QUALIA AND MENTAL CAUSATION IN A PHYSICAL WORLD

Themes from the Philosophy of Jaegwon Kim

EDITED BY TERENCE HORGAN,
MARCELO SABATÉS, AND DAVID SOSA

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107077836

© Cambridge University Press 2015

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2015

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-07783-6 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

List of contributors *page* vii
Preface ix

1 Reality and reduction: What's really at stake in the causal exclusion debate 1
Louise M. Antony

2 Two property theories and the causal conundrum for physicalism 25
Frank Jackson

3 Mental causation: The free lunch 40
Barry Loewer

4 Does mental causation require psychophysical identities? 64
Brian P. McLaughlin

5 The Canberra Plan neglects ground 105
Ned Block

6 Microrealization and the mental 134
Sydney Shoemaker

7 Supervenience and the causal explanation of behavior 154
Fred Dretske

8 Visual awareness and visual qualia 167
Christopher Hill

9 Phenomenal externalism, Lolita, and the planet Xenon 190
Michael Tye

10 Troubles for radical transparency 209
James Van Cleve

11 How theories work: Open questions for methodological philosophy of science 231
Lawrence Sklar

References 252
Index 262

Contributors

LOUISE M. ANTONY is Professor of Philosophy at the University of Massachusetts, Amherst.

NED BLOCK is Silver Professor of Philosophy, Psychology, and Neural Science at New York University.

FRED DRETSKE was, at the time of his passing in 2013, Senior Research Scholar in Philosophy at Duke University.

CHRISTOPHER HILL is Professor of Philosophy at Brown University.

FRANK JACKSON is Distinguished Professor in the School of Philosophy at the Australian National University and Visiting Professor of Philosophy at Princeton University.

BARRY LOEWER is Distinguished Professor of Philosophy and Director of the Rutgers Center for Philosophy of Science at Rutgers University.

BRIAN P. MCLAUGHLIN is Distinguished Professor of Philosophy and Cognitive Science at Rutgers University.

SYDNEY SHOEMAKER is Susan Linn Sage Professor Emeritus of Philosophy at Cornell University.

LAWRENCE SKLAR is Carl G. Hempel and William K. Frankena Distinguished University Professor of Philosophy at University of Michigan, Ann Arbor.

MICHAEL TYE is the Dallas TACA Centennial Professor in Liberal Arts at the University of Texas, Austin.

JAMES VAN CLEVE is Professor of Philosophy at the University of Southern California.

> Without the widowing of Xanthippe, the subsequent cooling of Socrates's body would not have occurred. (For in that case he would not have died when he did.) . . . [This] is [not] a genuine case of causal dependence. Instantaneous . . . causation [is] not so very easy! (Lewis, 1986b, p. 263)

Of course, all functional properties are extrinsic properties, and if mental externalism is true for certain mental properties, then the functional properties with which the NRP theorist wishes to identify those mental properties will be very highly extrinsic.

The idea that there are functional events in the role-functionalist sense does not jibe with Lewis's counterfactual theory of causation. Lewis, as we saw earlier, did not countenance functional events in that sense. The preceding discussion might well indicate reasons in addition to those that he explicitly gives when rejecting the idea that events have functional essences. I myself do not take such functional events to pose a problem for the thesis that a counterfactual dependency between distinct (in Lewis's sense) events suffices for causation. I think, rather, that a proponent of that thesis should follow Lewis in denying that there are functional events in the NRP theorist's sense.

Obviously, I have not established that there are no functional events in the role-functionalist sense. But I hope to have provided some reasons to believe that there are no such events, that functional properties are not constitutive properties of events, that events do not have functional essences, that quantification is not an event-forming operation. The reasons cannot be defeated simply by eschewing the idea that there is "causal oomph." Intuitions about "oomphish" causation play no role in the preceding considerations. At the very least, I have posed a challenge to NRP theorists: make a case that quantification is an event-forming operation and that we should countenance events with functional constitutive properties. They have yet to discharge that dialectical obligation. Like Kim, I am deeply skeptical that they can. Even if quantification is a property-forming operation (a matter left open here), it remains for NRP theorists to make a case that it is an event-forming operation.

CHAPTER 5

The Canberra Plan neglects ground

Ned Block

I. Introduction

According to the Canberra Plan, the first step in a reductive physicalist enterprise is to functionally define the property to be reduced, and the second step is to find the physical property that fills that functional role. Reductive physicalism is true for the mind if both steps can always be carried out for mental properties. This picture of what reductive physicalism is stems from J. J. C. Smart's (1959) "topic-neutral" analyses and has been advocated in one form or another by Armstrong, Chalmers, Jackson, Kim, Levine, and Lewis, even though these figures differ from one another on whether they are proponents or opponents of reductive physicalism. Smart's 1959 article is also the source of a different and perhaps incompatible picture, the mind-body identity view, according to which reductive physicalism about the mind should be modeled on "theoretical identities" such as light = electromagnetic radiation (of wavelength 400–700 nm). This chapter will argue that the point of view of the Canberra Plan neglects ground. I will consider a few attempts to graft an account of the physical/functional ground of mind onto the Canberra Plan, arguing that such attempts lead nowhere.

Terminological note: my main point is that reductive physicalism requires that, for any phenomenological similarity between two mental states, that similarity must hold in virtue of a physical similarity that explains or constitutes the phenomenal similarity; that is, the fact that there is a phenomenal similarity has as its ground the fact that there is a physical similarity. This terminology, in which variants of *ground*, *in virtue of*, *explains* (understood metaphysically rather than epistemically) are used to connect sentences, fits the way of speaking of ground in the work of Kit Fine. However, I also use an abbreviatory device not used in Fine's papers in which I speak of a phenomenal similarity as grounded in a physical

similarity. That is, I talk of properties being grounded in other properties as well as of facts being grounded in other facts.

A second item of terminology is that I use the terms *second order* and *first order* not in the sense of properties of properties versus properties of individuals, but as follows: a second-order property is a property that has a true definition in terms of having some other properties that meet a certain sort of condition. And a first-order property does not have such a true definition. A functional property is a special kind of second-order property in which the definition specifies causal relations to other properties and to inputs and outputs. I am supposing that the "other" properties quantified over in the second-order definition are themselves first-order properties.

II. The Canberra Plan

I start by describing some of the views that neglect ground.

Jaegwon Kim, in his landmark works on mental causation and reduction of mind (Kim, 1972, 1993a, 1998b, 2005), argues for a model of reductive physicalism as functional reduction. The first step in reducing water to H_2O, light to electromagnetic radiation of 400–700 nm, or the property of being a gene to molecular aspects of DNA is to functionally define the property to be reduced. For example, the property of being a gene might be defined in terms of its role in encoding and transmitting genetic information (Kim, 1998b, p. 25; 2005, pp. 101–2). The next step is to find the realizers of the functional role that has been defined. For example, DNA molecules encode and transmit genetic information.[1] The third stage is explaining the mechanism by which the realizer accomplishes that function, for example, how the DNA molecule actually does the job of encoding and transmitting.

Joe Levine (1993, p. 132) holds that: "Stage 1 involves the (relatively? Quasi?) a priori process of working the concept of the property to be reduced" into shape "for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are."[2]

[1] As has often been pointed out, theories at the "upper level" are often incompatible with the theories of the realizers, so this description is highly idealized. There is a good deal of disagreement in the literature about the consequence of this fact for reductionism. See Godfrey-Smith (2000), Hull (1974), Kim (1993a), and Schaffner (1969).

[2] See p. 551 of the version of Levine's paper reprinted in Block, Flanagan, and Güzeldere (1997). These ideas are discussed in Block and Stalnaker (1999).

Frank Jackson (1994, 1998b) tells a similar story. He sketches (1998b, p. 59) the following argument for the reduction of temperature to mean molecular kinetic energy:

Pr.1 [NB: premise 1]: Temperature in gases = that which plays the temperature . . . role in gases. (Conceptual claim)

Pr.2: That which plays the temperature role in gases = mean molecular kinetic energy. (Empirical discovery)

Conc.: Temperature in gases = mean molecular kinetic energy. (Transitivity of '=')

Jackson notes that premise 1, the conceptual analysis in functional terms, can be thought of in either of two ways: as a claim of synonymy or as capturing an a priori reference-fixing claim.

These views share some crucial features with the "Canberra Plan" movement of J. J. C. Smart, David Armstrong, and David Lewis (1966, 1970, 1972). Lewis held that the meanings of mental state terms can be analyzed via definitions of the following form: the state with causal role *R*. If mental state *M* can be seen via a priori analysis to be the state with causal role *R*, and if brain state *B* is found empirically to have causal role *R*, it follows that $M = B$. Lewis regarded supposed identities such as "pain = C-fiber stimulation" as contingent rather than necessary. The idea is that the term *pain* is a (nonrigid) definite description that picks out the contextually indicated property that occupies causal role *R*. One physical property might be picked out in the context of human pain, another in the context of octopus pain. Note also that Lewis regarded these identities as "type-type," that is, as identifying the property pain with the property C-fiber stimulation, and not as identifying this particular pain with this particular instance of C-fiber stimulation.[3]

Since he held that (assuming brain state *B* has causal role *R*) $M = B$, Lewis is often said to be a physicalist – including by Lewis himself (prior to "Mad Pain and Martian Pain" [1980], which advocates a mixture of functionalism and physicalism). And since he accepted a priori causal role analyses of mental state terms, he is often considered a functionalist. Some

[3] This combination of views makes it a bit obscure what we are supposed to make of generics such as "pain is distracting" in a context that doesn't single out any particular kind of creature. Brian Weatherson's Lewisblog from April 2006 has an extended discussion of this issue (in which the point I just made is attributed to Eric Hiddleston). As is noted there, "the President of the US," while often used to pick out the current president, can also be used generically, as in "the President of the US lives in the White House," so perhaps whatever solution applies to the latter case can also be used to help Lewis out.

reserve the term *functionalist* for those who identify mental states with their causal roles, and on that definition Lewis is not a functionalist since he identified mental states with the realizers of those roles, not the roles themselves.

David Chalmers (2012, p. 362) describes his view as "at least a close relative of the Canberra Plan," even though he is skeptical of the explicit definition aspect of the view. However, he has in the past endorsed something that sounds very Canberrish, for example, here:

> For the most interesting phenomena that require explanation, including phenomena such as reproduction and learning, the relevant notions can usually be analyzed functionally. The core of such notions can be characterized in terms of the performance of some function or functions (where "function" is taken causally rather than teleologically), or in terms of the capacity to perform those functions. It follows that once we have explained how those functions are performed, then we have explained the phenomenon in question. Once we have explained how an organism performs the function of producing another organism, we have explained reproduction, for all it means to reproduce is to perform that function. (Chalmers, 1996, p. 43)[4]

Although Chalmers is skeptical about explicit definitions, his "scrutability" framework shares the features of the Canberra Plan that I will be criticizing. The key similarity with the Canberra Plan is that reductive accounts are always accounts of determination by the reductive base without consideration of the ground of similarities in cases in which similar facts are determined by different reductive bases. In other words, Chalmers's vision of the reductive physicalism that he rejects does not require that phenomenological similarities with different scrutability bases be explained by physical similarities in the scrutability bases.

The Canberra Plan as I have been construing it is functionalist in that mental states are analyzed functionally in terms of their causal role. And it is physicalist in that mental states are said to be the physical occupants of these roles.

Kim (1998b) has given what may seem to be a direct argument for the identity of functional properties with physical properties. Kim presents the functional model of reduction as follows:

> To recapitulate: to reduce a property M to a domain of base properties, we must first "prime" M for reduction by construing, or reconstruing, *it relationally or extrinsically*. This turns M into a relational/extrinsic property. For functional reduction, we construe M as a second-order property defined by its causal role – that is, by a causal specification H describing its (typical) causes and effects. So M is now the property of having a property with such and such causal potentials, and it turns out that property P is exactly the property that fits the causal specification. And this grounds the identification of M with P. M is the property of having some property that meets specification H, and P is the property that meets H. So M is the property of having p. But in general the property of having property Q = property Q. (pp. 98–9)

[4] Kim quotes this passage as well.

To say that *M* is the property of having some property that meets specification *H* is to say that *M* is a second-order property in the sense used here. A first-order property in the sense used here is one that does not have a true characterization in terms of having some other properties that meet a certain sort of condition. So, obviously, a second-order property cannot be identical to a first-order property. A functional property can be thought of as a special case of second-order property that is constituted by the having of some other properties that have certain causal relations to one another and to inputs and outputs.[5] The "other" properties are known as the realizers of the functional properties. When Kim says that *P* is the property that meets the specification *H*, he is saying that *P* is the realizer of *H*, and of course realizers can be (and are assumed here to be) first order.[6] So, it may seem that Kim is arguing for a straightforward contradiction: that a second-order property is identical to its first-order realizer.

Later, I describe more of the context surrounding Kim's argument, which reveals that the natural interpretation just given is not the right one. The passage is misleading, as I will explain.

I have been describing armchair philosophical views that purport to be versions of reductive physicalism. However, there are more science-based versions of these views that seem to suggest grounding the mind in functional or computational properties. Recent neuroscience is strongly computational, and a computational view of the mind is often seen as a version of a functionalist view of the mind. So, a view of the mind in neuroscientific terms would seem to be both physicalistic and functionalistic. Further, it is often said that all science is functional. Lewis (1970) held that all terms of science should be defined functionally. Daniel Dennett (2001) says that functionalism is true generally, for all of science, and that the most general functional descriptions are at the level of physics:

[5] See Kim (1998b, p. 20). I have sometimes defined a functional property as a property that is functional in the sense in the text and that in addition involves causal relations to inputs and outputs. Of course there is no issue of fact as between these definitions.

[6] A suggestion to the contrary is made in Block (1990).

> Functionalism is the idea enshrined in the old proverb: handsome is as handsome does. Matter matters only because of what matter can do. Functionalism in this broadest sense is so ubiquitous in science that it is tantamount to a reigning presumption of all of science. (p. 233)

Physics is of course a science and also physicalist, at least about the entities with which it is concerned, so according to the point of view just mentioned, physics is both physicalist and functionalist.

The mental includes – at least – states, events, processes, entities, and properties. In the way of looking at the mind-body problem that I will be promoting, properties are key, and so I focus on them.

Why are properties important? The main reason is that from a physicalist perspective, phenomenal similarities must be grounded in physical similarities. Similarities are just shared properties, so of course properties are important.

Also, properties are the locus of an important issue concerning causation. The Queen of the Night sings "Der Hölle Rache kocht in meinem Herzen," which shatters the glass. But her words cause the glass to shatter in virtue of their volume and frequency rather than in virtue of their semantic properties (Dretske, 1989; Sosa, 1984). More generally, when one event causes another, some properties of the cause are causally efficacious in respect to (certain properties of) the effect, and others are not. Since functionalism is a causal thesis, there is good reason for a discussion of the functionalist approach to the mind-body problem to pay attention to properties. For simplicity, I take state types to be properties, albeit temporarily instantiated properties, and I think of an event (which I won't discuss much) as a thing's having a property at a time.

III. Brief refresher course

This section provides some elementary exposition on what functionalism is. Readers who are familiar with this material might still read the last two paragraphs of the section.[7]

Suppose we have a theory of mental states that specifies all the causal relations among the mental states, sensory inputs, and behavioral outputs. Focusing on pain as a sample mental state, it might say among other things that sitting on a tack causes pain and that pain causes anxiety and the pain

[7] Readers who wish to see a longer exposition could look at Block (1997b). A somewhat revised version is available at http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.pdf and in volume 1 of my collected papers (Block, 2007a).

jointly with the anxiety cause saying "ouch." Let us agree for the sake of the example to go along with this silly theory. Functionalism would then say that we could define *pain* as follows: being in pain = being in *some* state, which is caused by sitting on tacks, and which in turn causes *some* other state, and the two states jointly cause "ouch." The two "somes" indicate existential quantification, which makes the definition second order, not in the sense of a property of properties, but in the sense of a property that has a true definition in terms of having some other properties that meet a certain condition. Making the quantification over states more explicit:

Being in pain = Being an x such that x is in pain = Being an x such that $\exists P \exists Q$ *(being stuck by a tack causes P & P causes Q and P and Q jointly cause emitting "ouch" & x is in P).*[8]

More generally, if T is a psychological theory with n mental terms of which the 17th is "pain," we can define "pain" relative to T as follows – the $F_1 \ldots F_n$ are variables that replace the n mental terms; and i_1, etc., are the input terms (such as "being stuck by a tack"); and o_1, etc., are the output terms (such as "emitting 'ouch'"):

Being in pain = Being an x such that x is in pain = Being an x such that $\exists F_1 \ldots \exists F_n$ *[T(*$F_1 \ldots F_n$*, i_1, etc., o_1, etc.) & x is in* F_{17}*]*

In this way, functionalism characterizes the mental in nonmental terms, that is, in terms that involve quantification over realizations of mental states but no explicit mention of them; thus functionalism characterizes the mental in terms of structures that are tacked down to reality only at the inputs and outputs. In characterizing the mental in nonmental terms, functionalism gains what has been seen as a benefit of behaviorism while nonetheless acknowledging mental states by quantifying over their realizations, and thereby improving on behaviorism.

It is often easier to think about the relation between first- and second-order properties by using an example of a simple disposition, for example, dormitivity. Dormitivity can be construed as a second-order property, the property constituted by the having of some first-order property or other that causes sleep. Of course, one could equally well construe dormitivity as a first-order property, the property of just causing sleep. But now that I have acknowledged that there is a first-order construal of dormitivity, the

[8] The symbol $\exists$ stands for "there is." $\exists xFx$ means there is something that is F. $\exists P \exists Q$ [sitting on a tack causes P & P causes Q] can be read as: there are two properties such that sitting on a tack causes one of them and it causes the other.

reader may wonder what the difference is and why anyone would construe dormitivity as a second-order property. In the first-order construal, the property F is dormitive just in case F causes sleep. But if we want to ascribe dormitivity to dormitive *things*, for example, pills, we have to use the second-order sense. What it is for a pill to be dormitive is for it, the pill, to *have such an F that causes sleep, that is, what it is for the pill to be dormitive is for it to have some property or other* that causes sleep. That is, x is a dormitive pill if and only if $\exists G$ (G causes sleep & x has G) – or putting this so as to eliminate the free variable, dormitivity in the sense in which it applies to pills is the property of being an x such that $\exists G$ (G causes sleep & x has G). That is, dormitivity = $(\lambda x)(\exists G$ [G causes sleep & x has G]).

(We could also think of a pill as dormitive just in case it, the pill, causes sleep, but recall that I mentioned at the outset that I would focus on properties both for metaphysical purposes and because properties are important in causation, so this construal is not relevant to the purpose at hand.) The homes of the two construals are in application to different types of items.

This point applies straightforwardly to the functionalist perspective on mentality. If we want a functional definition of mental property terms that apply to properties, the first-order variant will do. For example, the pain-property can be thought of as the property of jointly causing certain outputs together with certain other (mental) properties, being caused by certain inputs, and so forth. But if we want to ascribe those properties to people, we need second-order properties. What it is for a person to have pain, according to the functionalist, is for the person to *have some property or other* that has certain causal relations to other (mental) properties and to inputs and outputs.

IV. Can a second-order property be a first-order property?

A second-order property is one that has a true characterization in terms of having some other properties that meet a certain condition. A first-order property is one that does not have such a true characterization. So, it is just a contradiction to claim that a second-order property is a physical property. Why do some appear to think otherwise?

Dormitivity – construed as a second-order property – has first-order chemical-realizing properties such as (having the) structure $C_{12}H_{12}N_2O_3$ (phenobarbital) that causes sleep. What is the relation between a second-order property and the disjunction of its first-order realizers if not identity? By the disjunction I mean the property that consists in being $C_{12}H_{12}N_2O_3$ (phenobarbital) or in being $C_{16}H_{13}ClN_2O$ (diazepam), or... But what

does the "..." *mean*? It seems that the "..." means something like: "or some other first-order properties that cause sleep." But the "some" reveals that the supposed first-order disjunctive property is *really second order*. Note that the identity claim amounts to something like this: the property constituted by being an x such that $\exists F$ (F causes sleep & x has F) = the property constituted by being an x such that (x has P and P causes sleep) or (x has Q and Q causes sleep) or x has *some* other property (maybe more than one) F such that x has F and F causes sleep. Again, the "some" shows us that the property expressed is second order.

But perhaps we can do without that clause with the "some" in it? We can just list the disjuncts. Suppose that there are exactly two first-order dormitive structures as a matter of physical law, $C_{12}H_{12}N_2O_3$ and $C_{16}H_{13}ClN_2O$. However, if there were another first-order property that caused sleep, it would be dormitive, according to the second-order definition, without being one of the first-order disjuncts. And a similar point holds for the infinite disjunction. So, there is a modal difference.

Even without a modal difference, the hyperintensionality of grounding leads to a similar conclusion. The existence of Socrates is the ground of the existence of the singleton of Socrates rather than vice versa (Fine, 2012) despite these facts obtaining in exactly the same worlds. The application is this: even if somehow a first-order physical property played the pain functional role in every possible world and nothing else played that role in any possible world, there would still be a question of whether the obtaining of the physical property was the ground of the obtaining of the mental property. That is, even if the mental and physical properties are coextensive across all possible worlds, the question still arises as to whether the mental property is instantiated in virtue of the instantiation of the physical property. I give an example toward the end of the chapter – in which the physical property is indexical – that should make this point vivid.

The reader may feel that the fact that a second-order property cannot be identical to a first-order property can mislead us with regard to physicalism, for the second-order property can itself be seen as in effect first-order physical *so long as all its realizers are first-order physical*. (In addition, it would have to be stipulated that there are no extra nonphysical "ghost" mental properties.) I explain the inadequacy of this view in the next section.

V. Metaphysics, ontology, and disjunctive ground

I use the term *metaphysics* to mean the study of ground, and I use the term *ontology* to concern what types of things exist. My use of the term *ontology* derives from Quine (1948). Quine and subsequent discussions influenced

by him speak of a person's or of a theory's ontological commitments, meaning commitments on what types of things exist. Sample ontological disagreements concern whether universals or souls or absolute space exists. The ontological issue of what it is like to experience pain is whether in adopting an ontological commitment to the experience of pain we adopt an ontological commitment to anything immaterial. An ontological physicalist says no. A metaphysical physicalist, by contrast, claims that the ground of the experience of pain – what all experiences of pains have in common in virtue of which they are experiences of pains – is a physical property that explains the experiential commonality of experiences of pains.

Metaphysical physicalism could fail even if ontological physicalism is true. The phenomenal commonalities between different pain-feeling creatures could fail to have a physical ground even without any immaterial souls. Metaphysical physicalism could be true even if ontological physicalism is false if, for example, our material minds have an immaterial adjunct that is part of a communication network with angels rather than part of the ground of our mental properties.

Dualism and physicalism are naturally understood as both ontological and metaphysical theses, but functionalism can be a metaphysical thesis without being an ontological thesis. Let me explain. Dualism and physicalism disagree on whether there is anything immaterial. But functionalism is compatible with both ontological positions because functionalism takes no stand on the occupant of the functional roles that define it. Functionalism can say that the ground of pain, what makes two pains both pains is a common functional role. This is a metaphysical, not an ontological, doctrine. Pains could have that functional role whether or not they involve nonphysical substances or properties, so long as the nonphysical substances or properties are causally efficacious in regard to other states, inputs, and outputs in the right ways and can be causally affected in the right ways.

Suppose there are souls in some adding machines that make them work. Still, the ground of something being an adding machine – to the extent that one can speak of something so nominal as a ground at all – is that adding machines' states function so as to add. If the soul stuff can function in this way, it doesn't make the metaphysical nature of adding in any way nonfunctional. There is nothing about the function that constitutes adding that requires a material basis. Similarly, a metaphysical functionalist should say that the existence of souls (ontological dualism) need not be relevant to the metaphysical issue of what grounds pain – the answer could be functional just as with adding.

As I mentioned, the physicalism of David Lewis (1966, 1970, 1972) derives from the idea that pain can be defined, a priori, on the basis of its causal role, *R*. Brain state *B in us* has *R* as a matter of fact, so pain = *B*. Suppose further that there are no nonphysical pains and that there never have been and never will be any nonphysical pains (Lewis, 1994). The result is a kind of ontological physicalism, but note that it does not amount to a metaphysical physicalism that grounds mentality in the physical. What is common to Martian pains, if there are any, and octopus and human pains that grounds the fact that they are both pains is not anything physical, on Lewis's perspective, but rather the fact that they are all instantiations of causal role *R*. So, on the metaphysical question of what grounds mind, Lewis (1966, 1970, 1972) should be seen as having no view, or perhaps as being a functionalist rather than a physicalist. In his preferred regimentation (not including Lewis, 1980, in which he adopts the weird "mixed" theory that I will be getting to), pain is physical, but having pain is identical to a second-order (functional) property, namely, the property of having some state or other that plays causal role *R*. A context-relative definite description of the form "the state that has causal role *R*" picks out one physical state in us, another in Martians, and so on. But "having pain" (on Lewis's regimentation) is not a context-relative designator but rather a rigid designator that always picks out the same second-order property, namely, the property of having some realizer or other that satisfies causal role *R*.

The upshot is that Lewis is an ontological physicalist and, to the extent that he had any metaphysics of mind at all, a metaphysical functionalist.

Note, incidentally, a point emphasized by Kim (1972), that the physical basis of pain can be sufficiently abstract so as to be shared by humans and octopi just as two physically very different substances can have the same temperature.

The story I have to tell about the views of Chalmers (1996) and Jackson (1994, 1998a) is much like that for Lewis – the physicalism that they are mainly concerned with (and that they are – were, in Jackson's case – mainly concerned to oppose) is ontological physicalism.

I can explain vividly the difference between metaphysical physicalism and ontological physicalism by reference to a fictional character, Commander Data (Block, 2002, 2007b). The TV series *Star Trek: The Next Generation* (26 February 1989) includes an episode ("The Measure of a Man") in which there is a trial to decide whether a human-like android, Commander Data, may legally be turned off and taken apart by someone who does not know whether he can put the parts together again. (The

technology that allowed the android to be built has been lost.) Let us think of Commander Data as defined as a merely superficial functional isomorph of us. A superficial isomorph of us is isomorphic to us in causal relations among mental states, inputs, and outputs to the extent that those causal relations are part of commonsense psychology. (That is, for every human mental state, input, and output acknowledged by common sense, there is a corresponding state – maybe mental, maybe not – input, and output of Commander Data; and for every causal relation among our states, inputs, and outputs that is acknowledged by common sense, there is a corresponding causal relation among Commander Data's states, inputs, and outputs – and conversely. One consequence is that Commander Data behaves just as we do as far as we can tell from the standpoint of commonsense psychology.)

Commander Data is a *merely* superficial isomorph of us. That means that he is not like us in physical realization of the superficial functional states he shares with us except to the extent that shared properties of physical realizations are required by superficial functional isomorphism. And Commander Data is not like us in detailed functional states, for example, functional states that involve functionalized psychology or neuroscience. For example, we may learn that conscious pain makes sounds appear to have higher pitch, but we cannot expect Commander Data to show that effect. We can assume that the only functional properties we share with Commander Data are the superficial ones mentioned earlier and that there are no shared physical properties that can explain any shared phenomenality without attributing phenomenality to things that don't have it. (I mean no shared first-order physical properties. Of course there is a shared functional property ensured by the superficial isomorphism, and I am also excluding heterogeneously disjunctive physical properties.) So, he is like us superficially, but not in any deep property that can plausibly be one that scientists will one day tell us is the physical ground of consciousness.

Suppose that – as seems conceivable – Commander Data is conscious. For vividness, suppose that Commander Data is exactly like us, phenomenologically speaking. That supposition leads immediately to metaphysical dualism about phenomenology. The case has been set up so as to preclude any substantive physical similarity between Commander Data and us that can ground the postulated phenomenal similarity between him and us. That is, a physicalist metaphysical account has to ground the phenomenology we share with Commander Data in a physical property that we share with Commander Data, but it can't. For, by hypothesis, he

does not share anything physical with us of the sort that a physicalist could appeal to.[9]

Further, the same point applies to psychofunctional or neurofunctional accounts, that is, accounts that appeal to detailed empirically oriented functional properties as solutions to the metaphysical mind-body problem.[10]

Here is the point: reductive physicalism of the sort that Lewis accepts and that Chalmers and Jackson most directly oppose is not troubled by the case of Commander Data, even if Commander Data is phenomenologically just like us. So long as the complete microphysical story (the complete nonmental part of Chalmers's scrutability basis) serves to entail that we are conscious and that Commander Data is conscious, the physicalism of the sort that Lewis holds and Chalmers and Jackson most directly opposed (and that Jackson has changed his mind on) is satisfied. So, there is a key question that that kind of reductive physicalism – ontological physicalism – does not ask or answer: what is it that creatures with the same phenomenology share that grounds that phenomenology? In sum, the kind of reductive physicalism acknowledged by the Canberra Plan is blind to the dualistic implications of Commander Data, so that version of reductive physicalism is inadequate.

Of course the mental could be claimed to be grounded in the commonsense *concepts* whose supposed a priori grasp provides the superficial functional organization that we share with Commander Data. However, that superficial functional organization would seem a merely nominal similarity between Data and us and so a poor candidate for any kind of substantive grounding claim.

Another suggestion would be "superficialism," the view that there is a substantive basis of mentality that – perhaps coincidentally – is the same as what is part of common sense or that we a priori grasp. I suppose that someone might think that we have the mental concepts we do because we have some sort of mental pipeline to the actual substantive nature of mentality. Let us return to the supposition that conscious pain raises the perceived pitch of sounds. That would be part of the functional role of pain, but not part of the superficialist functional role, since the fact that conscious pain raises pitch is not known to common sense. But why

[9] See papers by Jakob Hohwy (2004) and Brian McLaughlin (2003) for a different point of view.

[10] Recall that a functionalist characterizes pain as follows: if T is a psychological theory with n mental terms of which the 17th is "pain," we can define "pain" relative to T as follows (the $F_1 \ldots F_n$ are variables that replace the n mental terms; and i_1, etc., are the inputs; and o_1, etc., are outputs): Being in pain = Being an x such that $\exists F_1 \ldots \exists F_n$ [$T(F_1 \ldots F_n, i_1$, etc., o_1, etc.) & x is in F_{17}]. Psychofunctionalism is a version of functionalism in which T is a theory of empirical psychology. (Psychofunctionalism was introduced in Block, 1978b.)

should pain be grounded in aspects of functional role that happen to be known to common sense, but not in other aspects of functional role? The mysterious pipeline hypothesis has an answer, but that speculative answer reminds us that the superficialist grounding thesis raises an explanatory problem that is not independently puzzling (Block, 2002, 2007b). The new explanatory problem is: how could there be such a pipeline? Of course these considerations do not show superficialism is false, but they do put the burden of proof on the superficialist to show how to avoid the problem I raised.

What about a ground of consciousness in a heterogeneously disjunctive property in which the disjuncts are Data's realization of the shared functional property and ours? Suppose a physicalist says that the explanation of the fact that my pain feels the same as yours is that your realization and my realization of pain are both part of a heterogeneously disjunctive realization; the explanation is that you instantiate your disjunct and I instantiate my disjunct. But that is to give no physicalistic explanation at all (Block, 2002). So, again one wonders whether the putative ground is a ground.

The familiar jade analogy might be helpful. The functional role associated with "jade" picks out one physical substance (jadeite) in one circumstance, another physical substance (nephrite) in another circumstance. But even if all the realizers of the jade role are physical, that does not establish a physical ground for jade.

I now move to discussing how these points interact with Kim's and Lewis's views.

As I mentioned, Kim (1998b) says something that sounds a lot like the contradictory identification of a second-order property with a first-order property. He says, "M is the property of having some property that meets specification H, and P is the property that meets H. So M is the property of having P. But in general the property of having property Q = property Q." He concludes that $M = P$. However, in a footnote to this passage (132) he says, "How could roles be identical with their occupants?" Later, when he explains what he means, it turns out that he is not interpreting M as a rigid designator (as I have been interpreting it) but rather as a definite description. What he is saying is that a functionally "specified" property is identical to a first-order property. A functionally specified property is one that is picked out by a functional definite description of the form "the occupant of causal role R." So, the claim that $M = P_1$ in one species but $M = P_2$ in another species doesn't mean what it seems to mean. It just means that the occupant of the M role is one thing in one species and the occupant of the M role is another thing in another species. In terms of

the dormitivity example, his point could be put like this: the (contextually indicated) property that causes sleep = $C_{12}H_{12}N_2O_3$. And that is not the claim that a second-order property is identical to a first-order property.[11]

Kim (2005, p. 111) gives the following as a reductive explanation of x's having M at time t:

1. x has P_i at t.
2. P_i satisfies causal role C (in systems like x).
3. Having $M =_{\text{def}}$ having some property satisfying causal role C.
4. Therefore, x has M at t.

In step 3, Kim is explicit about his functionalism. He is a functionalist about those mental states that can be functionally defined and a dualist about those phenomenal mental states that cannot be functionally defined.

A second point about this passage illustrates the fact that Kim is not a metaphysical physicalist. Kim's paradigm of reductive explanation – as early as Kim (1972) – is one that relativizes to "systems like x." He is not arguing that there is any physical state in common to pain-feeling organisms that grounds their being in pain. Kim (1998b) makes clear that the only physicalistic mind-body identities he accepts are structure restricted – that is, restricted to specific realizations of the functional organization that defines the mental (to the extent that the mental can be functionally defined).

Let me return to the peculiar "mixed" functionalist/physicalist theory of Lewis (1980). Lewis starts with the "opinion" that both Martians and Madmen have pain. Martians are functionally just like us (at a superficial level) but are physically very unlike us. Instead of the neural basis of pain that we have, Martians have smallish inflatable cavities throughout their bodies whose inflation plays the functional role of pain. Madmen have our neural realizer of pain, but instead of causing winces and distraction, it causes finger-snapping and thoughts of mathematics. These "opinions" are supposed to show that there is something right about both physicalism and functionalism. Lewis runs with these opinions, building them into an account of pain according to which x is in pain if and only if x is in the

[11] There is more to it than that – he distinguishes between "sparse" and latitudinarian views of properties. On a latitudinarian view of properties, second-order properties cannot be identical to first-order properties. However, on a sparse conception of properties, the question is whether second-order properties have causal powers of their own, or whether there is no more to their causal powers than the causal powers of their realizers. If the latter, then, according to Kim, there is no further fact of the matter as to whether something has a second-order property as compared with its first-order realizer. I believe that second-order properties are indeed causally efficacious, but that issue is beyond the scope of this chapter. See Block (1990).

state that occupies the characteristic causal role of pain for the appropriate population. This account leads to problems of a very weird technical sort – for example, what to say about someone who is Mad, Martian, and different from others in the population. But there is no need to go into these issues here. Lewis says that maybe the Madman is in pain in one sense of the term *pain*, whereas the Martian is in pain in another sense of the term, but he also states unequivocally that the theory is meant to be a theory of the phenomenal character of experience. It is unclear whether what it is like to be the Madman is the same as what it is like to be the Martian. Perhaps Lewis would have rejected interpersonal comparisons of this sort (Stalnaker, 1999).

Lewis's view is certainly hard to swallow, something that I think Lewis was aware of. He thought that with the folk concept of phenomenology, as with many folk concepts, nothing totally satisfies it. The best fit – albeit perhaps not a very good one – is supposed to be given by his mixture of physicalism and functionalism. However, I think the unacceptability of the result does tell us something important about what is wrong with Lewis's methodology, namely, that it is a mistake to put so much weight on "opinions." Lewis should have subjected these opinions to scrutiny and rejected one or both of them. However, for current purposes, what matters is that Lewis's account is neither a form of physicalism nor of functionalism. If the Madman and the Martian are said to have the same phenomenal character, then the view is neither metaphysical functionalist nor metaphysical physicalist, because it ascribes the same phenomenal character to two creatures (the Madman and the Martian) who are neither functionally nor physically identical.

Of course, what is important to Lewis's reductionism is not that shared phenomenal states entail shared reductive states (e.g., physical – or functional – states) but rather a view according to which differences in phenomenal states entail differences in reductive states. That is, Lewis's reductionism is a supervenience thesis. But the claim that the mental supervenes on the physical is not a form of metaphysical physicalism. The property of being an adder supervenes on the physical because if one thing is an adder and another isn't, there must be a physical difference between them. But that is not to say that there is any physical property in common to adders that grounds the fact that they are adders. The property of being a wrong act supervenes on the physical, but that is not to say that there is any physical property in common to wrong acts that grounds their wrongness. Wrongness can supervene on the physical without being grounded in the physical.

It should be noted that one can be a metaphysical physicalist about one kind of mental state and a metaphysical functionalist about another. For example, what it is to be gregarious may be a functional or even behavioral state even if what it is to have a certain phenomenal quality is a neurological state.

My defense of the importance of what I am calling the metaphysical basis of mind is not meant to downgrade the importance of what I am calling the ontological basis of mind: both are important. My point, however, is that even if ontological physicalism is true, if metaphysical physicalism is false, there is an important respect in which the reductive physicalist program has failed.

Kim (1992, 1998b) advocates a reductive physicalist approach to the mind that in effect rejects the metaphysical physicalist point of view. He takes mental properties to be merely nominal properties – indeed, hardly properties at all if one's criterion of reality for properties is causal efficacy. The idea is that the similarities among pain-feeling creatures that grounds their being in pain is not anything deep, but merely that they all instantiate a functional or even behavioral concept. The most fundamental grounding is superficial. He says:

> "Sharp pains administered at random intervals cause anxiety reactions." Suppose this generalization has been well confirmed for humans. Should we expect *on that basis* that it will hold also for Martians whose psychology is implemented (we assume) by a vastly different physical mechanism? . . . The reason the law is true for humans is due to the way the human brain is "wired"; the Martians have a brain with a different wiring plan, and we certainly should not expect the regularity to hold for them just because it does for humans . . . "Pains cause anxiety reactions" may turn out to possess no more unity as a scientific law than does "Jade is green." (Kim, 1992, p. 16)

The assumption that I want to draw attention to is that Kim assumes that pain is a merely nominal property, along the lines that I would construe gregariousness. Kim's version of reductive physicalism (1992, 1998) is close to eliminativism, and of course a reductive physicalist who is an eliminativist about pain does not have to be concerned with the metaphysical grounding of pain. But if instead of pain Kim had applied his nominalizing "functionalizing" technique to the phenomenal quality of my current pain, *Q*, this line of thought would not sound so plausible. If Martians could have states with *Q*, it would not be so plausible that there need be no deep physical property shared by humans and Martians that grounds and explains the phenomenal similarity.

Judging from his (2005), Kim might agree about his earlier view. In the 2005 book, he poses the issue of reducibility starkly, saying, "That a property is functionalizable, that is, it can be defined in terms of causal role – is necessary and sufficient for functional reducibility. It is only when we want to claim that the property has been reduced . . . that we need to have identified its physical realizer" (p. 165). He then goes on to pose the question of whether mental properties are functionalizable. "The answer . . . is yes and no. No for qualitative characters of experience, or 'qualia', and yes, or probably yes, for the rest" (p. 165). No for qualitative characters of experience because of inverted-spectrum issues – it is metaphysically possible for functionally identical states to be different in qualitative character. The overall argument is that reductive physicalism fails for qualia – because they don't fit Kim's picture of reductive physicalism. However, there is another picture of reductive physicalism that has some merit, to which I turn in the next section.

So, Kim departs from the Canberra Plan precisely for mental properties whose substantive nature and need for metaphysical grounding are most obvious, favoring dualism. I think his view here is clearly superior to the views of Lewis, which treat all mental properties as equally lacking in ground.

To sum up, the Canberra Plan does not adequately capture the physicalist reductionist point of view because it neglects ground; that is, it does not involve any sort of metaphysical physicalism.

VI. Theoretical identity, reductive physicalism, and grounding

If we want to know why water = H_2O, freezing = molecular lattice formation, heat = molecular kinetic energy, temperature = mean molecular kinetic energy, and so forth, we have to start with the fact that water, temperature, heat, freezing, and other magnitudes form a family of causally interrelated "macro" properties. This macrofamily is mirrored by a family of "micro" properties: H_2O, mean molecular kinetic energy, molecular kinetic energy, and formation of a lattice of H_2O molecules. (Of course a given level can be micro with respect to one level, macro with respect to another.) The key fact is that the causal and explanatory relations among the macroproperties can be explained if we suppose that the following relations hold between the families: that water = H_2O, temperature = mean molecular kinetic energy, heat = molecular kinetic energy, and freezing = lattice formation. For example, why does decreasing the temperature of water cause it to freeze? Why does ice float on water? Here is a sketch of the explanation:

the oxygen atom in the H_2O molecule has two pairs of unmated electrons that attract the hydrogen atoms on other H_2O molecules. When the kinetic energy of the molecules decreases (i.e., the temperature decreases), each oxygen atom tends to attract two hydrogen atoms on the ends of two other H_2O molecules. When this process is complete, the result is a lattice in which each oxygen atom is attached to four hydrogen atoms. Ice is this lattice, and freezing is the formation of such a lattice, which is why decreasing temperature causes water to freeze. Because of the geometry of the bonds, the lattice has an open, less dense structure than does amorphously structured H_2O (viz., liquid water) – which is why ice (frozen water) floats on liquid water.

Suppose we reject the assumption that temperature is identical to mean molecular kinetic energy in favor of the assumption that temperature is merely correlated with mean molecular kinetic energy? And suppose we reject the claim that freezing is lattice formation in favor of a correlation thesis. And likewise for water/ H_2O. Then we would have an explanation for how something that is *correlated* with decreasing temperature causes something that is *correlated* with frozen water to float on something *correlated* with liquid water, which is not all that we want. The reason to think that the identities are true is that assuming them gives us explanations that we would not otherwise have and does not deprive us of explanations that we already have or raise explanatory puzzles that would not otherwise arise. The idea is not that our reason for thinking these identities are true is that it would be convenient if they were true. Rather, it is that assuming that they are true yields the *most explanatory overall picture*. In other words, the epistemology of theoretical identity is just a special case of inference to the best explanation. (See Block, 1978a, 2002; and Block and Stalnaker, 1999.)

As I mentioned, Kim, Lewis, Chalmers, and Jackson all have a rather different picture of theoretical identity than the one sketched here. Focusing on Kim, as I explained earlier Kim sees the role of identities as really a matter of specifying a realizer of the functional role of a mental state rather than capturing the metaphysical nature of a mental state. And this difference reflects a view of reductive explanation in which the role of reduction of, say, water, is not to find the physical ground of water but rather a matter of finding what plays the water role here and now.

Kim (2005) has objected to pictures of the epistemology of theoretical identity of the sort that I have been sketching. He says that identities such as "freezing = lattice formation" "serve only as *rewrite rules*, and they are not implicated in the explanatory activity" (p. 145). He allows that identities are important in the derivation of explanatory and causal claims mentioned

earlier, but he insists that "this is not an explanatory derivation; rather it is a derivation in which we put 'equals for equals', and thereby redescribe in folk vocabulary a phenomenon that has already been explained" (pp. 145–6). In terms of the example I just described, Kim's argument would be that the explanation of why ice floats on water is just a redescription in folk vocabulary of the explanation given in microterms for why a lattice of H_2O molecules floats on an amorphous conglomeration of H_2O molecules, so the fact that the identities allow us to give the former explanation when we already have the latter one is not an explanatory reason to believe the identities.

I agree with Kim's remark about equals for equals, but I don't think it establishes his conclusion. Let me explain. In a common regimentation that I think does have resonance with the way these terms are used, explanation is usually thought of as determining an "opaque" context, whereas causation is often thought of as determining a "transparent" context. Just as knowledge of the fact that freezing happened is not knowledge of the fact that lattice formation happened, so also an explanation of the fact that freezing happened is not an explanation of the fact that lattice formation happened. By contrast: just as the time at which freezing happened is the time at which lattice formation happened, so the cause of freezing is also the cause of lattice formation.

But I don't want to make too much of this linguistic fact, if it is a fact. Instead, we should be liberal, allowing both an opaque and a transparent sense of "explain." So, in the transparent sense of "explain," Kim is right, and in the opaque sense he is not. And that is enough for my point: in one sense of "explanation," the identities allow explanations one would not have without them. Is explanation in *that* sense enough to ground inference to the best explanation? Yes. It is a fact that ice floats on water, and a view that allows an explanation of that fact – even if only an opaque explanation of it – is thereby made more reasonable to believe than views that do not allow such explanations.

But so far, I have not gotten to the root of the disagreement. I think Kim might agree that identities allow opaque explanations that we would not otherwise have, and perhaps he would even agree that this is a reason to believe in the identities. However, he would not agree that this reason is of the sort that figures in science. That is, the root of the disagreement between me and Kim is not the issue of whether opaque explanation is legitimate explanation but whether it is explanation of the sort that is given in science. In discussing the issue of whether identities explain correlations, Kim (2005) says that "the kind of" explanation "seems entirely unlike

scientific explanations of correlations," the explanation of correlations that science gives being accounts of the mechanism of the correlation (p. 134). And he also notes that "not even Smart, perhaps the most sanguine of the contemporary materialists, thought that the choice between . . . physicalism and dualism was a matter to be decided by science" (p. 142).[12]

However, the same kind of inference to the best explanation reasoning that I gave earlier is repeated *within* science. For example, notions of heat, temperature, pressure, entropy, and enthalpy are notions within what is often called in thermodynamics textbooks "phenomenological thermodynamics," a science that was well developed even before the molecular nature of matter was understood. These phenomenological thermodynamic properties are molecular properties, or as it is sometimes put by scientists, the phenomenological thermodynamic concepts can be defined in molecular terms. Within phenomenological thermodynamics, entropy can be understood in a number of ways, for example, as the amount of energy not available to do work. But entropy is identical to a molecular property, as Boltzmann showed in the late nineteenth century. Entropy is a measure of the number of ways particles can be arranged in a given state without changing the total energy. No one would say entropy as defined thermodynamically is merely correlated with entropy as defined in molecular terms. They are the same thing. The rationale for accepting these identities in terms of families of macroproperties and corresponding families of microproperties is entirely *within science* rather than being a matter of the relation between a folk theory and a scientific theory. Similar points could be made about the relation between Mendelian genetics and molecular genetics; between geometrical optics and electromagnetic theory; between ordinary chemistry and physical chemistry; and between Newtonian rigid body mechanics and Newtonian point particle mechanics. Harkening back to what Kim says about Smart, the choice between reductionism about entropy and dualism (i.e., antireductionism) about entropy is indeed a matter to be decided by science.

Thinking of reductive physicalism in terms of theoretical identities is more conducive to grounding than the picture of reductive physicalism embedded in the Canberra Plan. But identities are not grounding claims. Identity is symmetrical and grounding is not. The identity claim that

[12] The three dots indicate something deleted from the quotation, namely, the word *type*, which I deem irrelevant because the kind of physicalism that we have been discussing all along is physicalism about properties and that is a version of type physicalism. That is, we can distinguish the view that each pain is a physical event from the view that pain, per se, is physical. The latter is type physicalism.

heat is molecular motion does not entail a commitment as to whether heat is grounded in molecular motion or whether molecular motion is grounded in heat. We can see how the fact that ice floats on water is grounded in microphysical facts only if we add to the explanation I gave earlier that water is grounded in H_2O, that temperature is grounded in mean molecular kinetic energy, and so forth. And once we have grounding we can do without identity. If instead of hypothesizing that water = H_2O, temperature = mean molecular kinetic energy, heat = molecular kinetic energy, and freezing = lattice formation, we hypothesized that water is grounded in H_2O, temperature is grounded in mean molecular kinetic energy, heat is grounded in molecular kinetic energy, and freezing is grounded in lattice formation, we would get explanations that are just as good as the ones described earlier. The explanation of why ice floats on water would go through just as well. And the case I made that the explanations that identities facilitate are scientific would go through equally for the grounding hypotheses.

VII. Grounding and multiple realization

Earlier, I mentioned – but did not address – the idea that since the physicalistic approaches to the mind are computational, the distinction between physicalism and functionalism dissolves. One thought along these lines is that a "multiple realization" issue can always be avoided by making one's functional description more detailed, for example, moving from "commonsense" functionalism to psychofunctionalism or neurofunctionalism. Consider, for example, the objection to functionalism that exploits the putative possibility that the functional description of a human might be realized by a group of people.[13] If the functionalist is uncomfortable with supposing groups have mentality, one way to proceed is to go neurofunctional in the hope that the more detailed functional description won't have a group realization – or, alternatively, that the "lower-level" functional description will more plausibly entail mentality even if the realization is a

[13] This example from Block (1978b) was derived from the mention in Putnam (1967) of the possibility that a swarm of bees might realize the functional organization typical of a single organism to which we want to ascribe mental states. Putnam stipulates that "no organism capable of feeling pain possesses a decomposition into parts which separately [are capable of feeling pain]" (p. 227). But what this amounts to is just the stipulation that no functional system of the right sort to be sufficient for pain can be composed of other such systems. Of course it is a mark against the view if it depends on such ad hoc stipulations. John Searle's Chinese Room example (Searle, 1980) is similar. Searle told me before his 1980 paper came out that he had read my 1978 paper before writing his "Chinese Room" paper.

group one. I mentioned this idea in Block (1978b), and it was taken further in Lycan (1981). In terms of the Ramsey approach mentioned earlier, the idea would be that the Ramsified theory T should not be a theory of common sense or of scientific psychology but rather a deeper theory of the neuroscience or physics of the brain. As we will see later, there is another multiple realization problem that operates at the lowest level of science.

I also mentioned Lewis's and Dennett's view that functionalism is just part of the fabric of science, so any scientific account will inevitably be functionalist. However, as I noted earlier, a first-order physical property and a second-order functional counterpart of it are always distinct. Given any first-order physical property, *P*, we can always define a functional property that is constituted by some property's having the functional role (with respect to some specific theory that specifies a level of analysis) that *P* occupies. But the latter property, being second order, is distinct from the first-order property (*P*) that realizes it.

The real thesis of "How to Define Theoretical Terms" is not the view that the meanings of theoretical terms are functional or that the properties science talks about are functional but rather the thesis that a useful regimentation of scientific language is one on which many meanings are functional. I suggest that whatever utility functional definitions of scientific terms have, it is not metaphysical utility and that functional definitions do not yield any sort of grounding.

After the passage quoted near the beginning of this chapter, Dennett (2001) goes on to explain that the level of detail in functional descriptions relevant to the mind – especially consciousness – is the level of detail that makes a difference in computational role. He sees the failure of AI-oriented research about the mind as one of thinking one could get away with too little of the functionalized detail, since functionalized neuroscience is required.

> The recent history of neuroscience can be seen as a series of triumphs for the lovers of detail. Yes, the specific geometry of the connectivity matters; yes, the location of specific neuromodulators and their effects matter; yes, the architecture matters; yes, the fine temporal rhythms of the spiking patterns matter, and so on. Many of the fond hopes of opportunistic minimalists have been dashed: they had hoped they could leave out various things, and they have learned that, no, if you leave out x, or y, or z, you can't explain how the mind works. This has left the mistaken impression in some quarters that the underlying idea of functionalism has been taking its lumps. Far from it. On the contrary, the reasons for accepting these new claims are precisely the reasons of functionalism. Neurochemistry matters because – and only because – we have discovered that the many different neuromodulators and other chemical messengers that diffuse through the brain have functional

> roles that make important differences. What those molecules do turns out to be important to the computational roles played by the neurons, so we have to pay attention to them after all. (pp. 234–5)

Of course, it is no recent discovery that has shown that what the molecules that make up neurons do is important to the *causal* role played by the neurons. What Dennett means is that what has been discovered is that neuromodulators are important to the functional or computational roles played by neurons. Computational roles are a species of causal roles that play a role in a specific kind of causal process, a computation. Thus the phosphors in an old-fashioned screen make it possible for us to see the output of the computer, but our seeing the output is not part of the computation itself. It can scarcely be thought that it is a new idea that neurotransmitters contribute to computational roles. The first discovery of a neurotransmitter was in 1921, and I don't think that those who have thought of the mind as computational would ever have denied that neurotransmitters are part of the implementation of those roles.

What AI got wrong – with respect to consciousness – was not seeing that consciousness is grounded in and must be understood at the neural level. If the scientific basis of consciousness is neural, it may not, however, be neurofunctional. That is an unwarranted further step. An argument for neurofunctionalism as a metaphysics of mind would be an argument to the effect that it is the roles, not the realizers, that are the ground of the mind. Dennett does not give any argument for that view other than the argument that claims that such a thesis is part of a general fact about science, and I dispute that argument later.

What was wrong with AI approaches to consciousness is invisible from the point of view of the Canberra Plan with its excessive focus on ontology at the expense of metaphysics. The flaw in traditional AI was metaphysical. It was not that the AI-ers failed to notice that neurotransmitters have important causal or computational roles. They had the mistaken view that the metaphysical problem of mind could be solved at a level of description that paid no attention to details of neuroscience. Now that mistake can be corrected in one of two ways:

1. Adopt a *physicalistic* approach to the mind, including consciousness, that includes detailed neuroscience.
2. Adopt a *functionalistic* approach to the mind, including consciousness, that includes more details of neuroscience in the functional roles. (Make the Ramsified theory T a neuroscientific theory.)

My complaint about Dennett's approach is that it is blind to option 1. Even after it is acknowledged that neuroscience is important to consciousness, we still have the same dispute that we had earlier between functionalism and physicalism.

But what is the "cash value" of the difference between 1 and 2? Does it really matter which we adopt? I argue later that it matters a lot for metaphysical purposes.

Of course Dennett is right that we care and know about things because of their causes and effects. But it would be a mistake to conclude that all properties are grounded in structures that include causes and effects. Anything that functions as a mousetrap is indeed a mousetrap, but something could function as a banana – in some respects and at least at one level of description – while being a mere ersatz banana. For example, it might be a member of another species that has many of the superficial properties of bananas. Dennett would no doubt agree, claiming that one can avoid the problem by specifying the causes and effects at a lower level, for example, a molecular level. However, exactly the same problem arises at other levels, maybe every level.

I gave an argument to this effect in "Troubles with Functionalism" (Block, 1978b), but it was not very clearly stated. I will try to correct that now, and I will also discuss briefly what Lewis said about the matter in a paper that was published posthumously. The argument is that the lowest level of all, that of basic level physics, is vulnerable to the same point. Putting the point in terms of the physics of fifty years ago (see Feynman, Leighton, and Sands, 1963), the causal role of neutrons is the same as that of antineutrons. If you formulate a functional role for a neutron, an antineutron will realize it too – assuming that the statement of a functional role cannot include indexicals, proper names, or terms such as *neutron*. As Feynman says, "The antineutron is distinguished from the neutron in this way: if we bring two neutrons together, they just stay as two neutrons, but if we bring a neutron and an antineutron together, they annihilate each other with a great explosion of energy being liberated" (p. 52–10) (In recent physics, I am told, there are symmetries that allow a more complex version of the same point.)

Put in terms of the Ramsey definitions mentioned before, the idea is that one could define *neutron* as follows:

Being a neutron = being an x such that $\exists F_1 \ldots \exists Fn\ [T(F_1 \ldots F_n, i_1, etc., o_1, etc.)\ \&\ x\ has\ F_{237}]$

where F_{237} is the variable that replaced *neutron* in the original physical theory. But "being an antineutron" would have a logically equivalent definition, since none of the relations mentioned in the Ramsey sentence would distinguish the variable that replaces *neutron* from the variable that replaces *antineutron*. Suppose the variable that replaces *antineutron* is F_{238}. The Ramsified theory would distinguish between F_{237} and F_{238} only by saying that when particles of type F_{237} meet particles of type F_{238} they annihilate one another. (And particles of type F_{237} do not annihilate particles of type F_{237}; particles of type F_{238} do not annihilate particles of type F_{238}.) One could put the point like this: *neutron* is defined in terms of having causal role R while not being identical to another type of particle that has a role exactly the same as R except that it includes being annihilated by collisions with particles of the first type, but not with particles of its own type. Or, more flamboyantly: what do neutrons say about what they are? They say, "I am characterized by causal role R, which includes annihilating another particle that also has causal role R but that is of a different type from me and not annihilating particles of my own type." If you were communicating by radio with a functionalist in a remote part of the universe, you would not be able to tell from what this person said about physics whether he/she lived in an antimatter part of the universe or a matter part of the universe.

Many suppose that it is conceivable that there be a realization of human functional organization that is mentally different from ours, for example, "inverted" or "absent" qualia.[14] The argument just given provides a case for multiple realization of even the lowest level of physics.

This point shows, contra Dennett, that there is no general functionalist metaphysics of mind that works for all of science.

Does it follow that there could be a world identical to this one in which matter and antimatter are switched and in which there is no consciousness? That is conceivable in the sense that there is no contradiction or incoherence in its supposition, but perhaps it is not metaphysically possible. Note that if it is true, it does not show that consciousness is causally impotent in our world. For example, for many computational structures, there are computationally equivalent electronic and hydraulic implementations. But it would be a mistake to conclude that the electrical properties or the hydraulic properties do not do anything significant (Block, 1980a). They have parallel causal efficacies.

This point does not refute the project of Lewis (1970), but it does show Lewis's point of view to be inadequate as an account of reductive

[14] These terms were used for the first time, I believe, in Block and Fodor (1972).

physicalism. Recall that Lewis would define *neutron* as the (contextually indicated) thing that has a certain causal role, where the causal role would be spelled out in terms of a Ramsified physics. Even if there is more than one thing that satisfies the neutron-role as spelled out in terms of a Ramsey sentence – as I have said – there is only one context-relative thing – in many contexts. Further, and importantly for Lewis, there is no actual context in which the definition picks out anything nonphysical. (In my correspondence with Lewis about this issue, he said it would be sufficient for his purposes that there is a pair of things – neutron, antineutron – that is picked out by the Ramsey definition, even ignoring the context relativity.) Lewis was not concerned with the question of whether there is a functional definition that tells us what grounds something being a neutron. He thought of himself as a physicalist (ignoring his 1980), not a functionalist.

It is odd that Lewis treated mental terms and terms of physics such as *neutron* as on a par since functionalism makes more sense for *neutron* than for *pain*. I objected to functionalist grounding of pain, saying it raised a puzzle about why the functional relations known to common sense are part of grounding but those that elude common sense are not. However, in the case of a functional reduction of *neutron*, there is no such issue since the scientifically important properties are to be included in the Ramsified theory.

In a paper published posthumously, "Ramseyan Humility," Lewis (2009) returns to these issues. He says:

> We have assumed that a true and complete final theory implicitly defines its theoretical terms. That means that it must have a unique actual realization. Should we worry about symmetries, for instance the symmetry between positive and negative charge? No: even if positive and negative charge were exactly alike in their nomological roles, it would still be true that negative charge is found in the outlying parts of the atoms hereabouts, and positive charge is found in the central parts. O-language[15] has the resources to say so, and we may assume that the postulate mentions whatever it takes to break such symmetries. Thus the theoretical roles of positive and negative charge are not purely nomological roles; they are locational roles as well. (p. 205)

A brief quibble: my neutron/antineutron example avoids the "location" issue mentioned in Lewis's "charge" example. Neutrons and antineutrons

[15] That is, the language of old terms, ones known before the introduction of T-terms via Ramsey definitions. Lewis obviously intends that the reader be reminded of the word *observational* while explicitly denying that there is any principled distinction between observational and theoretical terms.

are both located in the nucleus of an atom in the same "place." Less superficially, the key term in Lewis's discussion is "hereabouts." What Lewis is saying – put in terms of my example – is that a neutron is the kind of particle that has causal role *R hereabouts*. That of course is compatible with a different particle having causal role *R* somewhere else in this world or in a different possible world. If one thinks of Lewis as offering a definition in terms of a definite description of the form "the particle with causal role *R*," then the role of the "hereabouts" is to make the context relativity I mentioned explicit. But there is another way to take his remark. One could take him as offering a definition of the metaphysical ground of being a neutron in terms of having causal role *R* and being hereabouts. That would involve introducing an indexical (alternatively, a name) into a Ramsey definition. I considered something like this move, noting, "One could avoid this difficulty by allowing names in one's physical theory. For example, one could identify protons as the particles with such and such properties contained in the nuclei of all atoms of the Empire State Building" (Block, 1978b, 1980b, p. 302). (I also mentioned the option of ostension.) However, if the purpose is metaphysical, the indexical or name would seem to ruin the project, bringing defeat for the Ramseyan approach. Assuming we are willing to allow the indexical fact or the fact about a named individual as part of a ground at all: the difference between the metaphysical ground of the property of being a neutron and being an antineutron is in part nonfunctional – and profoundly unsatisfactory as an account of what grounds the particle properties.

This is the example alluded to in Section 4 where I said that since grounding is hyperintensional, even if a first-order physical property played the pain functional role in every possible world and nothing else played that role, that would not show that the first-order physical property is the metaphysical ground of pain. Arguably, an indexical fact or a named individual fact in a putative ground precludes the grounding relation. But why? One might suppose that it is because if one wants to know the explanation of the difference between particles and antiparticles, it does not help to be told that particles are the occupants of causal role *R* hereabouts and antiparticles are the occupants of causal role *R* thereabouts. Though it does not help to be told that the difference is that between hereabouts and thereabouts, it does raise the question of what *would* help. Perhaps the irreducibly different quiddities of particles and antiparticles provide the ultimate difference in ground. Or perhaps there is no "bottom" level, with an ever-descending chain of grounding relations (Block, 1997a).

Block (1978b, 1980d) said including names or devices of ostension in Ramsified theories was contrary to the idea behind functionalism.[16] Lewis was willing to allow "hereabouts" because, although he saw himself as a physicalist, he ignored physical grounding. Indeed, though he was a functionalist about the meanings of theoretical terms, he ignored functional grounding as well. Judging from the views of his discussed here, he was simply blind to what I am calling metaphysics.

VIII. Conclusion

The Canberra Plan is supposed to be a model of reductive physicalism, but it neglects ground, sacrificing what I am calling metaphysics on the altar of ontology. In particular, it has no room for an account of the physicalistic ground of mentality. I mentioned that the kind of reductive physicalism acknowledged by the Canberra Plan is blind to the dualistic implications of the Commander Data case, so the account of reductive physicalism is inadequate.

The a priori functional analyses mentioned by many adherents of the Canberra Plan would provide only a nominal ground of the mental, not anything substantive, and Kim in his most recent writings on the topic abandons the Canberra Plan for those reasons. Putative disjunctive grounds are explanatorily inadequate for the reasons I gave. One might try to revive functionalism as an account of ground along the lines proposed by Dennett or proposed by Lewis's view of theoretical terms, but the resulting functionalized science would require at the most basic level indexical or name-related facts as part of ground. If there is no substantive physical or functional ground of mind, in an important sense dualism is true, but the Canberra Plan neglects dualism in that sense.[17]

[16] Georges Rey (1997) has also argued for this view, concluding, "I think both friend and foe of the functionalist strategy would agree that this would violate its spirit" (p. 176).

[17] I am grateful to Eliza Block, David Chalmers, Kit Fine, Jaegwon Kim, David Sosa, and Jared Warren for critiques of an earlier draft.