in diverse scientific fields. Hence, to the degree that one renders plausible the claim that which properties are explanatorily relevant to a given explanandum is (at least partly) audience-relative, to at least that degree one diminishes the plausibility of Kim's argument for local reductionism.

As for the eliminativists, if I have indeed discerned an ur-argument underpinning their case, it is to the effect that lower-level predicates have greater explanatory efficacy than do higher-level predicates because the former convey more detailed information as to the causal history of the explanandum than do the latter. That is why eliminativists favor the rejection of higher-level folk psychological predicates in favor of more specific physical predicates. Now the very same point which militates against Kim's argument for local reductionism also militates against the eliminativist ur-argument. For the eliminativist

assumes that a more specific causal explanation is always superior to a less specific one without considering that the degree of specificity explanatorily relevant, even to one and the same explanandum, may be partly audience-relative.

One can also understand the eliminativist/localist ur-argument in the following terms. Given the layered view of the world discussed in Chapter One, a worldview to which functionalists evidently subscribe, all causally relevant properties strongly supervene upon physical properties. Equivalently, all nonphysical causal relations depend upon physical causal relations. Once the physical causal facts are fixed, all the causal facts are fixed. Furthermore, cognitive scientists are presumably only seeking causal explanations. Perhaps one could argue against this claim, but I will adhere to it in order to give the eliminativists and localists as much ground as possible. The third premise in the eliminativist/localist ur-argument is derived from functionalist metaphysics, viz. that any mental property is diversely physically realizable. Given multiple realizability and the dependence of all causal relations upon physical causal relations, it follows that physical predicates provide more specific information as to the causal ancestry of any token event than do psychological predicates. In light of the assumption that cognitive science requires causal explanations, eliminativists and localists conclude that all explanations in cognitive

science should be couched in terms of physical predicates. For only physical predicates provide information about the explanandum's etiology which is maximally specific. Kim makes this same point vis-à-vis properties instead of predicates. According to Kim, only physical properties are truly explanatory, since they alone are maximally uniform in terms of causal powers.

I believe this argument to be unsound and, even though I do not claim to be able to make an indubitable case against it, I do believe that I can raise considerations which diminish whatever prima facie plausibility it might possess. In Section 4.1, I argued that any claim of the ostensible form

(1) **E** explains **x**

is elliptical for some claim of the form

(2) **E** explains **x** to **A** when cited by **B**.

Accordingly, constraints on (2)-claims are also constraints on (1)-claims. In Section 4.1, I argued that there is a cognitive constraint on any (2)-claim and hence also on any (1)-claim. Namely, **A** must have the capacity to grasp the semantic content of **E**.

Identifying other constraints on (2)-claims will remove

any prima facie appeal of the eliminativist/localist argument. Specifically, it is plausible to hold that in order for a given (2)-claim to be true, B's citing of E to A must satisfy certain of A's interests. Why this is plausible and just what some of those interests are can be understood by considering a classic example of the interest-relativity of explanation.

In asking why an automobile collision took place, different sorts of causal factors can be deemed explanatory. For example, the oil on the road, the diminished visibility due to the rain, the inebriation of the driver, the overly worn tires on the car, and the sharp turn in the road are each a causal factor that can be cited in an explanation of the collision. Each causal factor can be construed as a candidate explanans. The combination of all the causal factors into a single sufficient condition of the collision can also be construed as a further candidate explanans.

How does one select from all these candidate explanantia a statement of fact which is genuinely appropriate as an explanation of the collision? An interest-relative view of explanation provides an appealing answer to this question: the interests of A provide the needed criterion. If A is interested in the collision by virtue of being a city planner, then A will be interested in the condition of the road and not in the condition of the driver or the vehicle. Accordingly, the sharp turn in the

road will be accepted by A as explanatory. The other factors will be rejected as irrelevant. By contrast, a personal injury lawyer might be interested in the condition of the vehicle for the sake of representing a client, and so would consider the condition of the tires alone to explain the collision. A fundamentalist preacher might only be interested in the inebriation of the driver as explaining the collision, and so on. On an interest-relative view, there need be no genuine disagreement among these people. Such a view simply recognizes that what counts as explanatory is partly a local matter.

Now, if one rejects such an interest-relative take on causal explanation, how does one decide which account explains the collision? Without appealing to local interests, what remains is metaphysics. On a nonpragmatic account of causal explanation, one will only accept as explanatory whatever account provides the maximal amount of information about the collision's etiology. This account would not only include exhaustive descriptions of all the aforementioned causal factors but exhaustive descriptions of all factors in the complete causal history of the collision going back at least as far as the Big Bang. Without appeal to contextual interests as a selective factor, there is no way to screen any of this out.

It is generally conceded that local interests are relevant to selecting an account as the appropriate

explanation of a given token event. This is partly due to the desire to have a unified analysis of explanation, a view in which (1)-claims really turn out to be disguised (2)-claims, thus requiring only a single analysis of the verb 'to explain'. On this view, the relevance of interests to (2)- and (3)-claims implies that interests are relevant to explanation simpliciter.

The attractiveness of a unified account, however, only convinces one of the interest-relativity of explanation if one is already convinced of the relevance of local interests to the truth-value of (2)-claims. Let us suppose that interests are not relevant to (2)-claims. Without being able to appeal to interests, how does one determine which candidate explanans is to be accepted? The cognitive constraint might be used to weed out a few, but surely it will not narrow the field down to a single candidate. For there are typically many candidate explanantia for a given explanandum which are equally comprehensible to the audience. Without appealing to interests, one must appeal to metaphysics.

In both cases, one does so by regarding as genuinely explanatory only that account which provides maximal causal information. As regards causal factors, one does this by lumping all the causal factors together and considering nothing less that that sum total to constitute the explanation. As regards levels of description, one does

this by choosing only that descriptive level which provides maximally specific information about the explanandum's etiology. On a nonpragmatic view, only the physical level of description is truly explanatory. All else is straw.

But this supposition is grossly counterintuitive. It requires us to evaluate virtually all (2)-claims as false. For one virtually never takes into account the complete causal ancestry of a token event in attempting to explain it. Moreover, in ordinary circumstances, one seldom attempts to explain a token event using only the predicates of basic physics. Hence, a nonpragmatic view forces us to reject as false many claims which we would commonly consider true. Conversely, the interest-relative view is much more consonant with our quotidian judgments about explanation.

Now, there need not be anything wrong in a philosophical theory providing such a skeptical or counterintuitive result. Perhaps common sense got this one wrong, and we should thus be thankful to philosophy for providing a corrective. Indeed, the sciences have falsified many commonsensical beliefs, such as the view that the earth is flat, that there is a non-relative up and down, etc. But theory of explanation is importantly different from theories about the nature of the earth and of space. The earth, space, and other phenomena investigated by the empirical sciences exist independently of social conventions (except, of course, when the empirical sciences are investigating

social phenomena themselves). They lie out there, their features awaiting discovery. Explanation, by contrast, is a social phenomenon. It is a matter of what we implicitly take it to be as evidenced by how we spontaneously assign truth-values to (1)-, (2)-, and (3)-claims. Hence, a theory of explanation which purports to show most of our evaluations of such claims to be false is itself a false theory.

It is for this reason that interests are commonly taken by philosophers to play a role in determining which causal factor is to be taken as explaining the explanandum in a given context. The same sort of reasoning, furthermore, applies to choosing among levels of description. It is as plausible to accept an interest-relative view in one case as in the other.

I conclude that which descriptive level contains the explanation of a token event is partly a contextual matter. At the very least, I believe I have raised positive considerations in favor of this view. Accordingly, I question the localist/eliminativist assumption to the effect that a token event must be explained simpliciter. It would appear, rather, that it is to be explained relative to local interests.

In Chapter Two, I have already discussed which sorts of interests are relevant to choosing a descriptive level. If one is doing a science in which a great premium is placed on

predictive precision, one will choose a taxonomy of categories which are maximally causally uniform. Hence, this consideration draws one toward a lower-level taxonomy. However, if one is concerned with locating the explanandum within a more abstract pattern of nature, one will be drawn toward a higher-level taxonomy.

Consider an example. Suppose that the explanandum is the decline in religious faith in a certain geographical region or community. Let us also suppose that this decline in religious faith has occurred on the heels of an increase in urbanization in that region or community. Now, it is indeed the case that an increase in urbanization usually does result in a decline in religious faith. One would appeal to this generalization in explaining this explanandum if one were interested in the fact that an increase in urbanization would greatly probabilify the explanandum even if the lower-level conditions upon which the urbanization supervenes were different. So long as an increase in urbanization occurred, whether its physical realizers were the actual ones or something different, the decline in religious faith would likely have occurred anyway. An interest in these relatively abstract, counterfactual supporting generalizations of probability inclines one toward opting for a higher level of description in devising an explanation.

However, an increase in urbanization does not <u>guarantee</u>

a decline in religious faith. If one's interest is in a science capable of greater predictive precision, one will go down to a more micro level thus taking into account the decisions, personalities, and beliefs of individual people. If one's interest in precision is maximal, one is drawn all the way down to the level of basic particles and fields.

The localist/eliminativist ur-argument is inadequate for not considering the possibility that explanation might be interest-relative in this way and the considerations favoring such a view.

4.4. A Possible Objection

I am here concerned with some philosophers who agree with my conclusion but who would reject the steps of reasoning I have taken to reach it. According to these philosophers, the eliminativist/localist argument fails because no psychological explanandum can take purely lower-level information as constituting its explanation. This differs from my view according to which either a lower-level explanans or a higher-level explanans is an option depending upon local interests.

One such philosopher is Alan Garfinkel who claims that the explanans must correspond to the same level of description as does the explanandum, for only at that level can one specify causal conditions for the explanandum which

are both necessary and sufficient.[13] By contrast, Garfinkel claims, an attempted explanation at a lower level is overly specific and so presents conditions which are sufficient but not necessary. The clearest attempt to illustrate this is Hilary Putnam's example of the peg and board:[14] Assume that the explanandum is the failure of a square peg slightly more than one inch across to fit through a square hole one inch across in a board. Apparently, an attempted explanans appealing to lower-level microstates would, even though presenting sufficient conditions for the explanandum, fail to provide necessary conditions for it. For the properties of being a rigid peg and of being a rigid board satisfying the above description are realizable by diverse physical microstates. Hence, a lower-level account would only concern one possible physical antecedent for the explanandum. Garfinkel concludes that a lower-level account fails to make clear which conditions are necessary as well as sufficient and so is inferior to at least one higher-level account.

Garfinkel is assuming that the virtues of a good explanation for any given explanandum are to be found at one level of description only. Accordingly, that level is the privileged level relative to the explanandum in question. I

---

[13] Alan Garfinkel, Forms of Explanation (New Haven, CT: Yale University Press, 1981) pp. 59-66.

[14] Hilary Putnam, "Reductionism and the Nature of Psychology", Cognition 2 (1973) 131-46.

believe, however, that the localists and eliminativists are correct in noting certain virtues which are more characteristic of lower than of higher levels. The presence of virtues on many levels recommends a pragmatic approach. By way of showing what some of these lower-level virtues are, consider Kim's localist argument. As noted in the second chapter, Kim claims that a multiply realizable or higher-level property fails to be an explanatory kind by virtue of subsuming physical properties which are too diverse in their causal powers. By appealing to the commonly accepted metaphysical principle that similar causes have similar effects, Kim concludes that a multiply realizable property, by subsuming so diverse a lot of physical properties, is too predictively unreliable to be projectible. That is to say, the dissimilarity of the possible physical causes corresponds to a wide dissimilarity in possible effects thus showing the multiply realizable property to be less predictable in its effects than any one of its physical realizers would be. Finally, due to the close link between projectibility and explanation, such a property is too predictively unreliable to be an explanatory kind. Hence, the low-level preference for a maximal specification of the explanandum's etiology rests on the belief that explanatory kinds must be maximally predictively reliable.

Although I am obviously not siding with Kim against

Garfinkel on the issue of physicalistic reductionism, I do believe that Kim makes a good metaphysical point which Garfinkel does not appreciate. Although philosophically appealing, Garfinkel's talk of necessary and sufficient conditions is too simple and clean to fit scientific practice. Scientists virtually never specify necessary and sufficient conditions for an explanandum. Instead, they recognize the statistical nature of virtually every generalization they countenance. Hence, Kim does have a point in appealing to the greater causal uniformity and hence predictive reliability of lower-level properties. Garfinkel, I believe, is mistaken in assuming that going to a lower level of description virtually never increases one's predictive precision. So Kim has presented a real virtue in moving to a lower level of description, viz. predictive reliability.

However, Garfinkel does come near to articulating a genuine virtue of the higher-level in his claim that anything lower is overspecific. Higher levels are more general in that they subsume and disregard a greater degree of possible physical dissimilarities in causal histories. Appeal to this level also has many virtues: epistemically speaking, higher-level properties are often more readily discoverable by humans; predictions made via them, although more likely to fail, are usually easier to make; and they are often more interesting by reason of corresponding more

closely to our commonsense categorial schemes.

Accordingly, as I claimed in Chapter Two, scientists working in different fields, such as neuroscience and psychology, would accept different explanations as to why Jones is in pain at t. Presumably, there is some difference in their interests which determines that different descriptive levels provide the explanation of the pain-instance. This difference in interests should be understood in terms of the degrees of predictive precision and generality which are expected. Psychologists seek generalizations which are more general than those sought by neuroscientists in that psychological generalizations subsume individuals exhibiting a greater range of physical dissimilarity than do neuroscientific generalizations. This is to say nothing more than that psychological generalizations are couched using predicates referring to properties which are more greatly multiply realizable than those of neuroscience. However, this greater degree of multiple realizability, as Kim has pointed out, corresponds to a diminution of causal homogeneity. Hence, psychological generalizations are less predictively reliable than neuroscientific ones.

The degree of predictive precision which one expects is counterbalanced by the degree of generality which one expects from one's generalizations. The net force of these expectations determines the level of description in which

one is interested. Therefore, if one is approaching the question as to why Jones feels pain at $\underline{t}$ from a neuroscientific perspective, one expects an answer couched in terms of predicates which can be used to make more reliable predictions than if one were approaching the question from a psychological perspective. On the other hand, if one is approaching the question from a psychological perspective, one expects an answer couched in terms of predicates of greater generality than those found in neuroscience.

This serves to clarify the point I made against Kim's position in Chapter Two, viz. that which properties are explanatory kinds is not simply a matter of the degree of homogeneity in their causal powers. The explanatory power of a property is partly determined by which descriptive level is relevant given the explanatory context. A more or less causally homogeneous property can be explanatorily relevant depending upon one's interest in more predictively reliable or more general predicates.

If, given one and the same explanandum throughout, higher and lower levels possess virtues not had by the other, as indeed I have tried to show is the case, this lends plausibility to the view that the explanatorily relevant level shifts according to which virtues are of interest to the pertinent audience. This is a pragmatic view whose denial is presupposed by the

eliminativist/localist argument. Its plausibility, accordingly, militates against these physicalistically reductionist positions.

4.5. Conclusion

This is a good point at which to review the course of the argument in this and the three previous chapters. The aim has been to show that functionalist metaphysics actually precludes psychophysical reductionism *pace* philosophers who claim it not only to be compatible with psychophysical reductionism but actually to imply it. Kim, for example, has argued that strong supervenience, one of functionalism's commitments, actually implies that psychophysical bridge laws can be formed provided that one is free to use disjunctive predicates in doing so. However, due to the enormous number of pain's possible physical realizers, the semantic content of the relevant disjunctive predicate would exceed the limits of human cognitive abilities. Hence, no human could comprehend the predicate thus rendering the bridge generalization nonexplanatory relative to a human audience. Without this explanatory efficacy, the generalization would not actually be a law, thus showing the formulation of psychophysical bridge laws indeed to be precluded by multiple realizability.

Local reductionism and eliminativism are attempts to show that multiple realizability impugns the explanatory

worth of mental properties thus implying another sort of psychophysical reductionism, one in which the mental properties familiarly posited by commonsense psychology are replaced by lower-level, more fine-grained properties in the explanation of behavior. In arguing for these positions, which are, in fact, indistinguishable theses, one assumes some invariant standard for the degree of homogeneity in causal powers required for a property to be an explanatory kind, and then attempts to show that multiply realizable properties fail to meet this standard. For example, according to Kim, since any mental property fails to have the degree of causal homogeneity necessary for physical kindhood, then the mental property fails to have sufficient causal homogeneity to be an explanatory kind simpliciter. According to the eliminativists, any account which lacks the degree of specificity vis-à-vis the explanandum's causal history to count as a physical explanation is simply not an explanation simpliciter or is at least inferior to some physical explanation. Both arguments are, in fact, equivalent and thus make the same mistaken assumption, viz. that explanatory kindhood is an entirely noncontextual, metaphysical matter and, hence, that something which does not meet the standards of a maximally specific (i.e., physical) causal explanation is either explanatorily inferior or nonexplanatory.

It is more reasonable, however, to adopt a pragmatic

view of explanatory kindhood such that a property's explanatory efficacy is partly a matter of whether one more greatly desires predictive precision or generality in one's choice of predicates. Accordingly, one may move from one level of description to another depending upon one's interests at the time. Hence, the failure of mental properties to be explanatory vis-à-vis the standards of physical explanatory efficacy does not impugn their explanatory efficacy vis-à-vis psychological standards.

None of this, however, discounts the possibility of some sort of reduction of psychology. In fact, functionalism itself, as noted in the Introduction, countenances the reduction of commonsense psychology to a computational theory of behavior via the type identification of commonsense psychological states with the relevant computational states. This computational theory can and, in fact, should be reduced to a nonphysical theory of some sort. This position will be defended in the following chapter.

CHAPTER FIVE

THE MOTIVATION FOR REDUCING COMPUTATIONALIST PSYCHOLOGY

One might gather from the preceding chapters that a computationalist psychology will prove simply to be irreducible without further qualification. But this does not, in fact, follow. In the Appendix, I defend the plausibility of reductively explaining psychology in terms of a field of science other than physics.

In this final chapter, my aim is to show that it is worthwhile to seek a reduction of computationalist psychology to some theory or field of science which is itself used to explain a wide range of phenomena. For, as I hope to make clear, the plausibility of functionalism at least partly depends upon the likelihood of such a reduction.

5.1. Functionalism's Reductive Commitment

My aim in this chapter is to show that functionalists should seek a reduction of computationalist psychology to some other field of science. If the computationalist psychology itself should prove to be irreducible to another theory which, in turn, explains a wide range of phenomena, this will cast doubt on the powers of the computationalist theory to reduce commonsense or folk psychology. In making this point, it is useful to begin by emphasizing the necessity

for functionalism of there being such a computationalist psychology which itself reduces folk psychology.

Although functionalism is perhaps associated with antireductionism in the minds of many, functionalism itself is clearly a kind of reductionism. Functionalists foresee scientists devising a theory that explains behavior, a theory consisting of laws which relate second-order properties, more specifically computational properties. Functionalists also foresee that folk or commonsense psychological laws will be logically derivable from these computationalist laws in conjunction with intertheoretic identifications, viz. the identification of commonsense psychological properties with these computational properties. Hence, it is a bit of an oversimplification to say that some functionalists consider psychology to be irreducible. They do, in fact, consider commonsense psychology to be reducible to some such computationalist theory.

That commonsense or folk psychology is a distinct theory from the computationalist psychology which is meant to reduce it, is evident from the fact that commonsense psychology is largely neutral as to the natures of the states that it posits. More specifically, it leaves open whether these are states characteristic of an unextended substance as in Cartesian metaphysics, whether talk of such states is simply shorthand for talk of dispositions for

observable behavior, or whether such states are computational states. The computationalist theory envisioned by functionalists, by contrast, is a theory as to the natures of the states which explain behavior. They are, of course, hypothesized to be computational states. Hence, the computationalist theory is distinct from folk psychology itself and is meant reductively to explain it.

It is often thought that multiple realizability poses a threat to the physical reduction of psychology simpliciter, but this is not so. Multiple realizability poses a threat to the physical reducibility of folk psychology only by posing a threat to the physical reducibility of the computationalist theory to which folk psychology itself is presumably reducible. As earlier noted, the properties posited by this computationalist theory, being second-order properties, exhibit what Ronald Endicott has termed "compositional plasticity,"[1] i.e., many radically diverse physical properties can satisfy the causal profile definitive of any such second-order property. It follows that the computationalist theory's posited properties exhibit the enormous and wide-ranging physical multiple realizability which precludes a physicalist reduction. Since the only means of reducing folk psychology to physical theory would be via a physicalist reduction of a theory of

---

[1] Ronald P. Endicott, "On Physical Multiple Realization" Pacific Philosophical Quarterly 70 (1989) 212-24.

the nature of folk psychological states, then, if indeed such a theory as to their nature is computationalist, the compositional plasticity of computational properties implies the irreducibility of folk psychology to a physical theory.

Functionalism is committed to the view that there is some such computationalist theory of behavior which itself has the power to explain commonsense psychology. For the computationalist theory putatively reduces folk psychology, and reductive power is a kind of explanatory power. Furthermore, due to the multiple realizability of computational properties, this computationalist theory, at least prima facie, has the appearance of being irreducible. However, it ill behooves the functionalist to assert that the computationalist theory is irreducible simpliciter. For if the computationalist theory fails to be reducible to another theory which itself explains a wide range of phenomena, this casts doubt on whether it has any genuine reductive power. This is, in fact, the primary claim of the present chapter. Hence, the fundamental claim of functionalism, that some computationalist theory explains commonsense psychology, is rendered significantly less plausible if the computationalist theory itself proves to be unexplainable in terms of some other science which itself explains a broad range of phenomena.

Briefly stated, my argument for this claim is as follows: A necessary condition for a scientific theory being

explanatory is that the entities or properties which it posits be real. This claim has an epistemic corollary, viz. that one is only justified in construing a theory to be explanatory if one has reason to interpret its ontology realistically. A theory positing unobservables (e.g., electrons, magnetic fields) merits a realistic interpretation only if the theory plays a role in unifying the sciences. More specifically, its ontology merits realistic interpretation only to the extent that positing that ontology serves to explain as great a variety of observable phenomena as possible in terms of the smallest variety of unobservable phenomena possible. Furthermore, a computationalist theory of the sort functionalists expect to reduce folk psychology, would be a theory of unobservables. If such a computationalist theory is itself irreducible, then it will probably fail to play a sufficient role in the unification of the sciences in order for us to be rationally justified in taking it as genuinely explanatory. This shows why it is worthwhile for anyone arguing for the physicalistic irreducibility of computationalist psychology to show that such a psychology is explainable in terms of some science or other even if not a physical one.

5.2. The Need for a Realist Construal of Computationalist Ontology

In support of the claim that a computationalist theory must

be reducible to a theory explaining a wide range of phenomena in order that the computationalist theory itself be able to reductively explain folk psychology, I assume the doctrine of explanatory realism. Kim's characterization of explanatory realism, moreover, is the one which I adopt.[2] On this view, both the explanans and the explanandum of any given explanation must posit properties or entities. Necessarily, a set of statements is an explanans for a putative explanandum only if the posits of the former stand in some determinate, objective relation to the posits of the latter such that the former determine the latter. The clearest illustration of explanatory realism involves the causal explanation of a token event e by appeal to another token event c. The explanans C expresses the claim that c occurred, and the explanandum E expresses the claim that e occurred. According to explanatory realism, C explains E only if e stands in an objective relation of dependency upon c. In this case, that relation is presumably the causal one. In other cases, the relations of supervenience or identity could constitute the explanatory relation. Explanatory realism is extendable to theoretical or reductive explanation as well. On the realist view, if theory T reductively explains theory T', then the properties

---

[2] Jaegwon Kim, "Explanatory Realism, Causal Realism, and Explanatory Exclusion", Midwest Studies in Philosophy 12 (1987) 225-39; reprinted in Ruben op. cit. 228-45. See pp. 229-31.

posited by T′ objectively depend upon those posited by T.

Given this realist view, a theoretical structure[3] only has explanatory power if the properties it posits are capable of standing in objective relations either to properties posited by another theoretical structure or to properties which are directly observable. From this it follows that an explanatory theoretical structure necessarily refers to properties which exist independently of our interests or procedures for devising such structures. For the dependency relations between properties can only have an objective existence if the properties themselves have an objective existence.

I will not argue for explanatory realism, but I will attempt to show that it is consistent with the pragmatic view of explanation presented in preceding chapters. This point is worth making, for a pragmatic view of explanation is often associated with theoretical irrealism, viz. the view that an explanatorily efficacious theory need not posit entities or properties having an existence independent of our interests. (This association is largely due to the historical accident of van Fraassen being both an explanatory pragmatist and an irrealist vis-à-vis all

---

[3] In previous chapters, I have used the term 'theory' to refer to a system of statements, lawlike in their form, which is genuinely explanatory. In the present context, I do not want to beg the question of whether a particular instance of such a structure is genuinely explanatory. Hence, I opt for the term 'theoretical structure' in order to leave this question open.

theoretical structures which posit unobservables.[4]) On an irrealist perspective, it is enough that a theory's implications be empirically adequate, i.e., that it imply much of what is observable and be consistent with all of it.

Now in discussing the pragmatic aspects of explanation in the preceding chapters, I was discussing constraints on explanatory efficacy. That an explanans' semantic content be apprehendable by the audience in question was one such pragmatic constraint. That the explanans be pitched at the level of description of interest to the audience in question was another. However, neither of these two constraints precludes the statements constituting the explanans from being interpreted literally. That the explanans be subject to a realistic interpretation is simply one more constraint, completely consonant with the previously discussed contextual constraints. Hence, explanatory realism is compatible with the pragmatic view of explanation defended earlier.

What bears emphasizing in this context is that explanatory realism implies that a genuinely reductive theoretical structure must posit an ontology which actually exists. This assumption is crucial for the argument of the current chapter. Now suppose that the functionalist dream of devising a computationalist theoretical structure has

---

[4] Bas C. van Fraassen, The Scientific Image (Oxford: Clarendon Press, 1980).

been realized such that, as the functionalist would project, the laws of folk psychology (in conjunction with bridge generalizations) can be derived from this structure. Would this show folk psychology to be reducible to the computationalist theoretical structure in question? Given explanatory realism, this would not be sufficient to show that there is such a reducibility. For in order that such a derivation be a genuine reduction, it must be an explanation. That is to say, the derivation of folk psychological laws from the computationalist theoretical structure must constitute an explanation of the truth of folk psychology in computationalist terms. However, given explanatory realism, a necessary condition for the computationalist theoretical structure's being reductive is that its ontology of computational properties be realistically construed. The mere derivability of folk psychological structure from computationalist structure is not enough to ensure this. For given explanatory realism, such a derivation is only a reduction if the putatively reductive theoretical structure's ontology actually exists.

This reflects Michael Friedman's distinction between two different sorts of theoretical derivation.[5] On one hand, there are theoretical derivations in which the derived theory is genuinely explained by the deriving theoretical

---

[5] Michael Friedman, "Theoretical Explanation", in Richard Healey (ed.), Reduction, Time and Reality (Cambridge: Cambridge University Press, 1981) 1-16.

structure by virtue of the reality of those properties and entities posited by the deriving structure. On the other, are those derivations in which the entities and properties of the deriving structure are fictional. In the latter case, the derivation is not a genuine reductive explanation.

There are examples of both in the sciences. Indeed, not all theoretical structures employed by scientists are meant to be taken literally. An example of a such a structure is that of color space as exemplified in the "color wheel." Various properties of colors can be derived from aspects of the color wheel, and yet no one is tempted to view such a derivation as a genuine reduction of color theory. Furthermore, it is evident that the nonliteralism of color space talk is sufficient to prevent it from being genuinely reductive. According to Friedman, in such a case as that of color space, the function of the deriving theoretical structure is to supply a mathematical model or representation for the phenomena captured by the derived theoretical structure. The latter phenomena are correlated with phenomena in the deriving theoretical structure as a means for making predictions or clarifying relations. In order for such a theoretical structure to be useful to science, it is enough that colors behave <u>as if</u> they were embedded within color space even though there is not literally any such space for them to be embedded within. Talk of phase space in mechanics and of sound space in

perceptual acoustics provides further examples of such nonliteral and hence nonreductive theoretical derivations.

This is in contrast to cases of theoretical derivation in which the deriving theoretical structure is meant to be taken literally thus justifying one in viewing the derivation as a genuine reductive explanation. The reduction of thermodynamics to statistical mechanics is a case in point. In such a derivation, temperature is not merely correlated with the mean kinetic energy of molecules. The former is literally identified with the latter. Our reasons for believing in the reality of molecules make us confident that the derivation is a genuine explanation.

It is possible, in fact, for one and the same theoretical structure to be either a genuinely explanatory theory or a mere mathematical model of phenomena, depending upon whether its claims are interpreted literally or nonliterally. Such was the case for Copernicanism shortly after Copernicus' death. Copernicus meant for his heliocentric hypothesis to reflect the actual arrangement of the planets in relation to the sun. Upon his death, however, there was an effort made to accept this theoretical structure merely as a useful device for predicting observable celestial phenomena without its being construed literally. I.e., it was proposed that the Copernican hypothesis be accepted for, say, navigational purposes without being taken literally as the denial of geocentrism.

On this interpretation, the theoretical structure would be useful as long as celestial phenomena appear in certain locations in the sky at certain times as if heliocentrism were true, even though one would be supposing that it is not.

On either interpretation, the theory could play a useful role in science. However, according to explanatory realism, Copernicanism is only explanatory when it is interpreted in the way Copernicus himself intended, viz. realistically. It is, moreover, intuitively appealing to say that Copernicanism fails to be genuinely explanatory if indeed it functions only as a calculus for making predictions. It would seem that it is necessary for its being explanatory for it to be taken literally.

Given that some systems of statements, of value in the sciences, are subject to realistic interpretations while others are not, the question arises as to whether a plausible theory of computationalist psychology would be amenable to a realistic interpretation. If not, it would fail to be explanatory and so fail to reduce folk psychology, thus falsifying functionalism. It is clear, moreover, that functionalists themselves wish for the computationalist theory to be taken literally. Zenon Pylyshyn, for example, has urged that computational models of mind be taken as literal theories as to how the mind works and not as mere computer simulations of mental

phenomena.[6] (This is in contrast to the nonliteralism of the nonfunctionalist John Searle who often insists that computer models of mind are nothing more than computer simulations of cognition analogous to computer simulations of the weather.[7] According to Searle, mental processes behave as if they were computational processes, but they are not literally so.) Given explanatory realism, it is essential for the functionalist to take the literalist position.

The task now is to ascertain what principle or principles of inference would rationally justify a literal interpretation of computationalist ontology. The functionalist obviously needs to appeal to such a principle in order to justify belief in the reductive power of the computationalist theory. If the second-order properties posited by such a theory were observable, presumably this alone would warrant our believing in their literal existence. Since the computationalist theory in question has yet to be articulated, it is still perhaps an open question whether structures will be found in the brain which evidently perform the posited functions. If such structures are indeed found, and their functions are not in dispute, I

---

[6] Zenon W. Pylyshyn, "Cognition and Computation: Issues in the Foundation of Cognitive Science" The Behavioral and Brain Sciences 3 (1980) 111-32.

[7] John Searle, The Rediscovery of the Mind (Cambridge, MA: MIT Press, 1992), p. 218.

submit that the matter will be resolved in favor of a literalist, and hence genuinely reductive, interpretation of the computationalist theoretical structure.

However, it is not clear that neuroscientists will in fact find such structures and be able to identify their relevant causal powers. Or, even if they do, they may do so long after the computationalist theoretical structure has been devised. In the meantime, cognitive scientists of the functionalist stripe will need independent grounds for taking the theoretical structure as genuinely explanatory. Such an eventuality is indeed not implausible, for cognitive scientists currently employ methods for ascertaining the nature of the computational architecture undergirding folk psychological phenomena, methods which do not involve direct inspection of neural structure. According to Pylyshyn, such methods involve the testing of a computationalist hypothesis by appeal to outwardly observable behavior, e.g., reaction times, characteristic errors and patterns of error, differences in performance skills at various points in maturation, etc.[*] With these sorts of constraints at hand, a computationalist theoretical structure may indeed be formulated prior to the neuroscientific means for confirming it observationally.

What one needs is a principle for inferring the existence of unobserved phenomena. More specifically, in

---

[*] Pylyshyn, op. cit.

the absence of the relevant observational data from neuroscience, one needs a reason for believing that psychological properties are most plausibly identifiable with computational properties. It has been thought that the principle of inference to the best available explanation is sufficient to secure the plausibility of this hypothesis. In fact, the inference to a computationalist theory of mind has been thought to be highly recommended by this principle. For, according to Fodor, the identification of cognition with computations performed upon representations is the only theory we have for the physical realization of inferential processes. It is the only means we know of for causal processes to be rationally coherent.* If Fodor is right, a computationalist model of mind would not only be the best available explanation but the only serious one. Hence, in the future when the specifics of such a model are fully spelled out, one supposes that we will be justified in construing its ontology realistically by reason of its being the best available explanation.

However, the notion that one should opt for the best available explanation even when there is only one serious explanation available is a highly liberal principle of inference. Friedman has raised plausible doubts as to whether such a liberal principle of inference is truly

---

* Psychosemantics, p. 20.

legitimate.[10] Friedman asks us to suppose that we are concerned to explain some observable phenomenon Q. Given that the only serious contender for explaining Q is the theoretical structure T_1, the principle of inference to the best available explanation would have us accept T_1, at least provisionally. Against this lenient principle, Friedman points out that any phenomenon whatsoever can be modeled by a theoretical structure containing a large enough universe of sets. That being so, it is possible further to provide a theoretical structure from which T_1 can be derived, viz. T_2. And, of course, T_2 can be modeled by T_3, and so on ad infinitum. This is meant to show that the principle of inference to the best available explanation, at least in its current highly inclusive form, is too inclusive. According to Friedman, this shows that the highly inclusive principle leads to the acceptance of virtually any theoretical structure. Accordingly, the principle of inference to the best available explanation, at least in this unqualified form, is illegitimate.

Therefore, the mere fact that a computational model is presumably the only serious candidate for a theory as to the physical implementation of rational processes is insufficient motivation for our taking such a theoretical structure literally. Hence, given that reductive power requires the literal truth of the theoretical structure in

---

[10] Op. cit., pp. 5-6.

question, it is not sufficient grounds for our taking such a computational model as providing a genuinely reductive explanation of commonsense psychology. Friedman himself makes this point against a specific type of computational model of mind, viz. Fodor's language of thought hypothesis.[11] Fodor has long claimed that the postulation of computational manipulations of sentence-like representations in the brain is the only serious means of explaining cognition. On the face of it, the truth of this claim would seem sufficient to warrant our taking these postulations literally. According to Friedman, however, it is not. It has not been ruled out that such a model is simply an "as if" theoretical structure. I.e., it is possible that the supposed reduction of cognition to such sentential computations is nothing more than the claim that cognitive processes behave <u>as if</u> they were sentence-crunching processes. This would be analogous to the instrumentalist interpretation of Copernicanism according to which celestial phenomena behave as if the planets and the earth revolve around the sun even while presupposing that they do not literally do so. Such instrumentalism does not imply the literal truth of the theoretical structure in question. It leaves open the possibility that the structure is merely a mathematical representation of the phenomena in

---

[11] Cf. Fodor, <u>The Language of Thought</u> (New York: Thomas Y. Crowell, 1975).

question but not a genuine explanation of them.

My point here is not to consider Fodor's language of thought hypothesis in particular. Rather, my point is that the computational model of mind, whether it posit sentence-crunching processes as basic to cognition or not, being the only serious model of mind does not secure its being confirmed as literally true. An instrumentalist interpretation of such a model has not been ruled out.

What is still needed is a reliable principle of inference that will serve to confirm at least partially the unobserved computationalist ontology. For without some degree of confirmation, there is no rational ground for construing such a model as genuinely explanatory and hence no rational ground for construing it as genuinely reductive of commonsense psychology.

5.3. No Explanation Without Unification

Friedman suggests a more satisfying criterion for construing a theoretical structure literally. According to Friedman, "A good or fruitful theoretical structure does not serve simply to provide a model for the particular phenomenon it was designed to explain; rather, in conjunction with other pieces of theoretical structure, it plays a role in the explanation of many other phenomena as well."[12] This principle receives support from the realist view of

---

[12] Ibid., p. 7.

explanation presupposed earlier. For on that view, a necessary condition for the explanatory efficacy of a theoretical structure is that its ontology be real. An epistemic corollary of this realist view is that one is only justified in taking a theoretical structure as genuinely explanatory when one has reason, at least provisionally, to accept it as well confirmed.

Friedman's principle, moreover, ensures that a theoretical structure will only be considered good or fruitful if it receives some degree of confirmation. For, as Friedman notes, a nonobservational theoretical structure is only confirmed to the extent that it plays some role in the explanation of phenomena other than those which it was originally designed to explain. Hence, the more such a structure plays a part in explaining other phenomena, the more one has reason to construe its ontology realistically and, consequently, the more one is justified in taking it to be genuinely explanatory.[13]

Friedman does not attempt to provide a precise answer to the monumentally important epistemological question as to precisely when a theory is confirmed. But he does make some general points relating to this issue. According to

---

[13] This is not, of course, the only criterion by which to judge an ontology realistically. The principle of inference to the best explanation does play some role here, but it must be constrained by the requirement that a theoretical structure play some role in the unification of the sciences.

Friedman, the derivation of the particular set of phenomena from a theoretical structure that that structure was specifically designed to explain is not sufficient to confirm that theoretical structure. This can be illustrated by considering the example of color space. The notion of color space was developed in order to render explicit the relations between certain chromatic properties. However, since its development, no other well established phenomena have been shown to be derivable from the color space theoretical structure. It is due to this lack of interaction with other domains that one feels no motivation to take talk of color space literally. Instead, one takes it to be merely an instrumentalistic calculus for predicting certain chromatic phenomena. However, if a theoretical structure plays some role in the derivation of other phenomena beyond those it was originally designed to account for, then it begins to acquire additional confirmation. It is surely a very difficult question as to how much additional confirmation is required before one should rationally take such a structure to be confirmed, but it is sufficiently clear that a mere mapping of the theoretical structure onto the aspects of the phenomena for which it was originally designed to explain does not suffice. For then there is no impetus for believing in the theoretical structure other than its accomodating those phenomena which it was specifically designed to accomodate.

This, according to Friedman, shows the virtue in the unity of science program. Put simply, the greater the number of observational phenomena which can be derived from the smallest amount of theoretical structure, the more reason we have for believing in the ontology posited by that structure. Accordingly, there is an epistemological motivation for striving for the unification of the sciences. And a particular theoretical structure is more greatly confirmed to the extent that it contributes to the project of unification.

Eventually, it will be necessary to evaluate computationalist psychology in terms of whether it plays a role in unifying the sciences. But before considering that issue, it is necessary to consider an ambiguity in Friedman's formulation of this necessary condition for theoretical confirmation. Resolving the ambiguity will play an important part in the subsequent discussion of functionalism.

There is both a strict and a more lenient interpretation of Friedman's principle. On the strict interpretation, a theoretical structure T only receives confirmation if it agrees with the following pattern:

(1) E <---- T ----> E'.

E is the domain of phenomena which theoretical structure T

was originally designed to explain. E′ is other phenomena which were subsequently found to be derivable from T perhaps in conjunction with other pieces of theoretical structures not appearing in the schema. The arrows represent relations of derivability such that E and E′ are each independently derivable from T. In a case fitting this description, T receives some confirmation because it has at least one explanandum in addition to the explanandum (or explananda) which it was initially designed to have. All of Friedman's examples of theoretical structures which play a unifying role fit this pattern. He adduces, for example, the molecular model of a gas. This model (T) not only explains the behavior of gases (E) but also explains chemical bonding (E′) when conjoined with the atomic theory of molecular structure and the identification of chemical elements with various kinds of atoms.

However, this is not the only way that a theoretical structure can play a role in unifying the sciences. For consider the following schema:

(2) E <---- T ----> T′ ----> E′.

In this sort of case, E′ is derivable from T but only via the theoretical structure T′. E′ is itself derivable from T′. In fact, E′ is the domain of phenomena which T′ was originally devised in order to explain. However, no other

domain of phenomena has been shown to be derivable from T'.

Does T' acquire additional confirmation from E? It is not clear that Friedman would say that it does. He does not include such a case in any of his examples. Hence, it is not clear that Friedman would consider T' to merit a realistic interpretation. However, I believe that Friedman's principle should be interpreted leniently enough to include T' as receiving some confirmation from E. For T' clearly contributes to the unification of the sciences since without T', E' would not be derivable from T. Moreover, T' does play a role in the explanation of E, since T' plays a role in the confirmation of T, and the more T is confirmed the more reason one has to construe T literally and thus as having the power to explain E. Thus, T' helps to explain E simply by virtue of contributing to the confirmation of T.

This sort of abstract talk may be somewhat persuasive, but an actual example from the sciences fitting the pattern of (2) is also needed. Such an example is to be found by considering the history of structural chemistry.[14] In this example, phenomenological chemistry, the chemistry consisting of laws relating purely observable chemical properties, takes the place of E'. What takes the place of T' is structural chemistry, viz. the postulation that different chemical compounds consist of atoms of certain

---

[14] I would like to thank to Prof. Shaughan Lavine for this example.

specified elements arranged in certain spatial arrays. (Each atom in the compound's molecule is represented by one or two letters, and the atomic symbols are connected by dashes representing the chemical bonds.) Structural chemistry is further reducible to atomic theory which takes the place of T.

It is important to note, however, that structural chemistry was developed by Friedrich Kekule in the mid nineteenth century, over half a century before the atomic structure of matter was confirmed. Without the confirmation of atomic theory, structural chemistry failed to be reductively explained in terms of atomic theory at that time. Because of this lack of reducibility to atomic theory, even though phenomenological chemistry was derivable from structural chemistry in the middle of the nineteenth century, scientists were hesitant to interpret the theoretical structure of structural chemistry literally. Some scientists did, but indeed many did not. At the time, it seemed plausible to construe the chains and rings of structural chemistry merely as useful devices for predicting observable chemical phenomena, e.g., predicting that water can be broken down into two parts hydrogen to one part oxygen. At the time, they were analogous to color space or phase space, modeling nothing other than phenomena in the domain which they were designed to represent. But with the independent confirmation of the atomic theory in the early

twentieth century, structural chemistry was found to be derivable from a theoretical structure which itself could be used to derive a wide range of phenomena (E). With this new development, scientists no longer hesitated to construe the chains and rings of structural chemistry as literal depictions of chemical microstructure even though structural chemistry still explained nothing more than phenomenological chemistry. Hence, it would appear that indeed a theoretical structure can pick up confirmation indirectly by being reducible to a theory which itself reduces a broad range of phenomena.

Therefore, there are two ways in which a theoretical structure can pick up confirmation and so be construed as genuinely reductive: it can, as depicted in (1), reduce more phenomena than simply the domain of phenomena which it was originally designed to explain, or it can itself be reduced to another theory which explains just such a broader range of phenomena and so receive indirect confirmation as depicted in (2).

5.4. The Confirmation of Computationalist Theory

In order for us to be justified in taking computationalist theory as reductively explaining commonsense psychology, it is not enough that computationalist theory be the only serious candidate explanans. We must also have grounds for construing the ontology of that computationalist theory

literally. As established in the previous section, the computationalist theory can receive confirmation directly by reducing more phenomena than its original explanandum-domain, or it can receive confirmation indirectly by being reducible to another theory which in turn plays the role of explaining the additional phenomena.

What counts as the domain of phenomena which such a computationalist theory will be designed to explain? An obvious initial answer is that the domain will consist in the truisms of commonsense or folk psychology. Hence, the computationalist theory could receive some confirmation by explaining some phenomena beyond that of folk psychology. In doing so, the computationalist theory would correspond to T in schema (1). Indeed, this is not an entirely implausible eventuality. The cognitive and experiential phenomena extending beyond the truisms of folk psychology which such a theory may explain could conceivably include sleep, dreaming, hallucinatory phenomena, schizophrenia, memory phenomena, how learning is possible, typical patterns of error, the lengths of reaction time, differing cognitive capacities at different points in maturation, etc.[15]

However, there are reasons for doubting that the explanation of these *nonfolk*-psychological phenomena will be sufficient to merit a realistic interpretation of the

---

[15] Generalizations concerning such phenomena as these are usually considered to be absent from folk psychology. See P. M. Churchland, Matter and Consciousness, pp. 58-9.

computationalist theory. One reason simply consists in the fact that it is unclear as to how much additional confirmation is necessary in order to justify a realistic interpretation of a theoretical structure. The derivation of a few additional, nonfolk-psychological phenomena from the computational model may not suffice to merit a realistic interpretation of its ontology. Another and more compelling reason for having doubts is that the computationalist model may actually be tailored to represent these phenomena in the first place. This is especially evident in the cases of reaction time, patterns of error, and differing cognitive capacities at different points in maturation; for these are the very observable phenomena which Pylyshyn suggests be used for determining the nature of the underlying computational architecture. Hence, the computational model may actually be devised with the intention of explaining some nonfolk-psychological cognitive phenomena as well. In this case, the derivability of the additional phenomena would not count as evidence for the literal truth of the computational model.

Therefore, it is far from clear that the computational model will receive sufficient confirmation directly as outlined in schema (1). It may turn out that the computational model does not generate any information other than descriptions of the phenomena which it was originally designed to model or, even if it does, it may not generate

enough to justify a realistic construal of its ontology.

The upshot is that it well behooves the functionalist to argue for the reducibility of such a computationalist theory to another theoretical structure, or, more specifically, to another field of science which in turn can be used to reduce a wide range of phenomena. That is to say, functionalists should be concerned to show that the computationalist theory will play the role of T′ in schema (2). In this way, the unobserved computational ontology could receive confirmation indirectly but powerfully just as the ontological commitments of structural chemistry began to receive confirmation indirectly but powerfully in the early twentieth century with their reduction to atomic theory.

5.5. Conclusion

The multiple realizability of computational properties has been thought by many who are sympathetic to functionalism to show the irreducibility of such a computational model to any other field of science. This accounts for the subtitle of Fodor's celebrated paper on multiple realizability and irreducibility, viz. "The Disunity of Science as a Working Hypothesis".[16] But, as I believe I have shown, if such a computationalist theory fails to contribute to the unity of science, this raises serious doubts as to whether it is actually confirmed. Furthermore, given a realistic

---

[16] "Special Sciences".

interpretation of reductive explanation, such a lack of confirmation implies the absence of sufficient reason to consider the computational account to be a literal reductive model of commonsense psychology. Hence, the irreducibility of computationalist theoretical structure to a theory with a wide domain of explananda calls into question the truth of functionalism itself.

Pace Fodor, functionalists should indeed adopt the unity of science project as a working hypothesis. Failure to do so would leave unanswered the question as to whether computationalism provides a literal and reductive model of commonsense psychology just as the irreducibility of structural chemistry in the nineteenth century left many scientists justifiably uncertain as to whether that theoretical structure was a literal and reductive model of the microstructure upon which observable chemical phenomena supervene.

However, the failure of a physicalist or microreductive explanation of computationalist psychology, a failure which has been argued for in the first four chapters, suggests to many the failure of computationalist psychology to be reducible to any other field of science whatsoever. It has been used to argue for the autonomy of psychology from other sciences. The task at hand, then, is to show that physical irreducibility and irreducibility simpliciter are not one and the same, that computationalist psychology can be

nonmicroreductively explained in terms of another theory which itself explains a wide range of phenomena, thus showing computationalist theory to fill the role of T′ in schema (2). I address this task in the Appendix.

CONCLUSION

The purpose of the dissertation has been to consider what reductive relation, if any, psychology will stand in to other sciences given certain functionalist assumptions. The issue is an important one, for it concerns whether psychology will play a part in the unification of the sciences or whether it will remain independent and autonomous as an irreducible science. I believe that I have shown that the classic view that multiple realizability precludes the reduction of psychology to physical science withstands recent attempts to refute it. For if one were to attempt to form a bridge law linking the predicate 'is in pain' to a physical predicate, one would be required either to use an enormously disjunctive physical predicate or to discard the psychological predicate as nonexplanatory. The disjunction strategy would violate a cognitive constraint on explanation thus rendering the generalization nonexplanatory and hence not genuinely a bridge law. The attempted justification for discarding psychological predicates as nonexplanatory in favor of predicates corresponding to nonmultiply-realizable properties consists in adhering to strictly invariant criteria for what counts as an explanatory predicate or an explanatory kind. The localist and eliminativist assume that any property which is not sufficiently causally homogeneous or, analogously, any

predicate which is not sufficiently specific in conveying information as to the explanandum's causal antecedents to count as physically explanatory is not explanatory simpliciter. They thus assume that the standards of physical explanatory efficacy apply to all forms of explanatory efficacy. I have argued that this nonpragmatic view of explanation is implausible and hence that the localist/eliminativist argument is unmotivated. Hence, both Nagelian and localist/eliminativist attempts to argue for psychophysical reductionism require implausible views of explanatory efficacy.

However, my defense of the physical irreducibility of psychology on the basis of multiple realizability does not commit me to the view that psychology is simply irreducible, i.e., that there is no other science which can explain it. In fact, functionalism would lose much of its plausibility if it were to imply that psychology is irreducible simpliciter. For functionalism is only plausible to the extent that scientists devise a computational model of commonsense psychology which, in turn, receives corroboration. The computational model, moreover, is only likely to receive corroboration to the extent that it plays a role in unifying the sciences, and it will only likely do that if it is reduced to another field of science which explains a wide range of phenomena.

In the Appendix, I suggest a possible reducer of

computational psychology.

BIBLIOGRAPHY

Block, N. (ed.), Readings in the Philosophy of Psychology, Volume I (Cambridge, MA: Harvard University Press, 1980).

Block, N., "Can the Mind Change the World?", in G. Boolos, 137-70.

Block, N. and J. Fodor, "What Psychological States Are Not", in N. Block (1980), 237-50.

Boolos, G. (ed.), Meaning and Method: Essays in Honor of Hilary Putnam (Cambridge: Cambridge University Press, 1990).

Brandon, R., Adaptation and Environment (Princeton: Princeton University Press, 1990).

Campbell, K., Abstract Particulars (Oxford: Basil Blackwell, 1990).

Churchland, P. M., Scientific Realism and the Plasticity of Mind (Cambridge: Cambridge University Press, 1979).

Churchland, P. M., "Eliminative Materialism and the Propositional Attitudes", The Journal of Philosophy 78 (1981); reprinted in P. M. Churchland (1989), 1-22.

Churchland, P. M., Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind Revised Edition (Cambridge, MA: MIT Press, 1988).

Churchland, P. M., A Neurocomputational Perspective (Cambridge, MA: MIT Press, 1989).

Churchland, P. M., "On the Nature of Explanation: A PDP Approach", in Churchland (1989), 197-230.

Churchland, P. M., "On the Nature of Theories: A Neurocomputational Perspective", in C. W. Savage; reprinted in P. M. Churchland (1989), 153-96.

Churchland, P. S., Neurophilosophy: Toward a Unified Science of the Mind-Brain (Cambridge, MA: MIT Press, 1986).

Clark, A., Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing (Cambridge, MA: MIT Press, 1989).

Clark, A., "Connectionist Minds", Proceedings of the Aristotelian Society (1990), 83-102.

Davidson, D., "Mental Events", in L. Foster and G. Swanson, 79-101.

Dawkins, R., The Selfish Gene (Oxford: Oxford University Press, 1976).

Dretske, F., Knowledge and the Flow of Information (Cambridge, MA: MIT Press, 1981).

Endicott, R. P., "On Physical Multiple Realization", Pacific Philosophical Quarterly 70 (1989), 212-24.

Elster, J., Explaining Technical Change (Cambridge: Cambridge University Press, 1983).

Flanagan, O., Consciousness Reconsidered (Cambridge, MA: MIT Press, 1992).

Fodor, J. A., "Special Sciences, or The Disunity of Science as a Working Hypothesis", Synthese 28 (1974), 97-115.

Fodor, J. A., The Language of Thought (New York: Thomas Y. Crowell, 1975).

Fodor, J. A., Psychosemantics: The Problem of Meaning in the Philosophy of Mind (Cambridge, MA: MIT Press, 1987).

Foster, L. and W. Swanson (ed.), Experience and Theory (Amherst: University of Massachusetts Press, 1970).

Friedman, M., "Theoretical Explanation", in R. Healey, 1-16.

Garfinkel, A., Forms of Explanation (New Haven, CT: Yale University Press, 1981).

Goodman, N., Fact, Fiction, and Forecast (Cambridge, MA: Harvard University Press, 1955).

Harman, G., "Inference to the Best Explanation", Philosophical Review (1965), 88-95.

Healey, R., Reduction, Time and Reality (Cambridge: Cambridge University Press, 1981).

Hempel, C. G., Aspects of Scientific Explanation (New York: Free Press, 1965).

Horgan, T. (ed.), "The Spindel Conference 1983: Supervenience", in The Southern Journal of Philosophy 22 (1984).

Jackson, F. and P. Pettit, "Functionalism and Broad

Content", Mind 97 (1988), 371-91.

Jackson, F. and P. Pettit, "In Defense of Explanatory Ecumenism", Economics and Philosophy 8 (1992), 1-21.

Kim, J., "Concepts of Supervenience", Philosophy and Phenomenological Research 45 (1984), 153-76; reprinted in Kim (1993), 53-78.

Kim, J., "Explanatory Realism, Causal Realism, and Explanatory Exclusion", Midwest Studies in Philosophy 12 (1987), 225-39; reprinted in D.-H. Ruben.

Kim, J., "Supervenience as a Philosophical Concept", Metaphilosophy 21 (1990), 1-27; reprinted in Kim (1993), 131-60.

Kim, J., "Multiple Realization and the Metaphysics of Reduction", Philosophy and Phenomenological Research 52 (1992), 1-26; reprinted in Kim (1993), 309-35.

Kim, J., Supervenience and Mind (Cambridge: Cambridge University Press, 1993).

Lakatos, I., "Falsification and the Methodology of Scientific Research Programmes", in I. Lakatos and A. Musgrave.

Lakatos, I. and A. Musgrave (eds.) Criticism and the Growth of Knowledge (Cambridge: Cambridge University Press, 1970).

Lewis, D., "Causal Explanation", Philosophical Papers, vol. 2 (Oxford: Oxford University Press, 1986), 214-40; reprinted in D.-H. Ruben, 182-206.

Lewontin, R. C., "Adaptation", Scientific American 238 no. 3 (1978) 213-30.

Marras, A., "Psychophysical Supervenience and Nonreductive Materialism", Synthese 95 (1993), 275-304.

Matthews, R. J., "Explaining and Explanation", American Philosophical Quarterly 18 (1981) 71-7; reprinted in D.-H. Ruben, 345-58.

Mayr, E., Animal Species and Evolution (Cambridge, MA: Harvard University Press, 1963).

McGinn, C., Mental Content (Oxford: Basil Blackwell, 1989).

Nagel, E., The Structure of Science (Indianapolis: Hackett Publishing Company, 1979).

Nagel, E., Teleology Revisited (New York: Columbia University Press, 1979).

Owens, D., "Disjunctive Laws", Analysis 49 (1989), 197-202.

Plotkin, H., The Nature of Knowledge: Concerning Adaptations, Instinct and the Evolution of Intelligence (London: Penguin Press, 1994).

Putnam, H., "Reductionism and the Nature of Psychology", Cognition 2 (1973), 131-46.

Putnam, H., "The Meaning of 'Meaning'", in H. Putnam (1975), 215-71.

Putnam, H., Mind, Language, and Reality. Philosophical Papers, Vol. 2, (Cambridge: Cambridge University Press, 1975).

Pylyshyn, Z. W., "Cognition and Computation: Issues in the Foundation of Cognitive Science", The Behavioral and Brain Sciences 3 (1980), 111-32.

Quine, W. V. O., "Natural Kinds", in Ontological Relativity and Other Essays (New York: Columbia University Press, 1969).

Ramsey, W., S. Stich, and J. Garon, "Connectionism, Eliminativism, and the Future of Folk Psychology", in W. Ramsey, S. Stich, and D. E. Rumelhart, 199-228.

Ramsey, W.; S. Stich, and D. E. Rumelhart (eds.), Philosophy and Connectionist Theory (Hillsdale, NJ: Lawrence Erlbaum Associates, 1991).

Richardson, R. C., "Functionalism and Reductionism", Philosophy of Science 46 (1979), 533-58.

Ruben, D.-H. (ed.), Explanation (Oxford: Oxford University Press, 1993).

Rumelhart, D. E.; J. L. McClelland, and the PDP Research Group Parallel Distributed Processing: Explorations in the Microstructure of Cognition vol. 1: Foundations (Cambridge, MA: MIT Press, 1986).

Salmon, W., Statistical Explanation and Statistical Relevance (Pittsburgh: University of Pittsburgh Press, 1971).

Savage, C. W. (ed.), The Nature of Theories, Minnesota Studies in the Philosophy of Science, vol. 14 (Minneapolis:

University of Minnesota Press, 1989).

Seager, W., "Disjunctive Laws and Supervenience", Analysis 51 (1991), 93-8.

Searle, J. R., The Rediscovery of the Mind (Cambridge, MA: MIT Press, 1992).

Sejnowski, T. J. and C. R. Rosenberg, "NETtalk: A Parallel Network that Learns to Read Aloud", Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01 (Baltimore: The Johns Hopkins University, 1986).

Smolensky, P., "On the Proper Treatment of Connectionism", Behavioral and Brain Sciences 11 (1988), 1-74.

Sober, E., "Screening-Off and the Units of Selection", Philosophy of Science 59 (1992), 142-52.

Stich, S. P., From Folk Psychology to Cognitive Science: The Case Against Belief, (Cambridge, MA: MIT Press, 1983).

Teller, P., "Comments on Kim's Paper", in T. Horgan, 57-61.

van Fraassen, B. C., The Scientific Image, (Oxford: Clarendon Press, 1980).

Wesson, R., Beyond Natural Selection (Cambridge, MA: MIT Press, 1991).

APPENDIX

THE PROSPECT OF AN EVOLUTIONARY REDUCTION

In this appendix, I consider the possibility of reductively explaining computationalist psychology in terms of evolutionary biology. I will not pretend to be able to offer a demonstration of the reducibility of psychology to evolutionary theory or any sort of program for doing so. My aim, a more modest one, is to show that an evolutionary reduction of computational psychology, whatever form it may happen to take, is immune to the fatal flaw of any attempted physical reduction. In other words, evolutionary or adaptationist explanation is free to remain at a relatively abstract level of description, rendering it indifferent to physical multiple realizability. Hence, an evolutionary reduction of psychology does not require an implausibly nonpragmatic notion of explanation as would a physical reduction. This will show that Darwinism is at least a candidate for reductively explaining computationalist psychology, whereas physics has been shown not to be.

Furthermore, taking some time to dwell on the prospect of reducing computationalist psychology will help to emphasize that my position is not antireductionist <u>tout court</u>. It will emphasize my agreement with Friedman that psychology should not be an independent science, that the unity of science program is a good thing for epistemological

reasons.

But why an adaptationist explanation of psychology as opposed to some other type of explanation? Indeed, I do not mean to rule out all other possibilities, at least not at this early stage. As established in Chapter Five, the criterion for choosing a reducer of psychology is that it also explain a sufficiently broad range of phenomena. As argued in the final chapter, computational psychology needs to be reduced to such a theory in order to contribute to the unification of the sciences and thus to pick up confirmation. Sciences other than evolutionary biology could conceivably fill this role.

My aim is simply to show that it is plausible to expect psychology to be reduced to some other science. Toward this end it suffices to argue for the feasibility of some one science reducing computational psychology. Doing so lends plausibility to functionalism. If it is also feasible that some other science can play this role, this is fine. For it is obviously not damning of the position that psychology is reducible. What matters to functionalism is that computational psychology be reducible to some science or other with a sufficiently broad range of explananda. It is not absolutely required that the reducer be evolutionary biology.

Nonetheless, evolutionary biology stands out as a likely candidate for explaining psychology. This is because

the psychological properties of organisms are almost certainly adaptations for living in a changing environment.[1] It is, I submit, difficult for anyone believing in evolutionary biology not to admit that psychological traits are adaptations. This fact alone provides a strong prima facie case for the susceptibility of psychology to an adaptationist explanation.

By way of adding even more plausibility, I shall argue for two points in the following sections respectively: not all reductions need be microreductions, for there can also be macroreductions; and an adaptationist explanation of psychology would indeed count as a macroreduction in virtue of being indifferent to physical multiple realizability.

The Very Idea of a Macroreduction

With a physical reduction ruled out, one seeks a nonphysical reduction of psychology. Before discussing the nature of a nonphysical reduction, it is worthwhile to consider once more what is meant by a physical reduction. It is helpful in answering this question to return to the characterization of the physical presented in the first chapter. There I appealed to what is known as "the layered view of the world." This metaphysical view predates modern scientific

---

[1] See Henry Plotkin, The Nature of Knowledge: Concerning Adaptations, Instinct and the Evolution of Intelligence, (London: Penguin Press, 1994), Chapters 5 and 6.

investigation. It is, in fact, discernable in Democritus and Epicurus. However, it was indeed a motivating assumption behind the scientific revolution including its methods of investigation, and empirical research to date provides much information to corroborate it and little or none to falsify it. I conclude that the layered view of the world should be accepted.

According to this view, observable objects are decomposable into smaller parts which are themselves decomposable into yet smaller parts until one reaches a point at which the resolution into smaller parts is impossible, i.e., contrary to laws of nature. The causally relevant properties of the larger objects, on this view, strongly supervene upon the causally relevant properties possessed by their smaller component parts. The causally relevant properties of the smallest possible components, the properties upon which all other causally relevant properties supervene are, by definition, physical properties.

Hence, a physicalist reduction of a field of science such as psychology is a <u>microreduction</u> of the dominant theory in that field, i.e. an explanation of the properties studied in the field by appeal to the microproperties upon which they strongly supervene. Therefore, functionalism rules out a microreduction of psychology by implying the enormous, perhaps infinite, physical multiple realizability of any given mental property.

I wish to argue for the position that reductions need not always be microreductions. I.e., one can reduce the field studying phenomenon x by relating it to or situating it within a broader system of phenomena rather than by dissecting x into its component parts. Now, this might strike one as a position which doesn't merit much serious discussion. After all, one might say, isn't 'reduction' a term of art? Whether or not the notion of macroreduction is coherent depends upon how philosophers have chosen to define 'reduction', one might further claim, and so the matter should be settled by consulting the prevailing philosophical view.

In fact, there is indeed a substantial question at issue. It is not merely a point of terminology. What is terminological is that reduction is the means of explaining one science T' in terms of another T or of replacing T' with T in order further to unify the sciences. Is it possible to help further scientific unity by explaining a science in terms of another science at an equally high level of description; that is the question, and it is a normative issue for scientific practice, not merely a question of how terminology is defined.

There has been a strong tendency among philosophers to answer the question negatively. They say, or, more typically, they implicitly assume, that any reduction must be a microreduction. This tendency is most evident in the

case of physicalist reductionist philosophers such as Kim and the Churchlands.[2] It is especially evident in the examples they use, most notably the explanation of the temperature of a substance in terms of the mean kinetic energy of its constituent molecules. But it is also evident in Fodor who is sceptical of a physical reduction of psychology. Given that functionalism precludes a microreduction of psychology, the identification of reduction with microreduction inevitably leads one to view functionalism as implying the irreducibility of psychology to any other science. This unqualified antireductionism is, evidently, the view of many functionalists.

However, in the previous chapter it was shown why the disunity of science is not a good working hypothesis. Even if the complete unification of the sciences, the reduction of all sciences to a single science, is not achieved, it is worthwhile for scientists to attempt to accomplish as much unification as possible. For, as Friedman has cogently argued, the degree to which the sciences are unified is the degree to which our theories of unobservables are confirmed.

---

[2] The boldest identifications of reduction with microreduction are to be found in the nonphilosophical literature. For example, according to the biologist Robert Wesson, "The principal method of science is to take things apart, reducing the complex to simpler components - reductionism, it is called. By studying the interactions of parts, one achieves understanding of more involved phenomena, and facts of one branch of science are made derivable from the results of a more basic science." See his Beyond Natural Selection (Cambridge, MA: MIT Press, 1991), pp. 2-3.

Given the plausible assumption of explanatory realism, the more these theories are confirmed the greater our rational confidence that they are genuinely explanatory. Hence, our attempt to increase scientific unification falls out of our desire to have rational confidence in the explanatory power of our theories.

As I attempted to show in the previous chapter, the plausibility of functionalism greatly depends upon the reduction of computationalist psychology to some other field of science which in turn explains a broad range of phenomena. Since the computational theory is not susceptible to a microreduction, the simple identification of reduction with microreduction would preclude this important source of corroboration for the computational theory. It follows that it is worthwhile for the functionalist to debunk the view that microreduction is the only feasible means of explaining one field of science in terms of another. The following is an attempt to do so.

One means of weakening one's confidence in this simple identification of reduction with microreduction is to show that the motivation for it largely stems from an implausible view. For if one understands the motivation behind a view to be a weak one, this should, all things being equal, decrease one's confidence in the view.[3] In fact, it seems

---

[3] Hence, I am not attempting to provide a demonstrative argument for the acceptability of macroreduction. Instead, I am attempting to show that the

likely that an implausibly nonpragmatic view of explanation underlies much of the appeal of this identification. On the realist view here assumed, as will be recalled, an explanation must express some objective, determinate relation of dependency such as the relation of causation or of supervenience. (The explanandum phenomenon must be objectively dependent upon or be determined by the explanans phenomenon.) However, on my view, there are also pragmatic elements to explanation, viz. an explanation must be comprehensible to the relevant audience and must also satisfy certain of its relevant interests.

Now, if one were simply to ignore the pragmatic aspects of explanation, then, assuming explanatory realism, one would naturally view an explanation as nothing more than the expression of some objective dependency relation. In the case of an antireductionist like Fodor it is less clear, but in the case of physicalist reductionists such as Kim and the Churchlands it would appear that this is exactly what they have done. That is to say, Kim and the Churchlands appear to believe that if there is some objective relation between macroproperties and microproperties such that the latter wholly determine the former, then it simply follows that the latter explain the former. The strong supervenience of the

---

prejudice against it stems from an implausible assumption. With that done, I believe that one has the intellectual right to appeal to the possibility of macroreduction unless (and until) someone provides some good reason for not doing so.

mental on the physical constitutes just such a relation and, on such a nonpragmatic view of explanation, would suffice to show the mental to be explainable in terms of the micro, the physical.[4]

Given this nonpragmatic view of explanation, the supervenience of all causally relevant properties upon physical properties suggests a certain form for the unification of the sciences to take. Each field of science is concerned with causally relevant properties. Therefore, all properties of interest to any scientific field supervene upon physical properties, and so all properties of any scientific field stand in an objective, determinate relation to physical properties, the sort of relation in which the physical determines the nonphysical. If, due to one's holding a nonpragmatic view of explanation, one simply identifies explanation with the expression of some objective relation between phenomena such that one phenomenon determines the other, then one will view all sciences as being explainable in terms of physics due simply to this relation of strong supervenience.

Given a nonpragmatic view of explanation and hence of reduction, the strong supervenience of all scientifically interesting properties upon physical properties implies there being one obvious means for unifying the sciences,

---

[4] I would like to thank Prof. Akeel Bilgrami for pointing out to me Kim's conflation of supervenience with reduction, a point very similar to the one being made here.

viz. microreduction. This, I submit, is why reduction is so often simply identified with microreduction: the strong supervenience of macroproperties on microproperties, in conjunction with a nonpragmatic view of explanation, implies that the sciences can be unified solely via microreduction. Hence, there would be no other motivation for the philosopher who values the unification of the sciences to posit any other form of reduction. Viewing the strong supervenience of all causally relevant properties upon physical properties as entailing the explainability of all sciences in terms of physics has resulted in philosophers simply identifying reduction with microreduction.

Considerations were raised in Chapter Four, however, militating against a nonpragmatic view of explanation. Hence, the existence of relations of supervenience between two sets of properties is not sufficient to show one to be explainable in terms of the other. Therefore, physics is no longer the obvious candidate for reducing all the sciences, and so there no longer appears to be any strong motivation for considering microreduction to be the only tool for unifying the sciences.

Indeed, Friedman has distinguished another form of reduction, <u>macroreduction</u>, the explanation of a phenomenon, not by considering the properties of its components, but by situating it within the context of a broader system of phenomena. As an example of a macroreduction, Friedman

adduces the explanation of the law of motion for freely falling particles in terms of general relativity.[5] A particle's obeying the law of motion is not explained by considering the properties exemplified by its constituent parts but by identifying the particle with a singularity in the gravitational field. This identification allows the derivation of the law of motion from Einstein's field equations.

Given that multiple realizability precludes a microreduction of computationalist psychology, the option of macroreducing psychology should prove attractive to the functionalist.

Physical Multiple Realizability and the Units of Selection

In order to lend plausibility to the claim that computational psychology is susceptible to an adaptionist explanation, I will attempt to show that adaptationist explanation can be macroreductive, i.e. that it is at least an option in explaining an adaptation that one remain at such a relatively abstract level of description that physical multiple realizability is rendered irrelevant. Hence, an adaptationist explanation of psychology would not require an implausibly nonpragmatic view of explanation as would a physical reduction.

There is a controversy within modern evolutionary

---

[5] Op. cit., p. 6.

biology which bears on this issue, the "units of selection" controversy: If we are to provide an adaptationist explanation for certain phenotypic[6] properties at a particular time, we must do so by appealing to the interaction between something, x, and the environment at an earlier time. For example, assume that the explanandum is the tolerance of a certain grass plant to heavy metals in the soil. This would be toxic to some grasses, but this particular grass is adapted to withstand it.[7] The possession of this property by the plant is to be explained by considering its ancestors and the environments in which they managed to reach reproductive maturity. Certain features of the ancestral plants interacted with their environments enabling them to survive long enough to reproduce. Thus the metal-tolerance of the descendant plant is to be explained.

The units of selection controversy concerns what should fill the role of x. The ancestral plants which survived to reproductive maturity differed from those of their peers which did not survive. They were different at the genotypic level as well as at the phenotypic level. For the only

---

[6] An organism's genotype consists in its genetic material and is, of course, only observable microscopically. An organism's phenotype consists in observable or readily ascertainable features of the organism which result from its genotype and environment.

[7] This particular example of adaptation is used in Robert Brandon, Adaptation and Environment (Princeton: Princeton University Press, 1990), pp. 140-2.

phenotypic differences which matter in natural selection are those which correspond to genotypic differences. Hence, those plants which were able to survive in the metal contaminated soil must have differed from those which failed to reach reproductive maturity not only in terms of phenotype but in terms of genotype as well. Hence, the struggle to survive in the face of an environment with certain features can conceivably be thought of as the struggle of a particular genotype to survive or of a particular phenotype to survive.

Those who take genotypes to be the units of selection, see adaptations as being explainable in terms of the interaction between ancestral genotypes and environments. By contrast, those who see organisms or phenotypes as being the units of selection see adaptations as being explainable in terms of the interaction of ancestral phenotypes and environments.* There are also those who take groups of organisms or even entire species to be the units of selection. But I will continue speaking as though the debate were simply over whether the units of selection are genotypes or phenotypes. The real issue, as will become apparent presently, is whether the units of selection are

---

* The reference to ancestral phenotypes oversimplifies the matter a bit. For one also partly explains an adaptation by appealing to the ability of the current phenotype or genotype to continue reproducing within the current environment. But this qualification is of no real concern to the strength of the argument.

smaller than phenotypes. For if they are aggregates of phenotypes, then multiple realizability is no stumbling block to devising an adaptationist explanation of psychology.

The question is a familiar one, viz. at what level of description should an explanation of a given phenomenon be couched? It is, then, not at all surprising that the issue of multiple realizability is relevant to the units of selection controversy. Let us suppose that the units of selection are necessarily genes. It follows that reference to the genes must be made in explaining adaptations. Hence, in providing an adaptationist explanation of certain phenotypic properties, one must refer to certain types of genes and how they have managed to survive in certain environments by being housed in certain bodies with certain appropriate adaptations. Viewing evolutionary explanation as lying in the interaction of genotype and environment results in one viewing the phenotype as a mere device for preserving genes. Richard Dawkins, who believes that genes are indeed the units of selection, has said, "A monkey is a machine which preserves genes up trees, a fish is a machine which preserves genes in the water; there is even a worm which preserves genes in German beer mats. DNA works in mysterious ways."*

---

* Richard Dawkins, The Selfish Gene (Oxford: Oxford University Press, 1976), p. 22.

Given this assumption, psychological properties must also be explained by appealing to the genes housed by the relevant organisms (and their ancestors). To explain a given psychological feature would require reference to the specific genes of the organism as well as the relevant environment. But since genes are small component parts of organisms, this means that an adaptationist explanation of psychology would contain a microreductive element after all. This is where the phenomenon of physical multiple realizability becomes relevant. There could be different species which are psychologically identical and yet which are genotypically distinct. A nonhuman species could develop many of the same psychological properties as we ourselves possess as adaptations to a similar environment and yet be microstructurally, genetically distinct.

That this is a genuine possibility becomes especially vivid in considering the phenomenon of convergence. Convergence consists in the development of the same functional property by unrelated groups of organisms as a means of adapting to similar environments.[10] The wings of birds and insects provide an example. Locomotion in water

---

[10] It is interesting to note that Block and Fodor have actually appealed to the phenomenon of convergence in arguing for the multiple realization of mental properties. They suggest that psychological features can arise as a similar solution to the same environmental problems even in species which are microstructurally dissimilar. See Block and Fodor, "What Psychological States Are Not", in Block, ed. Readings in Philosophy of Psychology, Volume I, (Cambridge, MA: Harvard University Press, 1980), p. 238.

provides another: many marine creatures have similar features enabling them to travel underwater. Fish have fins, while sea snakes possess a flat cross section.[11]

What is important to note here is the similarity in environment which produces convergence and the resultant identity of phenotypic adaptation. A single type of phenotypic adaptation, say possessing wings, can correspond to chemically distinct genetic materials in different species. Hence, if genes are necessarily the units of selection, the phenomenon of wings in general cannot receive a single nonconjunctive explanation. Instead, there would have to be a distinct explanation of the possession of wings for each distinct phylogeny. One can imagine a similar problem for the explanation of psychological properties. The same psychological property distributed over different species would have to receive distinct explanations, species-specific explanations. The explanation of psychology in general, in fact, would have to be a conjunction of such explanantia for all possible species possessing psychological properties. This conjunctive explanation would indeed be so enormously conjunctive as to fail to count as a reductive explanation. It would violate the cognitive constraint on explanation discussed in the first chapter.

---

[11] See Richard C. Lewontin, "Adaptation", Scientific American 238 no. 3 (1978) 213-30.

Now, one might want to argue that Dawkins' "gene-centered" view of evolution only has such a consequence if one assumes a certain view of how genes are individuated. More specifically, in order to reach this pessimistic conclusion, one must assume that genes are individuated according to their intrinsic chemical natures rather than functionally. One could, one might wish to suggest, individuate genes according to their phenotypic effects. For example, birds and insects would be genetically indistinguishable according to this view since both species possess the same "wing-making" gene. Similarly, if a collection of psychological properties are to be explained, one might suggest that their genetic causes are functionally identical. Hence, so one might argue, the genes'-eye view does not compel one to provide a complex conjunctive explanans for a psychological explanandum.

However, a single phenotypic property, such as possessing wings, does not exist, on the gene-centric view, simply for the sake of the gene which produces that trait. It, along with all other adaptive properties, exists for the sake of the entire genotype. So, when two groups of organisms possessing the same phenotypic trait are genotypically distinct enough so as to constitute two distinct species, then the gene-centrist does indeed have to deal with a kind of multiple realizability problem. The very same adaptation to a relevantly similar environment

must be given distinct explanations in distinct species.

This point becomes even more vivid once one considers the fact that the bulk of any organism's genotype has no phenotypic expression whatsoever.[12] Any phenotypic property exists for the sake of the whole bulk of the genotype, even though most of it lacks a phenotypic expression. This bulk, moreover, could certainly vary from species to species, especially if one considers the possibility of extraterrestrial life, while the relevant phenotypic properties remain constant. Hence, an adaptationist explanation of psychology in all its possible manifestations would indeed involve a massively conjunctive explanans, each conjunct corresponding to some genotypic difference. This would, as already noted, violate the cognitive constraint on explanation and so not count as a reduction.

Therefore, the view that adaptationist explanation requires reference to the interaction of genetic material and environment would be damning to the project of providing an adaptationist explanation of psychology. It would show that adaptationist reduction would indeed be microreduction and, hence, that an evolutionary explanation would also require an implausibly nonpragmatic view of explanation. In order to show that the prospect of providing an

---

[12] I would like to thank Prof. Arthur Markman for pointing out this interesting fact to me.

adaptationist explanation for psychology does not meet this fate, one must show that genes are not necessarily the units of selection.

Robert Brandon has argued that phenotypes, not genotypes, play the explanatory role,[13] and I believe that Brandon's argument is basically right in spirit, although I will later raise some criticisms against it. Recall that an adaptationist explanation of some feature requires reference to the interaction between some phenomenon, x, and the environment. The units of selection controversy concerns what x is, genes or phenotypic properties. Brandon's argument is an elaboration of Ernst Mayr's claim that the environment interacts directly with the organism's phenotype but only indirectly with its genotype.[14] According to Mayr, this means that the phenotypic level provides the real explanation of any adaptation. Brandon claims that Mayr's view can be characterized in terms of Wesley Salmon's idea of statistical screening-off.[15] The notion of screening-off as Salmon applies it to explanation is as follows: if event or property A renders event or property B statistically irrelevant to the outcome E but not vice

---

[13] Op. cit., pp. 82-8.

[14] Ernst Mayr, Animal Species and Evolution, (Cambridge, MA: Harvard University Press, 1963). See p. 184.

[15] Wesley Salmon, Statistical Explanation and Statistical Relevance, (Pittsburgh: University of Pittsburgh Press, 1971).

versa, then A affords a superior causal explanation of E than does B. Brandon's interest lies in the case in which the survival and reproductive success of the organism is E, A is the organism's phenotypic properties, and B is its genotypic properties. Letting 'O' refer to the organism, 'P' to its phenotypic properties, and 'G' to its genotypic properties, Brandon's claim that phenotypic properties causally screen off genotypic properties can be stated in formal terms as follows:

> Pr(O survives / O has phenotype P and O has genotype G) is equal to
>
> Pr(O survives / O has phenotype P) but is not equal to
>
> Pr(O survives / O has genotype G).

('Pr(E / A and B)' is to be read as "the probability of E given A and B.")

What this says is that the influence of the phenotypic properties upon the probability of the organism's survival is indifferent to the presence or absence of the specific genotypic properties. The opposite, however, is not the case: the effect of genotypic properties upon organismic survival only acts through the phenotype's effect. Some genotypes are more successful at replicating than are

others, but this difference in degree of success is due to the different degrees of success in which the organisms reproduce. I.e., genotypic success in replication is due to phenotypic success in surviving to reproductive maturity. Phenotypic differences directly causally determine differences in reproductive success, and genotypic differences only do so via the phenotypic differences. Hence, the phenotypic level of explanation provides a superior explanation of natural selection and hence of adaptations.

The point here simply is that proximate causes screen-off remote causes from their effects. Phenotypic properties interact more directly with the environment than do genotypic ones. In virtue of this, Brandon claims that phenotypic causes afford superior causal explanations of natural selection than do the more remote genotypic causes. Brandon concludes that explaining a trait in virtue of natural selection is to appeal to the phenotypic properties of the organism's ancestors and not to their genotypic properties. I.e., genotypes cause phenotypes, and phenotypes cause phenotypes; but phenotypes cause phenotypes more directly than do genotypes. Hence, Brandon concludes, appeal to the phenotypic level of description affords a better explanation of organismic adaptations.

However, the argument in its current form is unacceptable. It assumes the same nonpragmatic view of

explanation implicit in Kim's argument for local reductionism and in the low-level preference argument for eliminative materialism. That is to say, it assumes that accounts which provide more specific causal information are always explanatorily superior to those providing less specific information. That is the rationale behind claiming that proximal causes are always more explanatorily efficacious than more distal ones. For knowledge of the screening-off cause allows for more certain predictive power or knowledge of the effect than does the screened-off cause alone.

Elliot Sober has, in fact, pointed to this nonpragmatic assumption as a criticism of Brandon's argument.[16] Sober provides a striking example illustrating the implausibility of this nonpragmatic view of explanation. Imagine a causal chain beginning with someone dialing a telephone number, some other person's telephone ringing as a result, and that other person subsequently answering their telephone as a further result. Let us suppose that the explanandum is the answering of the telephone. Now, the ringing screens off the dialing from the explanandum, i.e. were the ringing to have occurred without the dialing, the answering would still have taken place. Sober asks rhetorically, "If someone wants to know why you answered your phone, is it really

---

[16] Elliot Sober, "Screening-Off and the Units of Selection", Philosophy of Science 59 (1992) 142-52. See pp. 148-9.

more explanatory to point out that your phone rang than to point out that I dialed your number? Once we disentangle the predictive power of a factor from its explanatory power, it is possible to doubt that screened-off factors are always less explanatory than the factors that screen them off." Here, Sober makes a point quite similar to the point I made about the pragmatics of explanation in Chapter Two, viz. that more specific causal information is only explanatorily superior when one is primarily concerned with predictive accuracy. Otherwise, it might actually be explanatorily inferior to the less causally specific account.

Given that Brandon's argument is unsound, does this mean that we are still uncertain as to whether phenotypes can be viewed as the units of selection? No, because Brandon's argument is not completely misguided. The argument can be repaired by reframing it in a conditional form. For it does nicely illustrate that there are some contexts in which the explanatory level is the phenotypic level, viz. those contexts in which predictive accuracy is of greater importance, as well as those contexts in which one desires to explain a trait appearing in a wide range of genetically distinct organisms. Contrary to Brandon's intentions, it is still left open as a possibility that there are some contexts (some audiences) for which the genetic level is the truly explanatory one. But this ecumenism is perfectly agreeable to me. All that matters is

that it is indeed an option to remain at the phenotypic level in explaining an adaptation. The considerations which Brandon raises clearly support this more modest conclusion. Hence, in providing an evolutionary explanation of an adaptation, it is at least an option to remain at the phenotypic level of description. This prevents us from falling into the pitfall of having to view the adaptationist explanation of psychology as being necessarily a microreduction of it.

It seems plausible to conclude that there are some contexts in which appealing to the "gene's-eye view" is permissible in devising an adaptationist explanation. The gene's-eye view is at least an option when one is not attempting to explain an adaptive trait as it is found in a wide number of species but simply as it is found in one species or even in one individual organism. In such a case as that, the cognitive constraint on explanation does not stop us from going down to the genic level of description. We are not required to consider all the possible genotypes which an organism having that adaptation can possess. However, when we wish to explain an adaptation in its full generality, i.e. as it is possessed by a wide range of microstructurally diverse species, then our adaptationist explanation must remain at the phenotypic level. What is important for present purposes is that the phenotypic level remain open as an option. Since the gene-centric view

itself is nothing more than an option, then the phenotypic level of explanatory description also remains an option.