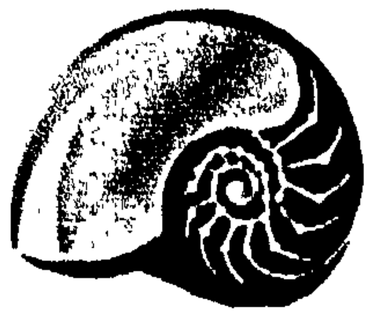
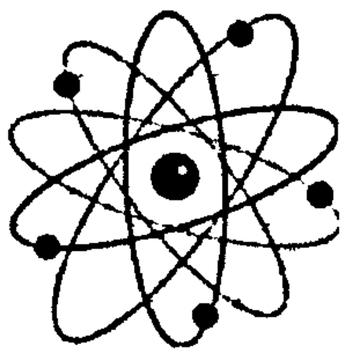
PSYCHOPHYSICALITY:

Rethinking the Physicalist Foundations of the Mind/Body Problem



Submitted to the University of Glasgow Department of Philosophy
in requirement of the degree of Ph.D
by
Cleodhna P. A. Nightshade
01/10/2001
Supervisor: Dr. Susan Stuart



Acknowledgements

No work of any worth is done in a vacuum. I am immensely grateful for all of the help and inspiration I have received along the way in constructing my little contribution to what is to my mind one of the most interesting problems there are. I must especially extend thanks to Susan Stuart, whose invaluable help and timely encouragement have been instrumental in the completion of this work – and for introducing me to the mind/body problem in the first place; to Pat Shaw, for his support and encouragement, and for making painless so many of the administrative factors involved with producing a doctoral thesis; to Kevin Clark, for support moral, emotional, geographical, poetical, and non-euclidean; Dr. Michael Charleston, Dr. Jonathan Sheps, Christine Atherton, and Steven Hanlon, whose friendship and assistance in specialities mathematical, biological, neurophysiological, and particle-physics-ical, has been of immense value; to Erik Kahn and Elizabeth Hale, for extending their help when it was most needed; Marcia Vance, my infinitely patient and capable student loans officer at the Vermont Student Assistance Agency; Beanscene, for making such *highly caffeinated* coffee and providing such a comfortable environment in which to work; the Reverend Dr. James E. Thomas, Joshua Pendragon, and most of all Sam Kington, without whose love and support I would not be a very happy or pleasant person, and my family (in which I include Ian Kozak, Skarrn Ryvnine, and Karl August Isselhardt), especially my sister Crystal, without whom I would long ago have assumed a primarily lapidoculous lifestyle.

Thank you all.

Table of Contents

Introduction

Mind in its purest play is like some bat
That beats about in caverns all alone,
Contriving by a kind of senseless wit
Not to conclude against a wall of stone.

It has no need to falter or explore;
Darkly it knows what obstacles are there,
And so may weave and flitter, dip and soar
In perfect courses through the blackest air.

And has this simile a like perfection?
The mind is like a bat. Precisely. Save
That in the very happiest intellection
A graceful error may correct the cave.

-Richard Wilbur, 'Mind' (1956)

RICHARD WILBUR evokes the image of the mind as a self-contained, fluttering thing, animated by an innate, uncanny insight into the environs that confine its coursings. By some power alien to the insensible caverns it inhabits, the mind is aware of the contours of its surroundings, so that unerringly it weaves a right path through a maze of possibilities. But beyond this point, the metaphor fails, for unlike as for a bat, an endless cavern is no prison for the mind. Unlike its winged simile, the mind is able to exert its own influence upon the walls that surround it and dictate its paths, and so carve for itself new ones. Where a passae may not lead to a destination it desires, the mind takes matters unto itself and willfully constructs a new way. The mind corrects the cave.

That the mind makes a difference in the physical world is beyond commonsensical questioning, but the problem of how it does so has been with us for centuries, even before Descartes, so captivated by the mechanical workings he perceived to be operative in the material world and the rational workings he understood to govern the right-thinking mind, declared the mental and the material to be of disparate essence. Descartes was never able to answer, to his own satisfaction or to anyone else's, precisely how an immaterial substance such as he held the mind to be could possibly exert a causal

influence upon the material substance, or how such influence would not violate the closed mechanical laws subsuming the material world. The insurmountable difficulty involved with explaining the interaction of an immaterial mind and the material world has led most modern philosophers of mind to abandon substance dualism in favour of monism, and overwhelmingly, materialist or physicalist monism. We now tend to believe that our minds are not substances apart from our bodies, but that minds are somehow realised in the complex workings of our brains and bodies The interaction problem is this dissolved, but the mind/body problem is far from solved by such a move. For us, the problem has ceased to be how a mental substance, given that it shares nothing in common with a physical substance, can exert its influence in the physical world, and has become the problem of how mental states, given that they are realised in physical states of the brain and body, can be causally significant in their own terms. Insofar as we hold on to the idea that it is our beliefs and our desires that motivate our actions, and that our actions are best explained in terms of our reasons for acting, there remains a tension between explanations for events involving mental causes and effects in terms of reasons, and explanations for the same events in terms of the physics of whatever it is in which the mental states involved are realised.

One move which has enjoyed some popularity in the debate on the mind/body problem is to question the validity of the physicalist position itself. If it can be shown that physicalism cannot be formulated so as to set up a distinction between the psychological and the physical sciences, then the tension between psychological and physical explanations may dissolve. Alternatively, if it is in principle impossible to differentiate between physical properties and nonphysical properties, then it is difficult to say what reasons we could have for excluding mental properties from being legitimately involved in explanations of events in their own terms *qua* mental. If we cannot differentiate between physical and nonphysical properties, and if there is no real distinction between the substance and methodologies of the special sciences and the physical sciences, then any privilege, ontological or explanatory, that we place in the physical sciences can be seen to be ill-founded.

It is my belief that the wholesale rejection of physicalism at this point in the debate is premature. The doctrine of physicalism has much to recommend it, foremost of its merits being that it is an expression of the rational requirement that *explanations should make sense* — that meaningful, comprehensible rules circumscribe all the world's relata; that the regularity we think we find in terms of the relations of cause to effect, of parts to wholes, or of the general to the specific, are not brute and miraculous features of a world of discrete particulars, but that some metaphysical glue, whatever that miht actually be and whether or not we could possibly fully comprehend its nature, actually does hold the world together and ground the order we perceive. So central are these notions to the way we conduct our enquiries into the workings of the world, whether in terms of the special sciences or in terms of basic physics, that the extent to which physicalism is motivated by them or motivates them deserves attention outside of the mind/body debate. Any insight into the question of physicalism, however, will have an impact upon the mind/body problem.

In this thesis, I shall examine the question of physicalism through two papers criticising the formulation of the doctrine. In the first chapter, I discuss Tim Crane's and D.H. Mellor's influential (1990) *There Is No Question of Physicalism,*[1] in which they argue that there are no real criteria by which the science of psychology can be separated from the paradigmatically physical sciences, and so no principled reason to suppose that the predicates of psychology do not describe real elements of the world's ontology whereas those of physics do. I shall explain why I find their arguments unconvincing, and to show how some of the reasons they consider not to support the noncontinuity of psychology with physics actually can support the distinction.

Crane and Mellor take physicalism to be an epistemological doctrine, according to which the empirical world "contains just what a true and complete physical science would say it contains".[2] Physicalism can, however, be taken as a metaphysical doctrine, and indeed I think that many modern physicalists do take it this way. In his (1998) *What Are Physical Properties?,*[3] Chris Daly argues that no principled distinction can be drawn between physical and nonphysical properties, and that therefore any metaphysical programme which assumes such a distinction is misguided. I shall agree with much of his reasoning, but not with his 'downbeat' conclusion:[4] while I agree that there are serious difficulties involved in setting constraints on the bounds of the physical, I think that enough can positively be said to make physicalism a meaningful position. Between the two papers, a fairly broad survey of some recent accounts of physicalism is made and these two distinct avenues explored: physicalism construed as a doctrine about science, and physicalism as a doctrine attempting to limit the contents of the world *a priori* through a definition of what it is to be a physical properties. All in all, I think that there is much to learn from these two papers, but not all of it is as negative, conclusive, or 'downbeat' as their authors might have intended. Rather, I think that some new directions are indicated by the failure of some of the avenues they explore.

In the third chapter, I shall consider some of the difficulties for physicalism and reductionist physicalism in particular as were indicated in the preceding discussions. Based largely upon the functional model of reductionism recently forwarded by Jaegwon Kim in his (1998) Mind in a Physical World,[5] I present a formulation of physicalism which I believe to be consonant with the primary aims and concerns of physicalists, and which is not faced with some of the difficulties presented for traditional reductionist physicalism. I shall argue that a minimal expression the doctrine of physicalism is best understood as encapsulated in two principles: the principle of the causal closure of the physical domain and that of the supervenience of all phenomena upon physical phenomena, where that supervenience is taken to be grounded in a metaphysical and explanatory relation between higher and lower level properties. I agree with Daly that it is not possible to set *a priori* constraints upon what properties are to count as physical properties, but I shall argue that this is neither necessary nor desirable: the constraints imposed upon what can and cannot be allowed into the physical ontology in the conjunction of the principle of the causal closure of the physical domain and the metaphysically and explanatorily grounded supervenience of all phenomena upon physical phenomena is sufficient to set physicalism up as a positive doctrine.

The arguments of the third chapter define minimalist physicalism, and indicate that functional reductionism is consonant with physicalism. In chapter 4, I further argue, again following Kim, that any version of nonreductive physicalism that construes mental properties and generalisations realistically is incoherent. Kim has argued[6] that nonreductive physicalism is in essence a version of emergentism, a doctrine which flourished in the earlier quarter of the 20th century, according to which some properties of composed wholes have simply to be accepted as brute, inexplicable in terms of the properties of the elements that constitute them. Because of some very fundamental differences I find between emergentist doctrine and nonreductive physicalism, I do not agree that modern nonreductive physicalism is a restatement of the earlier doctrine. However, I do believe that, as for emergentism, nonreductive physicalists cannot consistently retain both a commitment to the real nature of mental properties and generalisations and to the principle of the causal closure of the physical domain. Therefore, as a physicalist doctrine, nonreductive physicalism is at odds with itself.

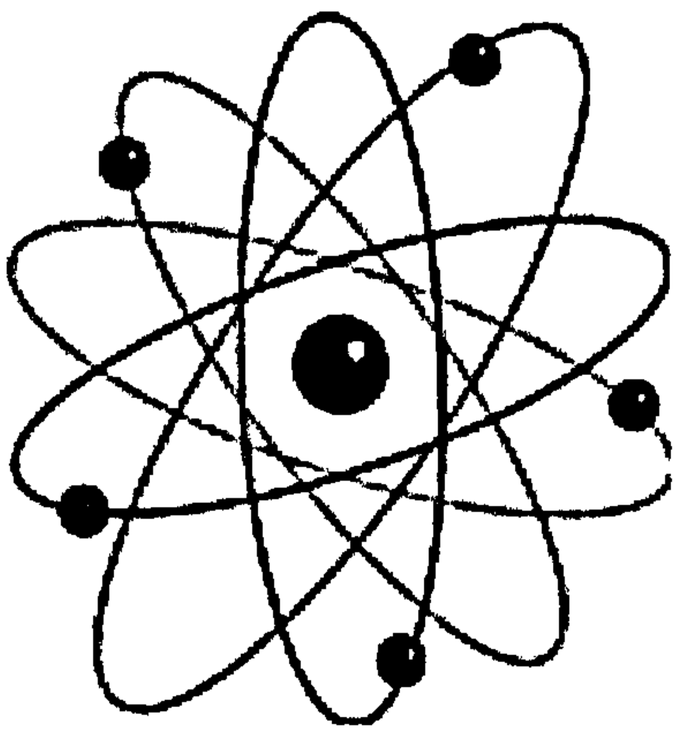
The conclusions of the fourth chapter are largely negative, and the discussion of the third chapter does not indicate any way in which the mental, *qua* mental, can be construed as causally effective in the physical world. In the fifth and final chapter, I shall attempt to address this deficiency. The notion of dynamicism has received quite a bit of recent attention in the fields of artificial intelligence and artificial life, and in the study of the eusocial hymenoptera. In dynamic systems, complicated, ordered, 'intelligent' behaviour can be bought very cheaply through the exploitation of interactions between dynamic components that follow very simple rules, and thus enable whole systems that can instantiate them to act in and interact with their environments in ways unanticipated by the properties of the least whole parts (such as a single ant, or a single brain cell) realising them. Further, higher-level dynamically emergent properties have the very interesting feature of self-perpetuation: once they are instantiated in a system, the components of the system tend to be constrained by the presence of the property so as to continue to realise the property. These features indicate that dynamically emergent properties are particularly robust properties of complex systems, and there are good reasons to believe that they are best understood in their own terms, at their own level, as features of the systems instantiating them. Dynamically emergent properties are, however, demonstrably unproblematic from a physicalist perspective: they are fully realised in and explicable (if not necessarily predictable in practical terms) in terms of their microstructures. Following Andy Clark and Brian Goodwin, I illuminate the notion of what I term 'dynamically emergent properties'. As John McCrone has indicated in his Going Inside, a dynamic understanding of how the brain works is taking hold in current neuroscience. Primarily following his descriptions, I show how it could be possible that one of the brain's abilities, the recognition of complex objects, is an ability that emerges dynamically in the actions and interactions of many highly specialised cells in different areas of the brain. If indeed the brain is a realiser of dynamically emergent properties, then we can begin to see how mental properties and the laws of psychology, taken realistically, might be unproblematically grounded in the physical and the laws of physics without being reducible to them.

There is much that I do not address. Most outstandingly, the problem of consciousness – of why any processes in the brain, no matter how complex or dynamic – should be accompanied by phenomenal experience is untouched by any of the discussions I have been able to forward. Indeed it is my belief that what David Chalmers has termed 'the Hard Problem' of consciousness is really the hardest problem facing modern philosophy of mind. It is my hope, however, that with a better understanding of the requirements of a physicalist system and with greater knowledge of the possibilities there are for the realisation of robust higher level properties in complex physical systems – an understanding of how the mind, by happy intellection or graceful error, may be able to correct its cave – we may better envision how to proceed.

1

Questions of Physicalism

1. Physicalism as an Epistemological Doctrine



THE DOCTRINE THAT EVERYTHING THAT EXISTS either is physical or is fully realised in physical things would of course be an empty one if it were not possible to draw a principled distinction between physical things and things that are not physical. This truism is the point of departure for Tim Crane's and D. H. Mellor's now classic (1990) *There Is No Question of Physicalism.* The physical or nonphysical status of the sciences appropriate to the study of various phenomena can be, and generally are, taken to be definitive of the physical or nonphysical status of the phenomena they pick out. Against the doctrine of physicalism, Crane and Mellor argue that there are no real criteria by which the science of psychology — the paradigm nonphysical science — can be separated from the paradigmatically physical sciences, and so there is no real reason to suppose that psychology, or by extension any other special science, is not contiguous with the physical sciences. If that is so, then the claim that everything is in one way or another susceptible to description and explanation in terms of the physical sciences, by virtue of being straightforwardly physical or being fully realised in physical phenomena, is empty.

Crane and Mellor take modern physicalism to be a doctrine about the empirical world to the effect that it contains "just what a true and complete physical science would say it contains".[1] Significantly, though they take it that it is descended from seventeenth-century materialism, which is a metaphysical doctrine, they take it that modern physicalism is not a metaphysical doctrine. The materialists of the classical tradition, they tell us, were concerned to limit the contents of the world *a priori* by describing what material things had to be like: specifically, that they had to be extended in space, impenetrable, conserved, and subject to deterministic mechanical law. Of course, as modern physics describes basic physical elements, they in no way resemble the seventeenth century materialist conception of fundamental material entities Accordingly,

modern physicalists "...have — understandably — lost their metaphysical nerve. No longer trying to limit physics a priori, they now take a more subservient attitude: the empirical world, they claim, contains just what a true and complete physical science would say it contains."[2]

According to Crane and Mellor, physicalism cannot coherently be formulated to do the work that it is supposed to do: to claim priority to physical facts as against nonphysical facts, and to claim for the physical sciences alone that their terms represent genuinely real entities. Their attack is not directed towards the falsification of physicalism in favour of dualism; rather, they hold that "there is no divide between the mental and the non-mental even to set physicalism up as a serious question."[3] They, it must be said, are not exempt from a general trend they open their paper in criticising: they do not give the doctrine of physicalism a very clear definition, except to qualify that the see it as a doctrine essentially about the mental, to the effect that all things mental are "really physical,"[4] and that they are interested in physicalism as about the empirical world, rather than as a doctrine that there are no universals. This lack of detail might be forgiven, considering that the task they have assigned for themselves is to examine possible justifications for the central claims of physicalism as have been offered in the past, to see where and how they fail. Yet what they perceive these justifications as failing to do answers to a predefined notion of the aims of physicalism that they do not expressly state, which is that physicalism is essentially a reductive program aimed at eliminating nonphysical entities, properties, and generalisations form its recognised ontology.

Thus the form of physicalism that Crane and Mellor are explicitly concerned to reject is eliminative physicalism, the thesis that, as they say, "all entities, properties, relations, and facts are those which are studied by physics or other physical sciences",[5] and that entities, properties, relations and facts studied by the special sciences, most notably psychology, are not *real.*" So much can be gathered from such phrasing as "Why should we suppose the existence of sub-atomic particles to require the non-existence of atoms, molecules, tables, trees or tennis rackets, figs or fast food restaurants — or animals or people with minds?".[6] The physical sciences then have a privileged epistemic position. They alone have "a unique ontological authority: the authority to tell us what there is".[7] Crane and Mellor undertake to question the source of this authority, by which physicalists might dismiss the special sciences as "non-physical, and thus ontologically inconsequential".[8] But not all physicalist systems are essentially reductive or eliminative. I am sure that Crane and Mellor would agree, but it is not immediately evident from the discussion that they do not see physicalism as essentially eliminative. However, all versions of physicalism do claim for the physical domain a certain ontological priority. This naturally will preclude special sciences from adding straightforwardly and in their own terms to the inventory of 'what there is', for the physicalist will claim for all facts featured in the special sciences that they are under some description physical facts.[9] The two claims are very different: it is one thing say of some paradigmatically non-physical thing — an economy, for example — that it must be realised in a physical system, quite another to say that only the physical system is real, the economy is not. Crane's and

Mellor's arguments might be extended to cover the weaker, noneliminativist claim as well, however, so I will proceed with the weaker claim primarily in mind.

Crane and Mellor rightly point out that to physicalists, the physical sciences have a certain epistemic superiority, or epistemic prioricity, as regards the special sciences, and quite rightly, they question the source from which this superiority arises, and what it is about special sciences, and their paradigm special science, psychology, that makes them 'epistemically suspect'. They are right to conclude that there is nothing in the quantity or quality of physical laws and the predictions and explanations based upon them that ought *prima facie* to render special science predictions and laws specious. Indeed, on a practical level, the predictions physics could make for phenomena normally dealt with by a proprietary special science would be useless, if not impossible: neuroscientifically based explanations of behaviour are a dream for some, but for the rest of us, and for all of us right now, good old belief-desire folk psychology carries the day for explanations of the actions of intentional agents. No, as Crane and Mellor point out, the appeal of physicalism cannot rest on the notion that the physical sciences are just *better* (at providing explanations) than the special sciences; if anyone thinks that, they must have reasons other than the obvious evidence for thinking it. "The bounds of the physical are set from the outside",[10] Crane and Mellor say. But, they believe, the bounds of the physical cannot be set so as to differentiate the physical from the non-physical so as to make physicalism a nonvacuous doctrine. They argue that there can be no principled reason why psychology (and presumably other special sciences by extension) are not actually continuous with the paradigmatically physical sciences, because, as they attempt to show, the same difficulties involved in the reconciliation of psychology with basic physics are present between paradigmatically physical sciences as well.

Physicalism, then, gives to physics and to physics alone the ontological authority to inform us as to what the world contains: the world "contains just what a true and complete physical science would say that it contains".[11] The task at hand is to assess this authority, and the aptness of the sciences to which it is given to bear it. What does physical science encompass, and whence comes its authority to dismiss special science concepts as empty? There are some sciences that are generally considered to be physical sciences, while clearly not being the sciences of the most basic entities as we currently understand such basic physical sciences: Crane and Mellor name such 'paradigmatically physical sciences' as mechanics, electromagnetism, thermodynamics, gravity, and particle physics, as well as sciences like chemistry and molecular biology. What principle allows us to consider these sciences as physical, and entitled to the ontological authority accorded to physics, while this authority is denied to psychology and other special sciences?

1.1. Reducibility to Physics

The standard response to which Crane and Mellor first direct their attention is that the paradigmatically physical sciences are directly reducible to physics, while other

special sciences, notably psychology, are not so reducible. The reduction in question is derivational: a science is reducible to physics if all of the generalisations in the science to be reduced can be derived from physics supplemented with suitable 'bridge principles' connecting the terms of the science to be reduced with the terms of physics. Of course, the reduction of any particular science to physics does not actually have to have been done for the science to be considered reducible; it is sufficient if the sciences in question are 'reducible in principle' to physics. If all the terms in any particular science are amenable to bridge principles linking them with terms in physics, such that all of the generalisations of that science could in theory be derivable from the laws of physics, then in principle, an explanation of any particular event in that science could be replaced by an explanation in physics with no loss of content. Qualifiedly reducible sciences thereby gather their authority to tell us what the world contains directly from physics. As long as we know the bridge principles, or at least as long as we know that there could be such principles, we can rest assured that these sciences are not adding any elements to our ontology that physics cannot accommodate. The principle according to which 'paradigmatically physical' sciences, and any others that may qualify, can be identified as physical sciences is dubbed, somewhat unfortunately, the RIP (Reducible In Principle) principle.

Crane and Mellor criticise this approach from three directions. First and most obviously, they ask: to what *physics* shall we say that reducible sciences are Reducible In Principle? Physics as a science is an ongoing project, subject to change or to supplementation as the science develops. This, obviously, will have an effect upon what sciences we find are reducible in principle to physics: current physics may not be able to accommodate some certain special science, and so by the RIP principle we would be obliged to judge that science not to be a physical science; if we are materialists of an eliminative demeanour, we will affirm that its terms do not refer to anything real. Yet some future development in physics may make that same science amenable to reduction. Would that salvage both the science and its terms? If so, how can we possibly know which of the special sciences we might now employ may, at some future point in the history of physics, be amenable to reduction?[12]

Crane and Mellor also point out that if the science to which eligible sciences are Reducible In Principle is current physics, then it is highly probable that any future developments in *physics* would render physics itself not a physical science, for it is highly improbable that any future physics will be reducible to current physics. It is more likely to be the other way around: current physics will be reducible to further, more encompassing, developments in physics. But neither can we appeal to some Unspecified Future Physics to tell us what sciences are to count as physical, because we do not, and cannot, know what sciences this Unspecified Future Physics might cover. Nor does the principle under discussion serve to illuminate us at all in that respect. If we apply the Reducible In Principle principle to the Unspecified Future Physics in an effort to divine what sciences the Unspecified Future Physics might cover, then we've made a mistake, for as we have seen, we cannot know what sciences might be Reducible In Principle to a physics upon which we cannot set constraints as to what properties it will or will not be

able to accommodate. Are we to take it that the Unspecified Future Physics will be able to encompass *all* the special sciences? If so, then physicalism comes out vacuous, and if not, we are left with precisely the problem with which we began. We do not know what sciences will be Reducible In Principle to an Unspecified Future Physics.

I agree that to define physicalism this way is highly problematic. However, I do not agree that this is what we are necessarily doing when we say that any science that counts as a physical science is in principle reducible to physics, and I would argue that, while the principle is interesting, it is not a reasonable way to *define* physicalism at all. It would seem that if we do define physicalism in terms of a true and complete physics then physicalism becomes either trivial, or very probably false: if the history of science has taught us anything at all, we can expect that physics will undergo serious revision in the future, and additions or alterations to its ontology may be made. Unless we allow physics room for growth, physicalism defined in terms of physics will be falsified as soon as physics changes. But if we do allow physics room for growth, it is unclear what sorts of facts it may cover, or what sorts of entities and properties it may admit. Unless there are independent principled reasons to exclude certain kinds of phenomena from subsumption under physics or independent constraints applied to Unspecified Future Physics, then conceivably, an Unspecified Future Physics might cover anything we like, and so physicalism becomes trivial.

Many have held, and held reasonably, that all physical sciences must be reducible, at least in principle, to physics,[13] and the question of to what physics they are to be reduced is well placed. Crane and Mellor have already given the answer when they tell us what physicalists say the world contains: exactly what a *true* and *complete* physics would say that it does. If it is a *true* and *complete* physical science that is supposed to tell us which entities and properties we might name are *real,* then it is unreasonable to postulate any current physics or any unspecified future, but presumably incomplete and possibly less than true, physics as the science to which all eligible sciences should in principle be reducible, for that would give these physics the authority to tell us what there is, which, if they are not true and complete, they are in no position to do.

I think that many physicalists would agree that if we had a true and complete physical science, the entities proprietary to that science would be the basic, fundamental entities at the bottom of the ontology, and that current physics and any unspecified future physics, as well as all of the paradigmatically physical sciences, would be reducible to it. That is not to say that it is the physics that sets the criteria as to what the basic physical elements are. Indeed, if we take on board the view that the object of physics is to discover things about the world, it ought to be the world which sets the constraints on what the true and complete physics will say the world contains. The idea that all sciences that are to count as physical sciences would have to be reducible to this hypothetical science is not terribly informative, and quite leaves open the question of whether or not there may be sciences that describe physical or physically instantiated systems that are not physical sciences.[14] Whether or not we can or ever will have a true and complete physics is an open question, and one that could certainly be argued; such argument is beyond the scope of this thesis. What is important is only that reducibility to a

hypothetically true and complete physical science cannot be a criterion for defining physicalism, or for identifying those sciences that can tell us what there is in the world, as we are not in a position to say what sciences might be so reducible. That physics as it stands is not complete, and that we do not know what physics will cover in the future, is no objection to the common physicalist idea that special sciences must in some sense be reducible to physical sciences. But it does mean that we cannot use the principle to evaluate which of any possible sciences would count as physical sciences, for we cannot say which among them could be reduced to the hypothetical true and complete physical science. To dismiss psychology and other special sciences as 'ontologically inconsequential' on the ground that they are irreducible to physics would, thus, be a mistake. But since we do not know even if *physics* is reducible to hypotetically true and complete future physics or whether or not psychology might be reducible to such a science, we have no basis upon which to claim ontological consequentiality for any science on the grounds of its in principle reducibility. Crane and Mellor are quite right to deny that Reducibility In Principle provides the physical sciences the ontological authority to tell us what the world contains.

Part of the appeal of the RIP, Crane and Mellor say, is that it is compatible with a view of the unity of science, or the goal of being able to derive all sciences from one great comprehensive science. That is certainly a dream alive and well in some quarters, but one that cannot be pursued, in Crane's and Mellor's opinion. Obviously, different kinds of sciences are needed to study different kinds of phenomena, and these methods do not represent anything unified:

> The world contains a vast number of very different kinds of entities, properties, and facts. That is why so many different sciences are needed to study them. No one could think astrophysics and genetics unified even in their methods, except under the most abstract descriptions of scientific methodology. And in their contents, they display no more unity than that of a conjunction. Nothing wrong with that: but why then cannot psychology supply another conjunct?[15]

Further, even physics, they note, is not unified; quantum mechanics and gravity have not been reconciled to anyone's satisfaction (yet, to my knowledge),[16] a possibility that Crane and Mellor do not seem willing to admit.[17] I do not think that this is quite the right way to view the dream of the unity of science. Even if a comprehensive physical science were developed, it would not mean that scientists would stop using the sciences and methodologies they now employ, or at least not as long as those methods remain useful. This is for practical reasons: many scientists, as well as other people, will unhesitatingly tell you that all the world is made out of quarks, the constituents of subatomic particles. Nobody believes that the physics of quarks is the appropriate science for the study of everything, for the practical consideration that the endeavour is neither humanly possible, nor would it give answers interpretable to human scientists interested in anything but quarks.

Further, even if the methodologies of various different sciences are not unified, it seems reasonable to think that the dream of the unity of science is motivated by a genuine unity exhibited by most, if not at some level all, of the empirical sciences. They

are all about real, tangible things in the world: things that can be seen, poked, prodded, taken apart, viewed under an electron microscope, whose smoke trails can be observed in accelerators, &c. The properties of atoms and molecules — as examples of such real tangible things in the world — are very important to such diverse sciences as, say, astrophysics and genetics. The two sciences may not focus on *the same* properties of atoms and molecules, but I hazard to guess that an astrophysicist will recognise a geneticist's description of carbon as awfully familiar, as easily as the geneticist will agree that the atoms in amino acids have their ultimate beginnings in stellar dynamics. The biologist is conscious of a continuity between her science and chemistry, microbiology, physics; so too even the psychologist, who does, after all, deal with physical human beings, and who must recognise that chemical facts are factors in the concerns of his subjects. All of the empirical sciences are, in that sense, unified in their subject matter. *None* of the sciences so considered supply an ontological conjunct to the inventory of the world (even if they do supply an epistemological one). It is hardly surprising if psychology also does not.

Well and good, but what, one might ask, are the characterising features of 'real, tangible things in the world' supposed to be? Is it that these things are physical? If so, then to define all sciences as physical because they are all about physical things is circular; moreover, it's false: many of the sciences are unconcerned with the *physical* particulars of their subject matter (psychology and economics being the paradigm examples). That, however, would be to stand the argument on its head. I only mean to express that the idea of the unifiability of science is not inconsistent if different sciences are not unified in their methodologies or in the particulars of their foci. Is the characterising feature of 'real, tangible things in the world' that they are empirically accessible? If so, then Crane and Mellor could surely argue that psychological facts are eminently accessible, as much so as anything else. I do not deny it, but I trust that Crane and Mellor would agree that psychological facts are accessible in a slightly different way from, say, rocks or electrons, and that physicalists will have to provide an account of how so. I do not think that they are prevented from doing so by the fact that different sciences employ different methodologies.

Crane and Mellor would probably not admit that the unity of subject matter I put forward represents any kind of genuine unity, for it assumes, essentially, that any science that studies real tangible things in the world is in a sense contributing, or contributed to by, other sciences which also study such things, and this is largely because the stuff of some sciences — molecular biology, for example — is composed of or comprised by the stuff of other sciences — chemistry, perhaps, or more basic sciences. This is the third source of the appeal for the RIP principle that they criticise.

Crane and Mellor identify microreduction as the idea that "there is really nothing more to things than the smallest particles they are made up of."[18] Since "the study of the smallest entities is indeed traditionally called 'physics': departments of physics have by long tradition cornered that particular market",[19] there is really nothing more to the world than physical things. That would of course entail that everything that is extended in space is physical. Of course, physics is not *only* the science of very little things, as Crane

and Mellor point out; it also is directed at very large things indeed. Thus physics must give an account of how and why things such as galaxies and quasars fall within its province while other kinds of large things do not, and the account that is given, according to Crane and Mellor, is that facts about such large things reduce to facts about the very small things that constitute them. If this notion is supposed to support the RIP principle, its appeal to reduction is fatal.

This is not an acceptable way to interpret the thesis of microreduction,[20] and the language employed is obfuscatory. It seems that Crane and Mellor are suggesting that the thesis of microreduction is, in essence, that everything is made out of small parts and descriptions of everything and of interactions between all things can be reduced to descriptions of small parts and their interactions. But the essence of the thesis of microreduction is not simply that everything is made out of small parts, as Crane and Mellor are aware. Nor is physics, not by long tradition, mere convention, or even by popular understanding the study of small things. Microreduction says of big things not just that they are made of and so can be reduced to small things, which might be gathered from the fact that the reduction of wholes to parts can be, and perhaps is normally, implied along with the notion that wholes are bigger than parts. But that is not the whole story, and arguably not the most important or interesting part of the story. More interesting, and more faithful to long scientific tradition, is the reduction of specific things and behaviours to general things and behaviours, or of complex things to basic things. The significance of this factor might be gathered from Terence Horgan's phraseology: "Paradigm examples of scientific reduction have a part/whole aspect: laws and phenomena involving complex wholes are explained in terms of laws and phenomena involving the parts of which those wholes are composed (this is called *micro*-reduction)."[21] The reason why this is interesting is that the laws and phenomena involved with parts ought to cover wholes as well, if parts are taken as wholly constitutive of wholes: the laws and phenomena involved with the simpler components of complex things are more general than the laws and phenomena involving the wholes constituted, those typically not picking out phenomena on the level of the parts. The reduction could be construed, therefore, as the reduction of the specific or complex to the more general or basic than to the small *per se*. Though it is obviously not a coincidence that basic things are often very small, it is by no means a necessary condition of basic things that they should be so. Indeed, they may be huge,[22] or of no particular size; they may be forces, fields, or even principles. Physics is not the study of *the smallest things* as a matter of definition, but is more properly to be considered the study of the most general and fundamental of things. That many small fundamental things are the parts of larger, non-fundamental things should not strike us as odd, nor should it strike us as definitional.

I suspect that Crane and Mellor would not concede this point, for the notion that facts about specific, non-basic things actually do reduce to general, basic facts, or that the entities individuated by the special sciences are composed by the entities individuated in physics, is required by microreduction. But microreduction, as they say, "...cannot have it. For unless the sciences of the very large, including psychology, reduce to microphysics, we shall still need to quantify over entities described in those science's

terms. But in fact... even the *physics* of the relatively large does not reduce to microphysics."[23]

Admitting that facts about parts often explain facts about wholes, they go on to say that:

> If for example we take the quantum mechanical description of a quantum ensemble to be complete (as orthodox interpretations do), the superposition principle entails that its properties will not be a function only of those of its isolated constituents plus relations between them. Orthodox quantum physics is not microreductive. And some physics is positively *macro*reductive: Mach's principle, for example, which makes the inertial mass even of microparticles depend on how matter is distributed throughout the universe. We realise of course that Mach's principle and orthodox quantum theory are controversial, and that a future physics might well abandon them. But they cannot be abandoned because they conflict with an MR [microreduction] entailed by modern microphysics: since, as they show, it entails no such thing.[24]

Well enough, but I am not sure that either the superposition principle or Mach's principle provides the counterexample to microreduction that Crane and Mellor seem to think that they do. I am far from an expert on quantum mechanics, or even well versed in the popular understandings of this extremely conceptually difficult science. But according to my understanding of John Gribbin,[25] who is, the superposition principle says that for any quantum particle with a certain probability of being in one of two (or more) states, the quantum is *really* in none of them as such, but is in a *superposition* of those states in which it could be until it is measured. While this is a property that certainly seems to be a little bizarre, I see no reason why it couldn't be a well behaved one, and I fail to see how it provides a counterexample to microreduction or entails that the properties of an ensemble are not a function only of the members of the ensemble and the relations between them. I suppose that it might be entailed if it were taken that any of the quanta in the ensemble were *entangled* with quanta outside the ensemble. Quantum entanglement is, according to my understanding, relatively uncontroversial in quantum physics these days, and the notion behind it is something like this: some particles exist in pairs, and the pairs are bound together such that, if one member of the pair has one of two properties when measured (for until at least one is measured, they are both in a superposition of states) — say, the first particle has spin up — then the other one has spin down, and this is true irrespective of where in the universe either member of the pair might find itself. Further, there is no principle restricting where the members of the pair might be — they may be in extreme and opposite ends of the universe. I suppose that if one member of an entangled pair were a member of a quantum ensemble taken as complete and the other were not, then the measurement of the second particle would cause[26] the collapse of the first into one of two states, and this effect would have its origin outside the ensemble. In that case, all of the properties of the ensemble would not entirely be explained by the properties of the members of the ensemble and their relations, and since the other member of the entangled pair could be anywhere in the universe, it would be quite impossible to set spatial boundaries that would enclose only the ensemble.

What this says to me, however, is that the superposition principle entails that isolated quantum mechanical ensembles cannot be *closed*. Facts about elements outside

the ensemble as isolated, which elements might be unlocated or in principle unlocatable, may potentially have an effect on elements within the ensemble. If, however, the isolated system is taken as being *the entire universe,* I doubt whether orthodox interpretations of quantum mechanics would make the same prediction. What this indicates is that the superposition principle entails not that microreduction is false for isolated quantum ensembles, but that quantum ensembles cannot be isolated within a universe containing other such ensembles such that they are *closed*.[27]

Mach's principle makes similar claims about the inertial mass of objects, as it says that the inertia of any object is relative to the positions and inertias of every other object in the universe. Again according to Gribbin,[28] the idea comes to us from Mach almost unchanged from its formulation in Berkeley's arguments against the absoluteness of space. Berkeley and Mach both held that space could not be taken as a reference against which to measure the motion of objects, and that motion could only be measured against other objects in the universe. The relation is transitive, so, theoretically, the inertia of anything — even microparticles — depends upon the relative inertia of everything else in the entire universe. But this does not entail that facts about inertia do not microreduce, if the frame of reference in which they are taken to be microreducible is the entire universe.

The fact that the physics of a particular 'closed' system do not microreduce to facts only within the system as isolated ought not to be surprising, for this kind of thing is taken as understood in physics all the time. Physicists routinely stipulate systems to be closed for practical reasons, all the while tacitly knowing that they are *really* not closed, but are affected in minute ways from without; often, the forces from without are so minute that they cannot practically figure. Gravity, for example, is currently understood to be a force with infinite range. The fact that predictions of the orbits of the moons of Jupiter do not include the gravitational influence of distant quasars does not make the theory that predicts either the orbits of the moons of Jupiter or the theory of gravity false, nor does it indicate that facts about the orbits of the moons of Jupiter do not microreduce.

What should boundaries of a closed system be taken to be? The idea that *the physical universe* constitutes such a system is common to physicalists, but the notion that the physics of systems taken in isolation within a particular boundary must microreduce to the physics of the elements within the same boundary is not plausible according to modern science (though the notion that *most* of them do *is*). Microreduction involves the idea that phenomena can be individuated on higher or lower levels of description. Though the physics of the very large may often reduce to the physics of the smaller, the boundaries of the system as picked out on the lower level may vastly outreach the boundaries of the system as it is isolated on a higher level, perhaps in principle even involving the entire physical universe. There is no *prima facie* reason why this is incompatible with microreduction, and no reason why we should take it that general principles and facts do not underlie more specific principles and facts.

I suppose that there are some who might take this as somewhat unsettling. If phenomena such as quantum state superposition and quantum entanglement can have

such an effect upon what we might have wanted to think were 'closed' systems, where does this leave us as regards the intrinsic properties of purely physical systems? Much of the literature having to do with psychophysical supervenience, for example, takes it as standard that if two individuals, x and y, are physically identical and embedded in identical environments, they must instantiate the same mental states. It is also a commonly accepted criteria upon causation that the causal powers of a system are intrinsic to that system. But what if there is no way to set the boundaries for the intrinsic properties of *any* system isolated within the universe?

This might appear to be a genuine problem, but I hazard that it isn't. The gravitational effects of distant quasars upon the moons of Jupiter are so small as to be insignificant, and so likewise are the effects of a single quantum, or even quite a sizeable number of quanta, changing states in systems as complicated as rocks, let alone human beings with mental states. In general, whenever a story is told about a macrophysical phenomenon in microphysical terms, one discovers that there is a quite a lot going on that isn't expressed in the macrophysical terms: a meandering river erodes its outside banks, but at the same time, water molecules are absorbed into the matter forming the outside banks; some chemical breakdowns happen; some evaporation into the air occurs; some subatomic particles decay; all kinds of things take place. It is hardly surprising that there aren't expressions for things like this going on on the level that describes 'rivers' and 'banks'. That doesn't mean that there are no rivers or banks, or that it isn't meaningful to say that a meandering river erodes its outside bank. It might be meaningful to say that, if no quantum boundaries can be set on an individual upon whose intrinsic properties some mental state is supposed to supervene, then the only way mental states can supervene is globally, which is arguably not enough for a robust physicalism. But neither is it, in my opinion, reasonable to set the constraints of the realisation of a mental state at a level so basic as a single quantum state or even several such states. Neural systems, relative to quanta, are really very large and complicated indeed. I dare say that the impact of changing quantum states in neural systems is probably about as significant to the functioning of those systems as the gravitational effects of distant quasars are on the orbits of the moons of Jupiter.

Crane and Mellor do not deny that some sciences can be reduced, either actually or in principle, to physical science, or that physics in its current incarnation might be reducible to some undefined future physics. They deny that this gives the physical sciences any ontological authority whatsoever or any right to provide conjuncts to the inventory of what there is; that "if the phenomena of psychology are less ontologically acceptable than those of physics and chemistry, it cannot be because psychology is irreducible to present or future physics."[29] I have argued that *none* of the non-basic physical sciences have that authority, and that the only physical science that would have that authority, the one to which all non-basic physical sciences are necessarily reducible, is not one which we are in a position to define. I agree with Crane and Mellor, therefore, that there are serious difficulties, probably insurmountable ones, involved with defining physicalism in terms of the possibility of reducing all empirical sciences, where that is taken as meaning all sciences that study things observable in the world, as reducible to

physics, but I do not agree that this implies that there is no reason why psychology cannot add to the world, in its own terms, its proprietary entities and generalisations. The examples offered to show how even physics does not microreduce are not conclusive without a further requirement that the boundaries of microsystems must match the boundaries of macrosystems for ensembles isolated within the physical universe, for which requirement I can foresee no convincing argument. Further, there is a question about the notion of reduction itself that Crane and Mellor employ. Their model for reduction is Hempel's,[30] according to which a science is reducible to physics if all of the laws of the target science can be derived from the laws of physics supplemented with appropriate 'bridge principles' linking the terms of the target science to the terms of physics.[31] This is very similar, almost identical in fact, to the Nagelian model of reduction by derivation via bridge laws, which has been standard in the literature for some time. Yet there is a question as to whether derivational reduction of this type is appropriate when some of the terms of higher-level sciences are interpreted functionally. Indeed, very convincing arguments against reductionist physicalism taking intertheoretic reduction as definitive of reduction have been forwarded on the basis that sciences featuring multiply instantiable functional kinds as their subject matter cannot be reduced to physics. This may not prove to be a difficulty according to a model of reduction sensitive to functional higher level properties,[32] and such models are possible. For example, Jaegwon Kim has recently forwarded a model of reductionism focusing on the reduction of properties rather than on the reduction of sciences. According to Kim's model, the key notion in whether or not a property described by a special science is reducible to physical properties is whether or not that property can be construed extrinsically and defined as a second-order functional property. This model will be a topic of more detailed discussion later,[33] so I propose to leave it here.

1.2. Mental Causation and Intentionality

The mark of the mental, as Brentano wrote,[34] is its intentionality: mental states are essentially representational states, individuated by their content properties, whereas physical states are typically (with caveats, of course) individuated by their intrinsic properties. The problem this creates for mental causation is that we generally expect the behaviour of a system to be determined by the intrinsic properties of that system in response to inputs to it, and not by relational, extrinsic properties of the system.

Crane and Mellor accept that extrinsic properties cannot feature in explanations of a system's behaviour, but they deny that this means that content properties cannot cause events. "All it means is that they must do so indirectly, via a mental representation, i.e. via some intrinsic non-relational property of the mental state."[35] Such an intrinsic property, correlated with its object in such a way as to give it "its right causes and effects",[36] they term a 'causal surrogate' for the object of the mental state, wherever that may happen to be and whatever its ontological status.

I must admit that I find this talk of causal surrogates a little strange. If a 'local

causal surrogate' is required to do the causal work of the extrinsic content properties of a mental state, then surely it is not the content properties of the mental state at all that are doing the causal work, but the intrinsic properties that are standing in as causal surrogates for the content properties only. Crane and Mellor might object that since the local causal surrogate is appropriately related to the object of the mental state, the object does in fact play a role in the determination of behaviour. Two obvious objections can be made to this approach, one from consideration of the 'causal exclusion problem' raised by Jaegwon Kim, and one from the moral of the well known 'Twin Earth' story as forwarded by Hilary Putnam.[37]

Kim formulates the causal exclusion problem as follows:

> Suppose then that mental event m, occurring at time t, causes physical event p. And let us suppose that this causal relation holds in virtue of the fact that m is an event of mental kind M and p an event of physical kind P. Does p also have a physical cause at t, an event of some physical kind N? ...to acknowledge that p has also a physical cause, p*, at t is to invite the question: given that p has a physical cause p*, what causal work is left for m to contribute?[38]

The application of this problem to the notion of a causal surrogate is transparent: given that an intrinsic causal surrogate is required to produce the behaviour of an intentional agent, surely it is all that needs to be cited in a complete explanation of that behaviour? Surely it is sufficient to give an explanation invoking *only* the causal surrogate properties, and not to refer to the content properties of the mental state, which would be cited in an explanation couched in psychological terms, at all? It might be objected that since the intrinsic property in question is somehow correlated with the extrinsic content property in question, that extrinsic property is a feature in a complete explanation of an intentional agent's behaviour. But if Hilary Putnam's Twin Earth scenario has taught us anything, it is that this sort of correlation is highly problematic indeed: however correlated, content 'just ain't in the head'.[39]

Crane and Mellor would not accept that these arguments create any difficulty for the causal efficacy of the mental *per se,* for they see no reason why intrinsic, nonrelational properties of individuals might not themselves be mental. Such intrinsic properties could, on their account, "be sensations, or visual or mental images or models".[40] This should immediately strike us as odd, for since Descartes, a major problem in the philosophy of mind has been how to make sense of causal interaction between physical things and things that have none of the properties that are apparently necessary to produce effects in physical things, and *vice versa.* But Crane and Mellor appear almost to see this difficulty as merely stipulative:

> Perhaps the physical just is the causal, and what physicalism really means is that the empirical world comprises all and only those entities, properties, relations, and facts which have causes or effects. This definition clearly underlies one familiar formulation of the mind-body problem: how can mental states have effects in a physical world? This question would not pose such a problem if it were not assumed that causation is essentially non-mental.
> ...But why should we assume this? It is surely obvious that there is plenty of mental causation.[41]

This is, to my mind, to miss the point of the physicalist's position. Physicalists do not typically deny that there is mental causation. It is, however, a desideratum of

physicalism that descriptions of chains of events involving physical properties be explanatorily sufficient: that they should, to put it crudely, make sense. Thus, physicalists are concerned to spell out precisely *how* the mental, *qua* mental, can be causally effective with regard to the physical. It is fairly well received that causal interaction between physical things and nonphysical things cannot meet with this desideratum. Thus, the physicalist asserts that if indeed mental properties are causally effective with regard to physical things, they must under some description themselves be physical things. This, I take it, is the source of the claim that "the physical just is the causal, and what physicalism really means is that that the empirical world comprises all and only those entities, properties, relations, and facts which have causes or effects", to which Crane and Mellor take exception. But to counter this with the observation that "there is plenty of mental causation" is simply to deny the physicalist desideratum of explanatory sufficiency. Recall that the object of Crane's and Mellor's argument here is to show that the notion of causal efficiency is not one that can set psychological properties, and so the sciences that describe them, apart from physical properties in such a way as to deny the psychological sciences the authority to add to the world's ontology, *in their own terms,* their proprietary elements. Their claim is that, while it is assumed that causation is essentially non-mental, there is really no reason for the assumption, and it is obvious that mental properties such as sensations or mental representations often cause physical events. To the physicalist's objection that this kind of interaction is fundamentally mysterious, Crane and Mellor answer: "It is indeed an old thought that mental causation is hard to make sense of, and especially causation linking the mental to the non-mental, because they seem to be so different. But why should that impress anyone who has learned from Hume that causation never 'makes sense': that it is always a matter of fact, not of reason?"[42]

There is, of course, not much to be said to a Humean sceptic on the subject of causation, but while the question of the ultimate status of the causal relation and the question of the desirability of explanations that make explanatory sense are related, they are hardly identical. One can, I think, subscribe to physicalism without having any commitment one way or the other about the status of the causal relation, but physicalism is motivated in large part by the desideratum that properties or events construed as related be related in a way that makes explanatory sense. To reject this and to assert that relations of a non-explanatory nature just simply obtain is ultimately to reject the possibility of explanation, even in the basic physical sciences: if nonphysical events can in their own terms interrupt purely physical chains of events, then explanations in physics can not be understood to be sufficient to describe all kinds of physical events. If Crane and Mellor are prepared to allow that causal interactions between the type described straightforwardly take place, then they cannot countenance the possibility of fully explanatory theories of anything, or of generalisations that would not be subject to falsification by miracles. Rather than demonstrating the continuity of psychology with the physical sciences, I think that this approach ultimately robs all the sciences of their explanatory power. This is not, I believe, a conclusion that many will find acceptable.

Another feature of intentional mental states that is thought to set them apart from

physical states is that they involve non-extensional contexts, or in other words, that coreferential substitutions for terms used in the ascription of a mental state are not necessarily truth-preserving. For example, Ed may believe that Gustav is a very nice man. But Ed may not believe that the King of Sweden is a very nice man, if he does not know that Gustav is the King of Sweden. So while we can truthfully substitute 'Gustav' for 'the King of Sweden' for statements in nonmental contexts — "Gustav is a very nice man" is just as true (or false) as "The King of Sweden is a very nice man" — we cannot make this substitution for mental contexts. If Ed does not believe that Gustav is the King of Sweden, then the truth values of "Ed thinks that Gustav is a very nice man" and "Ed thinks that the King of Sweden is a very nice man" are completely independent. He may believe one, the other, both, or neither, even if Gustav is indeed the King of Sweden and is also a nice man.

Crane and Mellor deny that this is a unique feature of the mental, "since non-extensionality occurs in physics too. This is because laws entail non-extensional conditionals."[43] They go on to give the following example:

> Suppose for example that *H* and *K* are the genes that give us hearts and kidneys. The fact that we all have both does not make 'anyone who had gene *H* would have a heart' entail either 'anyone who had gene *K* would have a heart' or 'anyone who had gene *H* would have a kidney'.[44]

I fail to see how this means that laws entail nonextensional conditionals. If it is as they say just a *fact* that we all have both hearts and kidneys, and so the genes that are responsible for our having them, and if it is as they seem to imply just a *fact* that we all have the genes responsible for our having hearts and at the same time those responsible for our having kidneys, then there is no law connecting our having hearts to our having kidneys, and so no law connecting our having gene H with our having gene K. Hence there is no law to create a nonextensional context.

The correlation between the our having hearts and our having kidneys might generate a nonextensional context if indeed it were a law, but that any science would countenance a law correlating the presence of fully formed hearts and kidneys in adult human beings is extremely unlikely. I suspect, however, that it is not just a *fact* that all human beings have both hearts and kidneys. Rather, it seems likely that there are some biological features that necessitate the presence of both hearts and kidneys in creatures such as ourselves.[45] We may, then, loosely grant that there *is* a law correlating the having of both hearts and kidneys in us. If there is such a law, however, it must be a law about the biological constitution of certain kinds of animals to the effect that they are so constituted that they have both hearts and kidneys, among other things. This law would be resolved on the level of species and individuals to species-specific laws about the constitution of individual creatures, and as such, it would correlate the mechanisms responsible for the having of both hearts and kidneys in healthy representatives of the species (or in an individual typical of the species). In such a case, 'anyone who had gene *H* would have a heart' *does* entail that anyone who had gene *H* would have a kidney, via the suppressed 'anyone who had gene *H* would have gene *K*', and 'anyone who had gene *K* would have a kidney'. Thus in extension, the substitution is truth-preserving, but is not

in intension, for if anyone did not know the generalisation, one could believe that all mammals had hearts, but not necessarily that they had kidneys, even though there is a law (if there is) to the effect that they all do.[46]

Crane and Mellor argue further that probabilistic laws such as are employed in quantum mechanics create nonextensional contexts. They say:

> The probabilistic laws of modern microphysics cannot be extensional for another reason, too, because '*p*(...)=n' is not extensional: for if it were, '*a* is the *F*' and the necessary truth '*p*(*a* is *a*)=1' would entail '*p* (*a* is the *F*)=1, which it clearly does not, on any view of probability (take for example *F*= 'next Prime Minister').[47]

This is a very curious claim, and Crane and Mellor do not explicate any further than the quotation indicates. If I read them correctly, what is meant is something like this: we cannot make truth-preserving substitutions between propositions and conjunctions involving those propositions where either involves a statement of probability. Take the two propositions 'Tony Blair is Tony Blair' and 'Tony Blair is the next Prime Minister'. That Tony Blair is himself is a necessary truth, but that he is the next Prime Minister is a lesser certainty. So, p(Tony Blair is Tony Blair)=1, but p(Tony Blair is the next Prime Minister)=1-n.

Now, suppose that Tony Blair *is* the next Prime Minister. Since, however, he is so with probability 1-n, there is a possible world in which he is not the next Prime Minister. So, although "Tony Blair" and "The next Prime Minister" refer to the same thing in this world, but not to the same thing in some alternate possible world where something other than Tony Blair is the next Prime Minister or in which there are no Prime Ministers, the conjunction is not truth-preserving across worlds. That is, neither

p((Tony Blair is Tony Blair)&(Tony Blair is the next Prime Minister))=1

nor

p((Tony Blair is Tony Blair)&(Tony Blair is the next Prime Minister))=1-n

is true.

If this is correct, then the argument might be very persuasive towards the conclusion that the creation of nonextensional contexts is not a feature of the mental that sets psychology apart from physics. But I think that to read the example as applicable to the probabilistic laws of quantum mechanics would be to misinterpret the role of probability in quantum mechanics. Given that Tony Blair is the next Prime Minister in some possible world W1, but that he is not the next Prime Minister in some other possible world W2, surely it is the case that *in W1,* p(Tony Blair is the next Prime Minister)=1, and that *in W2,* p(Tony Blair is the next Prime Minister)=0, or that, within possible worlds, the substitution of 'Tony Blair' for 'The Next Prime Minister' is truth-preserving in nonmental contexts, for probability statements involving whether or not

Tony Blair is the next Prime Minister do not hold cross-worlds. However, this is not the case for quantum events with probabilistic outcomes. Suppose that it is a law of quantum mechanics that particles of type α decay within a certain time t with probability 1-n.

Then if x is a particle of type α, p(x will decay within t)=1-n is true across all possible worlds, because it is an intrinsic feature of such particles that the laws concerning when they decay are probabilistic. It seems deeply unlikely that there is any such intrinsic feature of Tony Blair or of any candidate for Prime Minister, but that rather whether or not someone is the next Prime Minister and the probability with which they are one or the other are relational properties individuals possess by virtue of their embedment within a society with certain electoral procedures and having members with certain political leanings. But if there were some intrinsic feature of candidates for Prime Minister having to do with the likelihood of their actual election, as there are intrinsic features of particles having to do with their actual decay, then p(Tony Blair is the next Prime Minister)=1-n would be true across possible worlds.

This would mean that when the property of having a certain probability to be F is an intrinsic property of a thing such that laws about events involving such things are stochastic, the conjunctions p[(*a* is *a*)&(*a* is the *F*)]=1 or that p[(*a* is *a*)&(*a* is the *F*)]=1-n are misformed. The conjunction properly ought to be [(*a* is *a*)&(p(*a* is the *F*)=1-n)], which, it should be clear, does not generate nonextensional contexts in nonmental cases. But since a subject could fail to be aware that (*a* is the *F*) with any degree of probability, the substitution of '*a*' for '*the F*' cannot be made for intensional states, as far as I can see, uniquely.

1.3. Psychological and Psychophysical Laws

Crane and Mellor next turn to the possibility that psychology can be denied ontological authority if there are no strict psychological or psychophysical laws. They say: "The ontological authority of a science arguably rests on the laws it discovers, which tell us what kinds of things there are, and what properties and relations distinguish them. But many agree with Davidson that the mental is 'anomalous': that strictly speaking there are no psychological or psychophysical laws. If that were so, psychology would add nothing to our ontology of non-mental kinds, with their distinctive properties and relations."[48] They consider four possible avenues from which such a denial might come: 1) the idea that laws are necessarily true while generalisations about the mental are not, and that there are no necessarily true generalisations linking mental terms to mental or to physical terms; 2) that physical laws are subject to excogitation whereas psychological generalisations are not; 3) that the mental is multiply instantiable; and 4) the arguments presented by Donald Davidson for anomalous monism. I agree with Crane and Mellor that 1) and 2) are unconvincing, if not entirely for the same reasons, and in the interests of brevity propose to let those stand. A defence of anomalous monism against Crane's and Mellor's highly general critique would be substantial diversion and is quite outwith

the scope of this thesis, though I believe that such a defence could be sustained. In any case, I am convinced by the arguments forwarded by Kim and others to the effect that anomalous monism is ultimately a form of epiphenomenalism, leaving no genuine causal work for the mental to perform in the world. Since the multiple instantiability of the mental is a topic to which we will return in detail, I would like to attend briefly to 3), the notion that the multiple instantiability of the mental could ground the assertion that there are no psychological or psychophysical laws.

Crane and Mellor write:

> Another bad reason for denying the existence of psychophysical laws is the so-called 'variable realisation' of mental states: the fact that 'the range of physical states fit to realise a given mental state can be indefinitely various'. That cannot stop psychophysical generalisations being laws. For if it did, there would be hardly any laws in physics either. States like masses, volumes, and temperatures are even more variously realised than mental states: one can have a gram or a litre of almost anything, at any one of an indenumerable infinity of temperatures. So if variable realisations does not rule out laws in mechanics and thermodynamics, it can hardly rule them out in psychology.[49]

This may immediately appear plausible, but in fact the multiple instantiability of the mental is quite a different form of multiple instantiability than that of such states as masses, temperatures, and volumes. The problem generated for physicalism by the multiple instantiability of mental states is that, if mental states are instantiable in wildly heterogeneous physical bases — if a particular mental state, such as a sensation of being in pain, can be realised in me, in my dog, in an octopus, or in a silicon-based Martian life form — then physical explanations of actions involving individuals in pain will take different forms. For me, explanations in terms of human neuroanatomy will be relevant; for my dog and for the octopus, dog and octopus neuroanatomy; for the Martian, explanations in terms of whatever physical mechanisms detect damage to the Martian body and initiate damage-avoidance behaviour routines will be relevant. There is no one physical theory that can explain *pain behaviour*. Thus pain, *qua* pain, is not reducible to any physical type, wherefore laws involving pain *qua* pain cannot be reduced to physics.

It should be clear that the heterogeneity of the realisation bases of such states as *having a mass of 1 gram*, even though all kinds of things can possess this property, does not indicate that sciences in which *having a mass of one gram* can feature are irreducible to physical sciences. Firstly, we do not generally think of things massing one gram as being kinds of things, or of the property *massing one gram* as a special property defined by its causal role, whereas we do tend to think of mental states as being kinds of things so defined. Secondly, explanations of causal transactions involving only such multiply realised properties as masses, volumes, and temperatures will take generally similar forms, no matter what it is that has some mass or some volume or some temperature: they will all most appropriately be explained in sciences of macroscopic mechanics. Further, though there may well be different explanations for these kinds of causal transactions, and indeed there may be different, mutually incompatible sciences of macroscopic mechanics, an explanation for *any* causal transaction involving masses, volumes, temperatures and the like could be given in *any* science of macroscopic mechanics. This could not be the case for explanations involving mental properties *qua* mental, if mental

properties are multiply instantiable. Though human neuroanatomy may be appropriate to explain some of my actions, it is useless for explaining any of a Martian's. *Qua being in pain,* our physical descriptions are far too diverse.

Unlike the multiple instantiability of mass, volume, temperature, and other such properties, the multiple instantiability of mental properties like *being in pain* in entities with very different physical descriptions prevents *being in pain* from being treated in general physical terms. This is enough to qualify sciences that deal with such multiply instantiable properties as noncontinuous with the physical sciences, if continuity is judged by the generality of laws about causal transactions.

1.4. Supervenience

The notion of mind/body supervenience is essential to most versions of nonreductive physicalism. Nonreductive physicalists claim that mental states are dependent upon and realised by physical states, but that they are in principle irreducible to physical states and relations. If they are right, then this irreducibility would ground a claim of the noncontinuity of psychology with the physical sciences.

A standard definition of the supervenience relation is as follows: A family of properties A is supervenient upon a family of properties B just in case for any property F in A, if some x has F, then there is some property G in B such that x has G; further, if any y has G then y has F. It follows that if any two things are A-identical, they are B-identical. For the psychophysical case, this means that there can be no change in the mental state of an individual without there being some change in the physical state of that individual. This is a rather intuitive notion that should appeal to any physicalistically inclined person.

Crane and Mellor argue that there is no empirical support for the notion of supervenience: "The evidence for it cannot be empirical, since the prospect of ever finding two things, complex enough to have psychological properties, type identical in every reasonable non-mental respect is very slight, to say the least."[50] They find the only sensible argument that could be made for supervenience the notion that there can be no unmediated action at a distance, but even this is not enough, for, as they had claimed earlier, the local causal surrogates for extrinsic properties could themselves be brute mental states, not supervenient upon physical states. If the only argument for psychophysical supervenience is the idea that there can be no unmediated action at a distance, but mental states can be local causes of behaviour in a system, then clearly there is no further reason to postulate a relation of supervenience between mental and physical properties. Further, they argue, mental states will still fail to supervene upon physical states, for the reasons given in Putnam's Twin Earth argument to the effect that the content of a mental state will not supervene directly upon the intrinsic properties of an individual, but upon the intrinsic properties of an individual and that individual's relation to their environment.

They do note that proponents of supervenience might attempt to formulate the

thesis so as to account for this, and indeed, such formulations are in the offing. Terence Horgan, for example, has constructed a tri-level version of supervenience as follows:

> Any two physically possible worlds which are exactly alike physically are also exactly alike in all other respects.
>
> For any two individuals *i* and *j*, either in distinct physically possible worlds or in a single such world, if *i* and *j* are exactly alike in all intrinsic physical respects then they are exactly alike in all other intrinsic respects.
>
> For any two individuals *i* and *j*, either in distinct physically possible worlds or a single such world, if (i) *i* and *j* are exactly alike in all intrinsic physical respects, (ii) *i* has a non-intrinsic property F, and (iii) *i* and *j* are exactly alike with respect to all non-physical features which are pertinent to *i*'s possession of F, then *j* also possesses F.[5]

Crane and Mellor object that consideration of the ways in which thinkers err will show the falsity of such formulations: "Suppose for example that you and your physically identical twin now look at the same elm, but that although this makes you think that it is an elm, it makes your twin think that it is an oak. Same intrinsic properties, same relations, same properties of the thing thought about: but different thoughts."

Such a scenario is precluded by Horgan's formulation, and would falsify the formulation if it could be shown to be possible. However, I do not think that it is possible; indeed, I find the scenario as Crane and Mellor describe it wildly counterintuitive. If *every property* of the two individuals is supposed to be the same, including all of their histories and educations regarding the identification of trees, and if they are related in identical ways to identical trees, how *possibly* could one individual think "That's an elm" and the other "That's an oak"? Perhaps the counterintuitive nature is better exposed if we take the scenario slightly differently: suppose that I am now standing in front of an elm on a clear, bright day, and, being in a frame of mind to name trees, I think 'That's an elm." Presumably, the series of events producing this thought take a little bit of time: at t1, I focus my attention in an elmward direction; light bounces off of the elm into my retinas; some rational thought process takes place wherein I compare the characteristics of the tree to various tree-concepts I possess; &c until I come out with 'That's an elm' at t2. Now suppose that time goes backwards to t1 again, and the same process occurs (in the original scenario, my twin is just as identical to me at t1 as this). How *possibly* can I now come out with 'That's an oak'? Even if it *is* an oak and I am making a mistake, I should make the *same* mistake upon repetition of the scenario: same tree in the same relation to me; same light; same rational thought process; same tree-concepts: same outcome. I believe that the burden is upon anyone who would assert that anything other that this could happen to show *how* it could happen.

This consideration does not falsify supervenience. However, Crane and Mellor claim that modern physics itself, with its probabilistic laws, does falsify supervenience. Their example is as follows:

> Suppose an intrinsic non-mental property P causes a mental property M indeterministically. (Say for example that one's being in M at t2 is 0.9 if one has just been P (at t1) and 0.1 if one has not). Now suppose that at t1 many people share *all* their intrinsic non-mental properties, including P. At t2, therefore, most but not all of them will be M: that is, some pairs of people, atom-for-atom alike at

> t1, will differ at t2 in this mental respect...
> In other words, modern indeterministic physics must predict that some pairs of people, atom-for-atom alike in all non-mental respects, will differ in some simultaneous mental respects: and will do so precisely because the properties involved are causally related."[5]

This does not work. First, there is an equivocation: in the first part of the example, the relation involved is causal: an instance of some physical state P *causing* some mental state M with a certain probability. In the second part, two atom-for-atom identical individuals are supposed to diverge in mental states because of the supposed indeterminism of the *supervenience* of a mental state by a physical state. But causation and supervenience are hardly the same relation. Supervenience is almost universally taken to be a synchronic relation, whereas the probabilistically interruptible relation required by Crane's and Mellor's argument would certainly have to be diachronic.

Crane and Mellor might change the example, and argue that the supervenience of rationality, or any other mental *process,* fails to supervene upon the physical because of the probabilistic causal relations between certain kinds of physical events. Suppose that two individuals *x* and *y* are in a particular physical state *P* at *t1*. When presented with some stimulus (let's call it *q*), physical state *P1* has a 0.9% probability of going to state *P2,* and a 0.1% probability of going to state *P3*. Then it is quite clearly possible for two physically identical individuals with identical causal histories to be presented with the exactly similar stimulus *q* and yet respond differently: 9 out of 10 individuals in *P1* at *t1,* when presented with *q,* will move to *P2,* while one will move to *P3*. This physical difference could ground a mental difference: some mental state *M2* might be supervenient upon *P2,* while a different mental state or no mental state at all might supervene upon *P3*.

This objection will not work either, for it isn't at all clear that supervenience cannot accommodate probabilistic physics. Consider the following:

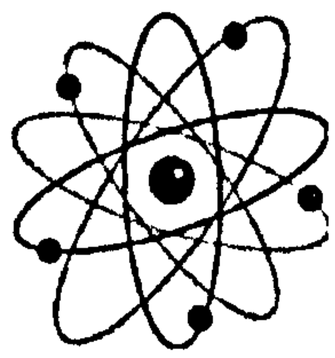
> Necessarily, for any object x and any property F in A, if x has F, then there exists some property G in B such that x has G, and necessarily, if any y has G, then there is a probability *n* that y has F.

I do not see that there is anything intrinsically wrong with this formulation. Some might be uncomfortable accepting it for the psychophysical case, for it would make the determination of mental states by physical states only probabilistic: there would be nothing to guarantee that rational thought might not be interrupted by a random quantum occurrence, and nothing to ensure that sightings of brown cows would produce brown-cow-thoughts, if some random quantum event could get in the way and replace the brown cow thought with, say, a John Major thought. But provided the probability of y's having F when y has G was high enough (say, 99.9999999%), this would make the worry all but ridiculous, and were the probability lower, it might explain quite a bit of human irrationality.

Further, it must again be pointed out that practically speaking, quanta and quantum events and their effects are *incredibly tiny* in comparison to anything that might reasonably be postulated to realise complex mental states. Thus, a particular random quantum event would have such a small impact upon rational thought or the realisation of

anything so grandly complex as a sensation or a mental representation as to be negligible, as the gravitational effects of distant quasars are theoretically a factor, albeit a negligible one, in tides on Earth. It is, of course, possible that enough co-ordinated quantum randomness could produce a quite unanticipated mental anomaly. I cannot myself perform the calculations, but I dare say that the probability of such an event occurring within the lifetime of the universe is vanishingly small.[3]

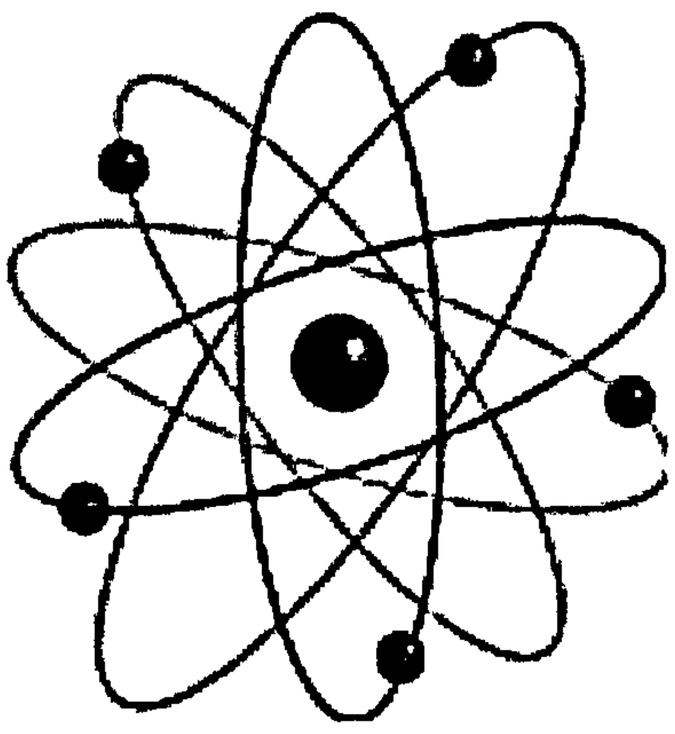
In conclusion, I find that Crane and Mellor have failed to provide convincing arguments against the noncontinuity of psychology with the physical sciences. They have failed satisfactorily to show that the idea of reducibility to physics is so ill defined as to suggest that the reducibility or irreducibility of a science to physics is insufficient ground to say that irreducible sciences are noncontinuous with physics. According to physicalism about the mind, the content properties of mental states must, indeed, *in some sense* be physical intrinsic properties of mindful systems, and that the postulation of causally effective brute sensations is, if not simply philosophically unsatisfying, explanatorily disastrous. Therefore the intentionality of mental states and the essential reference psychology makes to the contents of mental states could very well ground the autonomy of psychology *vis-à-vis* physics. Crane's and Mellor's arguments against the thesis of supervenience have proven to be inadequate. So if indeed the relation of supervenience can make sensible the thesis that mental properties can be realised in the physical, but be irreducible to physical phenomena subsumed by physical laws (though there is some question as to whether indeed it can do this), then the nonreductive physicalists have a case for the autonomy of psychology. In short: there is still very much a question of the epistemological doctrine of physicalism.



2

The Physical

2. Physicalism as a Metaphysical Doctrine



CRANE AND MELLOR HAD TAKEN PHYSICALISM to be an epistemological rather than a metaphysical doctrine, focusing their arguments on the possibility of the contiguity or noncontiguity of the special sciences, particularly psychology, with physics. However, physicalism can also be taken as a metaphysical doctrine limiting the contents of the world to one prime ontological category in which all phenomena consist or are realised. In a recent paper,[1] Chris Daly argues that no principled and well defined distinction between physical and nonphysical properties can be drawn, and that therefore, all metaphysical projects that assume such a distinction should be abandoned. Correctly, Daly asserts that physicalism must provide such a distinction in order to avoid vacuousness.[2] His arguments for the rejection of physicalism as a viable doctrine are thus quite different in character from those presented by Crane and Mellor.

Three claims characterise the thesis of physicalism for Daly. Given that physicalism must provide a distinction between two classes of properties, (1) physical properties and (2) properties which are not physical, he identifies the central claims of physicalism as follows:

> (i) (at least) all actual individuals are physical individuals, and all actual properties are physical properties;
>
> (ii) there is a relation (or a family of relations) R such that if any property F of type (2) exists at a world w, there are properties G1,..., Gn of type (1) at world w, and R holds between F and G1..., Gn at w, then F is a physical property at w; and

(iii) For every property F of type (2) which exists at the actual world, there are properties G1,..., Gn of type (1) which exist at the actual world, such that R holds between F and G1,..., Gn at the actual world.[3]

In order to be nonvacuous, physicalism must be able to distinguish between properties of type(1), physical properties, and properties of type (2), properties that are not physical. (ii) claims that any property F of type (2) — that is, any nonphysical property — *is in fact a property of type (1) — a physical property* — if it exists at a world such that G1,..., Gn also exist at that world and some relation (Daly is ambivalent as to what, specifically, this relation is specified to be) obtains between G1,..., Gn and F. (iii) claims that in the actual world, all nonphysical properties are so related to physical properties in such a way and so are in fact physical properties.

This formulation is a bit too strong for the minimal requirements of physicalism. It is not necessary to say that all apparently nonphysical properties are *really* physical properties. It is consistent with minimal physicalism to assert that there are *really* properties that are nonphysical, requiring only that any such properties be wholly realised or constituted by physical properties. Physically realised properties must have a complete physical description, but in themselves, may have no interesting features identifiable on the physical level, and so in themselves may be considered nonphysical properties. For example, the property *being a token of a monetary unit* is not a straightforwardly physical property, because any number of physical things might serve to realise that property: bits of paper, pieces of metal, even wholly electronic processes, can suffice for being a token of a monetary unit, as potentially could be all manner of physical thing. It follows that there is no way of identifying a monetary unit merely by its physical features, and so that *being a monetary unit* is not straightforwardly a physical property. However, there are never tokens of monetary units that do not have something physical about them, so each instantiation of the property *being a monetary unit* is ontologically grounded in some physical property or other. It seems that meaningful things can be said about monetary units and their interactions in economic systems without necessary reference to the physical phenomena with which they correlate, and the idea that the property *being a monetary unit* — or any similarly multiply realisable property, such as *being a belief that Ralph Nader should win the election* — is *really* a physical property is far too strong. I would prefer to weaken Daly's formulation to reflect this notion:

(i*a*) (at least) all actual objects, properties and states either are physical objects, properties or states or are nonphysical ones wholly realised by physical objects, properties and states;[4]

(ii*a*) there is a relation (or a family of relations) R such that if any nonphysical property exists at a world w, there are properties G1,..., Gn of type (1) at world w, and R holds between F and G1..., Gn at w, then F is wholly realised by physical properties; and

(iii*a*) For every nonphysical property F which exists at the actual world, there are physical properties G1,..., Gn which exist at the actual world such that R holds between F and G1,..., Gn at the actual world.

I believe that this weaker formulation will answer to most physicalists' requirements without burdening them with implications that they would rather not countenance, and is therefore truer to physicalism. It should be noted that this weakening of the formulation of physicalism does not affect Daly's arguments, so anyone objecting to his formulation of the physicalist thesis could read his arguments as applying to this perhaps more agreeable version.

Daly considers several methods of defining what it is for a property to be a physical property. He immediately rejects the notion that paradigm physical properties might simply be picked out by ostention, for the simple reason that physical properties are "a large and varied bunch",[5] not all of which are observable, and many of which currently recognised physical properties are not the sorts of properties we might intuitively classify as physical. Neither can physical properties be picked out by reference to the kinds of laws in which they feature, as for any kind of constraint we put upon such laws, we could certainly find examples of nonphysical properties that feature in similar calculations of lawlike interactions between such properties.

He is quite right to reject both of these approaches. As he says, we could not possibly pick out all the physical properties by ostention, and further, without adding a constructive element to the description of what it is to be a physical property by which we could project nonoccurrent but possible physical properties, this method would be in danger of failing to recognise merely possible physical properties. I take it that such a property as *'being a sphere of uranium five miles in diameter'* would be a perfectly well behaved physical property, even if it is not one that is instantiated in the actual world. It would not do for a formulation of physicalism to exclude such properties from its ontology. Daly is likewise right to reject the idea that physical properties miht be picked out as uniquely featuring in functional laws. He considers that one might take it that physical properties are characterised by featuring in functional laws, and so by being represented by differential equations. But surely anything to which a quantity could be assigned could be treated in just the same way as physical quantities for the purposes of study.[6] Clearly, we can assign quantities to all manner of nonphysical things: values to monetary units, numerical probabilities to degrees of belief, and so on, and so render these paradigmatically nonphysical properties subject to treatment in terms of functional laws expressed in differential equations.

These simpler accounts rejected, Daly goes on to consider some more complex and promising means of distinguishing physical properties: an account from family resemblances, credited to Frank Jackson and David Papineau;[7] an identification of the physical as an extremely basic natural kind constructed by Paul Snowdon; an identification of the physical from a pre-theoretically given class of paradigm physical effects forwarded in Papineau's Philosophical Naturalism;[8] and the idea of physical properties as those properties studied by the physical sciences, citing formulations by

Geoffrey Hellman[9] and Jeffrey Poland.[10] Finding these accounts also wanting, Daly concludes that we have no satisfactory definition of what it is to be a physical property that will serve the purposes of analytical metaphysics, and that therefore all theses that assume such a distinction are not well defined. In light of this, he urges that philosophical effort be concentrated on those problems that would arise independently of an assumed distinction between the physical and the nonphysical. I intend to examine these accounts and Daly's criticisms in some detail, and ask whether Daly's 'downbeat' conclusions are indeed motivated by his aruguments.

2.1. The Family Resemblance Account

According to the 'Family Resemblance' account of Jackson and Papineau,

> if *A*, *B* and *C* are paradigm physical properties at the actual word, then a property $\mathcal{J}$ is also a physical property if there is a series of properties *D*, *E*, *F*, *G*, *H*, *I* which carry a chain of suitable family resemblances from any of *A*, *B*, or *C* through to $\mathcal{J}$.[11]

This account has a few features in its favour: for one, it does not require that physical properties all share any particular characteristic, but only that they are all related by some chain of resemblances to any of the paradigm physical properties {A, B, C}. It might express how we actually apply the concept of the physical: Daly notes that the account could be supported by Kuhn's argument to the effect that scientists do not acquire abstract theoretical concepts by studying strict definitions and principles, but rather become familiar with the categories in their purview through experience and exposure in their work.[12] However, Daly raises problems for the account which cast serious doubt over its fitness for physicalist purposes.

First of all, it is all very well and good to say that physical properties are related through chains of resemblances, but then we need to know in what at least some of these resemblance relations consist, and we need a convincing reason to believe that this relation will do the work that it must do to make physicalism a nonvacuous metaphysical doctrine — to provide a principled distinction between physical properties and nonphysical properties. But as Daly says, "As matters stand, we have no reason to suppose that any two physical properties are linked by suitable family resemblances, and that these same resemblances do not also hold between the members of of some pair of properties, at least one of which undoubtedly is not a physical property."[13]

Suppose that there are physical properties *{A, B, C}*, *D*, *E*, *F*, *G*, *H*, *I*, $\mathcal{J}$ in the actual world, and that *D*, *E*, *F*, *G*, *H*, *I*, $\mathcal{J}$ are all linked by suitable family resemblance relations to *A*, *B*, and *C*, which will serve as our paradigm physical properties. For clarity, Daly asks us to suppose that $\mathcal{J}$ is the property *being an electromagnetic wave,* as that is a property that we would unhesitatingly classify as a physical property, but which is not among the historically paradigmatically physical properties (as, presumably, such properties as *being*

extended in space and *having location* are). Now imagine an alternative possible world W_1 which contains fewer properties: this world has only *A, B, C,* and *J*. In W_1 there are not enough properties to carry the family resemblance relation through to *J*, so on the family resemblance account, *J* is not a physical property in W_1, but we have specified that it is. Now consider a paradigmatically nonphysical property *Z,* for which in the actual world there are not properties carrying the resemblance relation from {*A, B, C*} through to *Z*. But there is another possible world W_2 which has so many properties, *A, B, C, D, E, F, G, H, I, J*... and so on, which carry teh resemblance relation to *Z*. Thus the family resemblance account mistakenly identifies *Z* as a physical property in W_2.

An advocate of the family resemblance account might answer Daly along two lines. They might a) restrict 'any possible world' to the actual world, or b) allow that if in some possible world a suitable family resemblance chain relating *A, B, C, D, E, F, G, H, I, J, K... Z* obtains, then *Z* is a physical property, but constrain the resemblance relation such that there is a principled reason why certain kinds of properties, which will be nonphysical in all possible worlds, cannot stand in such a relation to the paradigmatically physical properties, no matter how far removed. Such constraints might be constructed from various avenues: the failure of qualitative information to 'fall out' of physical descriptions as illustrated by Levine,[14] or the intractable conceptual difference between the subjective and the objective as outlined by Thomas Nagel,[15] for example.

Of these two possible avenues, a) will not work. To draw upon the actual world for criteria by which to identify physical properties across possible worlds trivialises physicalism for the actual world: if the properties that are physical in the actual world are the properties that are physical in all possible worlds regardless of whether or not there exists a chain of properties linking them via resemblance relations to the paradigmatically physical properties, and those properties which are not physical in the actual world are not physical in possible worlds in which such a chain of resemblance relations does obtain, then it seems that the family resemblance account only works at the actual world, and only because those properties that are physical at the actual world are physical at the actual world. Daly considers this approach as well, and points out that rigidification with respect to the actual world would fail to identify merely possible physical properties: any property P is a physical property if it is linked by a chain of family resemblances to a paradigmatically physical property. This requires that P exists at the actual world, so any merely possible property in an alternative possible world — such as *being a sphere of uranium five miles in diameter* — is not a physical property, which seems wrong.

Line b) might appear to be more promising, but there are serious difficulties with this line of reasoning as well. First, a description of the characteristics of properties that would be precluded from standing in the resemblance relation must be given, and we would need to see reasons why such characteristics would do what the family resemblance theorist would want them to do. Immediately, it seems that it would be difficult to provide this description without it being too *ad hoc* on the one hand, merely stipulating that properties which have a suspiciously mental tone be excluded, letting anything else be a physical property, or else in danger of disallowing genuine physical

properties identified by a more advanced physics than that which we now have. At a previous state in physics, *being an electromagnetic wave* would not have counted as a physical property under the suggested constraints given above, as a property that would not 'fall out' of explanations made in terms of classical mechanical physics. Along those lines, we would not wish to exclude properties that might be countenanced by a later physics: physicists currently are willing to entertain some very strange properties. We have every reason to believe that in the future they may very well countenance properties that are stranger still.

The family resemblance account could be developed further, but in light of the difficulties Daly has raised for the account and the further ones I have presented, I think it unlikely that the account could be developed to a satisfactory degree. It may work as a 'rough and ready' guide to picking out physical properties, but it cannot provide a complete definition of 'physical' that will suit metaphysical analysis.

2.2. The Physical as a Basic Natural Kind

Paul Snowdon gives an account of what it is to be a physical property according to which 'physical' refers to an extremely basic natural kind with which we are familiar through encounter. He writes:

> We apply the term 'physical' to a range of objects, for example, tables, chairs, and stones which we think we encounter in perception. The term is intended to mark out their most basic and shared essential features. That is, 'physical' is an extremely basic natural kind term, a term for the most all-embracing (natural) kind with which we are acquainted. The discovery (if such is possible) of the essence of the kind is a posteriori. So, on this suggestion, the term's restrictions flow in the same way that those of a natural kind do in other cases.[16]

Thus the predication 'physical' of objects. For a property to count as physical, Snowdon gives the following condition:

> A property is a physical property if it can be instantiated in a domain consisting only of physical objects.[17]

Daly criticises Snowdon's account from several directions. Waiving the objection that the account draws essentially on the Putnam-Kripke theory of natural kind terms such that substantive arguments against that theory would have implications for Snowdon's account of the physical, Daly objects that it is not at all easy to see what could be the 'basic and shared essential features' that all and only physical things have in common. The two features offered by Snowdon — that physical objects should be able to exist independently of perception and occupy space — are intuitive, but are problematic. First of all Daly notes that some paradigmatically nonphysical objects putatively are supposed to be able to exist unperceived: Cartesian souls, Platonic forms, and the like. Secondly, Daly objects to the use of the idea of occupation of space as a criterion for the definition of what it is to be a physical object, for what is meant by

'space' is surely 'physical space'. If so, then to avoid circularity some account needs to be given of 'physical space' that does not draw on physical concepts.[18]

A further difficulty could be raised with the first part of the account given above: that 'we apply the term physical to objects which we believe we encounter in perception and use the term to mark out the most basic, shared essential features of those things'. This is an extremely forgiving account of the attribution of physicality to objects, and without a number of psychological caveats, could possibly admit quite a bit too much into the physical ontology. For example, many people believe that they have encountered in perception such things as ghosts and apparitions. Are the most basic and shared essential features of all things to include features shared by these things as well? If so, then it seems that, unless we can apply some limitations on what we can and cannot genuinely encounter in perception, all manner of reports might contribute to the physical ontology. We might add a constraint to the effect that such observations of entities be repeatable, but consider an argument from a cryptozoologist against a quantum physicist that might run as follows: "You ask me to believe in quarks, but you refuse to believe in ghosts. Show me a quark, and then I will believe". I agree that the physicist probably has a more coherent case than the cryptozoologist for the existence of quarks, but it is not one that he could present without turning the cryptozoologist into a quantum physicist: how else would the cryptozoologist be convinced that smoke trails in particle accelerators and mathematical predictions demonstrate the existence of quarks? And if the quantum physicist were to submit to an equal amount of instruction by the cryptozoologist, perhaps she would find herself, at the end of a three year postdoctorate project, convinced of the existence of ghosts. Stranger things have happened. Thus it appears that, without some limitations, 'things that we believe we encounter in perception' is rather too broad to work as a ground for the formulation as given.[19]

To be fair to Snowdon, he is aware that there are difficulties in setting *a priori* constraints upon the physical, and particularly of the ones he had given: "...it is hard to be confident of even the most general notions, such as occupation of space, such that, as understood, they *must* apply to all physical objects..."[20] The criteria criticised by Daly are offered by Snowdon in response to a possible idealist argument that the most basic and shared essential features of objects, such as tables, chairs, and stones, that we think we encounter in perception might not be a physical essence, but a nonphysical one. Nonetheless, it does seem that, given that there are obvious difficulties in constraining the physical *a priori,* it will not help the physicalist case against idealism to posit *a priori* constraints that the idealist denies.

Daly criticises the account offered by Snowdon that 'a property is a physical property if it can be instantiated in a domain consisting only of physical objects' because it would seem to identify all kinds of properties, such as biological properties, economical properties, and even mental properties, as physical. Consider tables and chairs, some of Snowdon's paradigmatically physical objects. Plainly (unless you are an idealist)they have physical properties, but they probably also have biological properties as well, such as *being made of cellulose,* as well as chemical properties, such as *being viscous* (these are Daly's examples: this author shrinks from the thought of a viscous chair). If these properties can

be instantiated in a domain consisting only of physical objects, then it would follow that they are physical properties. So, too, would be mental properties: human beings are physical objects, and presumably can be instantiated in a domain consisting only of physical objects. Thus on Snowdon's account, mental properties would turn out to be physical properties.

It does indeed seem that Snowdon is committed to this view.[21] But that is a very strong claim, particularly for a minimalist physicalism, the central claims of which it had been Snowdon's initial aim to abstract. Further, it would seem to make all properties physical trivially, as long as they are properties that some physical object can have. Further, as Daly notes, "even if it is true that these properties are physical properties, it should not be a matter of definition that they are."[22] Merely defining properties as physical as we feel inclined will not save physicalism from vacuousness.

It will be interesting to look at how Snowdon constructs his minimalist physicalism so as to understand the context in which the formulations given above arise. The object of Snowdon's paper is to provide a minimal formulation of both materialism and dualism, in order to investigate whether or not there might be viable positions in between. From the outset he is not concerned to construct an exhaustive account: he states that his is but one possible account, whose primary virtue is that it is one which is "intuitively plausible both in the notions it employs and in its conception of materialism, but one which faces problems of clarification and elucidation."[23] For Snowdon physicalism involves a notion of 'weak reductionism', weak because he does not wish to imply by the use of the term anything to which a minimalist physicalist would not agree. With that caveat he writes that "Physicalism is the thesis that we can reduce mental facts (states of affairs) to physical facts (states of affairs); dualism is the denial of this thesis."[24] Weak reductionism involves the notion of the 'exhaustive constitution' of one state of affairs by an arrangement of other states of affairs. Thus mental states will be weakly reducible to physical states if they are exhaustively constituted by physical states.

With this understood, it is easy to see why Snowdon has no difficulty with the formulation of a physical property being one that can be instantiated in a domain consisting only of physical objects: a domain consisting only of physical objects may be arranged such that all manner of things might be 'exhaustively constituted' by that arrangement, but in every case, whatever is so constituted will be reducible to physical states of affairs. He is aware, of course, that the constitution thesis is empty if there is not some definition of what is to count as a physical property and what non-physical, specifically, mental. He offers the following idea for elucidating 'physical':

> One possible response is to explain what the import of 'physical' is in a way which is like that standardly employed for the term 'mental'. That way is to cite samples or paradigm cases while saying that mental states are states of this kind. It is allowed that this is enough to explain (or convey) a notion, which is then used with a fair degree of intersubjective agreement. So, we can try to explain by a 'physical fact' we mean facts which are like these — and then we would list a characteristic group... it provides no discursive account of what 'physical' means, and it provides no identification of the essence of the physical. But if it succeeds in generating a recognitional capacity, then it makes physicalism contentful and non-vacuous.[25]

Snowdon thinks it unlikely that we should be able to give an *a priori* account of what counts as a physical property: "the reason is that it is hard to be confident of even the most general notions, such as occupation of space, that, as understood, they *must* apply to all physical objects, given the surprising entities physicists are prepared to endorse",[26] and further, that even if we could provide some criteria, it is doubtful that they would provide sufficient conditions for a property's counting as a physical property. What is needed, he says, is "an elucidation of 'physical' which allows its elucidation to be sensitive to the empirical exploration of the physical world".[27] He rejects the idea that either current physical science or some unspecified future physics should be the measure of what is physical, because it is impossible to say what properties an unspecified future physics might admit, as will be familiar from above. It is in this context that he offers the formulation criticised by Daly.

I understand Snowdon's formulation in the following way: when he says 'We apply the term 'physical' to a range of objects... which we think we encounter in perception', Snowdon means to capture the basic, intuitive sense in which we know how to identify a physical thing, and apply the term to those things that we can see, touch, smell, poke &c., and not to things that we cannot encounter in perception, such as other people's mental states, which must be derived from things that we *can* encounter in perception, such as their speaking, acting, and so forth. Whatever our best science tells us about the basic nature of these things, no matter how surprising it may be, it is these things which we will consider to be physical, and should we ever be able to specify the essence of such things — which, presumably, a true and complete physics would be able to do — then we will have a full definition of what the essence of the physical is. Snowdon's definition is intended to perform three functions: to set up a principled distinction between the mental and the physical; to specify something that generates a recognitional capacity for what is to count as 'physical'; and to leave whatever is the essence of the physical sensitive to and (possibly) discoverable through *a posteriori,* empirical investigation of the world.

With the context of Snowdon's formulation understood, and especially with the understanding that he is not offering a defence of any particular version of physicalism, but only attempting to lay out and clarify its central claims, I think it is easier to see the places in which the phrases extracted by Daly for criticism fit, and why Snowdon makes the somewhat surprising claims he does about any property that can be instantiated in a physical domain *really being* a physical property, about which Daly had said " Snowdon provides no independent support for this apparently mistaken consequence of his view".[28] On Snowdon's account, if any property is instantiated in a domain consisting *only* of physical objects, then it is *exhaustively constituted by* the obtaining of purely physical states of affairs. This notion of exhaustive constitution can easily give way to identity, so long as it is individual instantiations of exhaustively constituted properties that are identified with the physical things that constitute them, rather than types of exhaustively constituted properties, to be sensitive to the multiple realisability of some types of properties. Thus Snowdon might say, as he does, "The obtaining of our mental states (our being the way we are mentally speaking) is exhaustively constituted by the obtaining

of purely physical states of affairs (by, that is, our being the way we are mentally speaking),"[29] and interpret that to mean that our being the way we are mentally *is* our being the way we are physically, and that mental properties *are* physical properties in this restricted sense. So it appears that Daly's criticisms are slightly unfair: Snowdon does not mean to define physical properties, for once and all, in the way that Daly suggests that he does, and neither is the problematic consequence that mental states are physical states entirely unsupported. However, while the notion of exhaustive constitution certainly can be construed this way, it does not seem that it must be so construed. It is not clear that the notion of exhaustive constitution cannot accommodate a notion of realisation as above outlined according to which some realised properties are genuinely nonphysical. Thus, without further development of the notion of exhaustive constitution, the account is susceptible to the objections raised by Daly.[30] There is also the matter of the actual identification of those properties to which we apply the term 'physical' — Snowdon's account, it has been shown, is in danger of admitting far too much, and identifying highly suspect properties as physical. Snowdon's account thus also fails to pass the test for providing a metaphysically satisfactory definition of the physical.

2.3. Physical Properties and Paradigm Physical Effects

Daly's next target is an account of the physical given by David Papineau in the first chapter of his Philosophical Naturalism,[31] in which he describes his interpretation of the physicalist thesis and attempts a defence of physicalism generally. I would like to take a slightly different approach to the analysis of Daly's criticism of this account. It will be instructive first to see exactly how Papineau's physicalism is constructed.

Physicalism for Papineau consists in the conjunction of two theses: *supervenience* and *token congruence*. The supervenience relation Papineau favours is local, strong supervenience, according to which two systems cannot differ in any intrinsic higher level respect without differing in some physical respects, and if two systems are physically identical, they will also be identical in all of their intrinsic higher level respects.[32] The *token congruence* of mental states with physical states is a statement of the identity of each dated mental occurrence with some dated physical occurrence. Papineau is not committed to the strict identity of higher-level properties and events with complexes of physical properties. Respecting the multiple instantiability of mental states, token congruence does not require that types of mental occurrences correlate with types of physical occurrences, though they certainly could. Papineau thus prefers to cash out the notion of token congruence in terms of the realisation of higher level properties by physical properties.[33] In those cases where type identity between higher and lower level properties fails, because the higher level properties are multiply realised, "special categories cannot even in principle be specified in physical terms. Nevertheless, physicalists will say, such special terminology is still just a way of describing complexes of physical stuff, and does not require us to recognise any non-physical substances."[34] The conjunction of these two theses is sufficient for physicalism, constraining all properties

and relations to physical properties and relations (thus excluding epiphenomena) and providing an account of the relation between physical and nonphysical properties which, while respecting the idea that special terms have real referents, does not have difficulties with the issue of overdetermination.[35]

Both the theses of supervenience and token congruence are independently motivated by what Papineau calls 'an important feature of physical science', the principle of the *causal closure of physics*. He writes:

> I take it that physics, unlike other special sciences, is *complete*, in the sense that all physical events are determined, or have their chances determined, by prior *physical* events according to *physical* laws. In other words, we never need to look beyond the realm of the physical in order to identify a set of antecedents which fixes the chances of any subsequent physical occurrence.[36]

Thus the causal closure of physics amounts to the proposition that physics is *complete*, or able to give complete explanations for physical phenomena. This principle, in conjunction with an intuitively plausible principle Papineau terms the *manifestability of the mental*, supports both the principles of psychophysical supervenience and token congruence.

The principle of the manifestability of the mental states that for each mental particular, there are possible contexts in which the possession of that mental particular by an agent would make a causal difference, and thus would be manifest: "...if two systems are mentally different, then there must be some physical contexts in which this difference would display itself in differential physical consequences, or chances thereof".[37] He does not argue for this principle, apparently taking its truth as obvious, but some explanation of the principle is appropriate. As I understand it, the key idea is not that each mental state *is in fact* causally effective, but that each mental state *has potential causal powers* such that, were it instantiated in an agent embedded in the right circumstances, it *would be* causally effective. No mental state (say, the belief that the kind of aliens who tend to abduct people from Earth speak Quebecois) can have zero causal potential, even if *in fact*, in the entire history of all sentient life anywhere in the actual world, it never features in any causal chain (no aliens ever abduct anyone from Earth, and no one considers the possibility that they might): had things been otherwise, the mental state would have featured in some causal chain with a behavioural outcome (some aliens abduct me, and believing that they speak Quebecois, I greet them in French, which they will understand). In other words: For each possible mental state, there is at least one possible world in which it features in some causal interaction.

This principle is similar to a principle expressive of realism, specifically, anti-epiphenomenalism about the mental which Kim has termed 'Alexander's Dictum'. It is the principle that "*to be real is to have causal powers,*"[38] or, if something has no causal powers, it may as well not exist at all. Kim quotes Alexander:

> [Epiphenomenalism] supposes something to exist in nature which has nothing to do, no purpose to serve, a species of *noblesse* which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time be abolished.[39]

Alexander's dictum is stronger than the principle of the manifestability of the mental, for differential epiphenomenal mental states could show themselves — or more appropriately perhaps, be shown — in differential physical consequences, in virtue of being epiphenomena of differential physical states, whereas according to Alexander's dictum, if that is the only way they are shown, they may as well be assumed not to exist. However the principle of the manifestability of the mental is forwarded by Papineau in the context of the conjunction of the principles of supervenience and token congruence, the latter of which is designed to rule out the possibility of epiphenomenalism. So, in essence, the principle of the manifestability of the mental as it functions in Papineau's account and Alexander's Dictum amount to the same thing: each real mental state has real causal powers, even if those powers are only potential.

Perhaps this principle would be easier to swallow if it were translated to the physical case, as is easily done, for it generalises over physical things as well. The notion of a physical thing which cannot interact with any other physical thing, ever, in any way, in any possible world under any circumstances, makes very little sense: what reason could there be to postulate such a thing? What purpose could it serve in any physical theory, or in any epistemological system? What reason could we give even for calling such a thing *physical,* as opposed to anything else? If it has no causal powers, then it can never be detected, never have any effect upon the world at all, never make a whit of difference. By contrast, consider a possible world in which the laws governing electromagnetism are exactly as they are in the actual world, but wherein there exists nothing but a single, solitary electron. In the entire history of this universe, no causal transactions take place,[40] for there is nothing with which the electron could interact. But if there had been a proton in suitable proximity to the electron, the two particles would have interacted, because it is in the nature of electrons and protons to do so. Their potential for causal interaction is a fact about their nature, even if in fact they never interact with anything.

Papineau's arguments for supervenience and for token congruence run as follows:

•According to the principle of the causal closure of physics, the effects of any physical event are determined by prior physical events: once the physical antecedents are given in any casual interaction, the outcome (or chances thereof) is fixed.

•According to the principle of the manifestability of the mental, any mental differences must be able to show themselves in differential physical effects.
So:
•It follows that mental differences without physical differences are impossible.

The argument in support of token congruence is similar:

•Every mental event is a potential cause of some physical event (the manifestability of the mental).

•All physical events have complete physical causes (the causal closure of physics).

So:

•It follows that all mental events are physical events.

Both of these arguments rest on the principles that mental states are or potentially are causes of physical effects, and that physical effects, or their chances, are fixed once all physical antecedents of those effects are fixed. In order to be causally effective, mental states must indeed *be* physical states: they must be token congruent with physical states.[42] This is not to say that they are strictly to be identified with mental states, for, as Papineau is aware, if any mental states are multiply instantiable, then they cannot strictly be identified with the types of physical states with which they are token congruent. Again, Papineau prefers to cash out the relations of supervenience and token congruence in terms of the realisation of one set of states (properties) by another set of states (properties), preserving a genuine difference between mental properties and the physical properties realising them whilst providing an explanation of how the physical effects of mental causes are not overdetermined:[43] "even though the two properties are not the same property, the instantiation of one is realised by the instantiation of the other."[43]

The next step in Papineau's argument is to provide a defence of the principle of the causal closure of physics, upon which his arguments for supervenience and token congruence depend. Papineau identifies the most obvious difficulty facing a defender of the principle to be to say what is meant by 'physics'. Current physics cannot be the physics involved, for current physics is surely incomplete and not suited to the identification of *all* the antecedents of some physical events. Neither can some future or ideal physics be meant, for we cannot know at this juncture in the history of physics what kinds of things physics will in the future countenance. However, Papineau makes use of the notion of an ideal physics, possibly very different from current physics, but stipulated to be complete and correct: "Suppose we simply *define* 'physics' as the science of whatever categories are needed to give full explanations for all physical effects".[44] Now we need to know what these 'physical effects' for which the ideal physics is supposed to provide explanations are. It is here that the passage to which Daly objects occurs. Papineau writes:

> I propose that we simply postulate some pre-theoretically given class of of paradigmatic physical effects, such as stones falling, the matter in our arms moving, and so on. If we take this class to be independently given, then we can effectively characterise the rest of physics as all the categories that need to be brought in to explain those paradigmatic physical effects.[45]

Those properties identified by physics will be physical properties. Thus, any category that needs to be brought in to explain the stipulated class of paradigmatically physical effects will be a physical property.[46]

Daly finds three major difficulties with the notion of deriving the import of 'physical' from a pre-theoretically given class of paradigm physical effects. First, the account may fail to identify merely possible physical properties, as these will not be needed to provide explanations of any paradigmatically physical effects. Daly considers two methods by which Papineau might overcome this difficulty: he might suggest that a possible nonphysical property P is physical just in case there is some possible world at which P is needed to explain some paradigmatically physical effects. This will not work, however, for there are possible worlds at which dualism is true, and in these worlds, nonphysical properties would need to be cited to explain some paradigmatically physical effects. To identify merely possible physical properties with reference to possible worlds in which they would be needed to explain paradigmatically physical effects would thus mistakenly identify nonphysical properties as physical in such worlds.[47]

Daly considers that Papineau could present an account of what counts as a determinable physical property, from which all truly physical properties would emanate as determinates. However, providing such an account would be extremely problematic. Indeed, it seems to me that it would be impossible to form a notion of a physical determinable without a prior grasp of the determinates that were to emanate from it. What sorts of properties could count as physical determinables? Perhaps properties such as *being a force,* from which bein gravitational, electromagnetic, strong, weak forces, &c., would emanate; perhaps *being a particle,* from which being electrons, muons, photons, tachyons, &c. would; and so forth. But what else might be specified by future physics? It is certainly possible that a future physics might postulate properties that do not fit into any determinable category now employed, thus implying a new determinable property, and so we cannot postulate a determinable that will cover them all.[48] In advance of knowledge as to what properties future physics will accommodate, we cannot constrain the determinables from which such properties will emanate. Such an account as Daly postulates might be able to accommodate possible physical properties such as could reasonably be projected from properties that are currently recognised as physical — *being a sphere of uranium five miles in diameter,* for example, would be included, but the account might exclude *actual* physical properties not recognised by current physics, but which physics could recognise at a later stage in its development, as well as possible physical properties that might not be projectable from any of the actual physical properties in this world.[49]

Daly's second objection is that Papineau's account is too narrow with respect to physical properties at the actual world, for some physical properties can be brought in to explain paradigmatic physical effects, but are not necessary to explain those effects. Daly notes that since *density* can explain certain kinds of paradigmatic physical effects which could just as easily be explained by reference to *mass* and *volume,* it follows that *density* is not a physical property. But surely Papineau could reply that since *density* just is the product of *mass* and *volume,* and explanations of a paradigm physical effect made in terms of one or the other are effectively the same explanations couched in different terms. From such an account a form of reductionism might be seen to follow: a property is a physical property just in case it is part of the set of categories that need to be brought in

to explain paradigmatically physical effects or is identical with or reducible to a member of that set.

This does indeed seem to be the spirit in which Papineau's account is offered, insofar as he intends it to provide an account of what is to count as a physical property. A class of pre-theoretically given paradigm physical effects is given, the explanations for which will be provided by a stipulatively complete ideal physics. This physics is "the science of whatever categories are needed to give full explanations for all physical effects."[50] Presumably, whatever this physics has to say about the microstructure of the paradigmatically physical effects and their causes will tell us what are the basic physical categories, and so in what 'being physical' consists.

This brings us to the third difficulty Daly identifies for Papineau's account. Daly writes: "...we know which effects to include in the selection of the requisite physical effects only to the extent that we know which properties are physical, and so what properties need to be brought in to explain those effects. But without a definition of what a physical property is, it is unclear which effects should be included in the selection."[51] Daly is quite right to call attention to this weakness. The notion of a paradigmatically physical effect must draw upon the notion of physical properties, for, as Daly says, "a paradigmatic physical effect is — paradigmatically — a change with respect to a paradigmatic physical property".[52] Thus we must draw upon a pre-theoretic notion of physical properties in order to be able to say what properties physics is to cover. Any attempt to define which properties are to count as physical properties by this method would be circular.

I believe Papineau would reject the charge of circularity on the grounds that the set of paradigmatic physical effects is taken as independently given such that what is to count as 'physics' is derived from this set. Whether or not this is a viable strategy remains to be seen. I am not certain, however, that in this section of his book Papineau is concerned with the definition of what is to count as a physical property. His intention is rather to provide a defence of physicalist intuitions and physicalism generally, according to which *all* properties either are physical properties, or else they are multiply instantiable, functional properties not strictly to be identified with any particular types of physical states, yet identical to instantiations of physical states with which they are token congruent. If he is right, we will not require a definition of what is to count as a physical property. Physics, stipulated to be complete, will in principle be adequate to the task of explaining all physical effects in physical terms, and any category physics employs will be a physical category. Thus he focuses his attention on the defence of the more problematic of the two main assumptions upon which his arguments depend: the principle that physics is closed or complete.

There are serious difficulties with this defence, however. We have seen that Papineau solves the difficulty involved with clarifying what is to count as a 'physical effect' for the purpose of identifying those effects for which stipulatively complete physics is supposed to account by postulating an independently given set of paradigmatically physical effects. But this threatens to trivialise physicalism. Recall that Papineau gives very loose criteria for the identification of the independently given class of physical

effect: "stones falling, the matter in our arms moving, and so on."[53] What are we to make of this 'and so on'? I take it that Papineau means to include at least all objectively observable macroscopic changes of states of affairs in the set. This involves mental causes as well: 'the matter in our arms moving' is presumably sometimes caused by our desires to move our arms. By definition, physics will be the science that provides explanations for all paradigmatically physical effects, including those for which mental causes can be cited. Thus mental causes, even if they are shown to be irreducible to purely physical states of affairs (as some people certainly believe that they are) will simply be included in the domain of physics. This simply makes it a matter of definition that physics is complete, and that physicalism is true in this world. It will also mistakenly identify nonphysical properties as physical properties in possible worlds at which dualism is true, as Daly had argued, noted above.

Papineau is aware that this definitional strategy is in danger of making the principle of the completeness of physics an 'empty analyticity', and of trivialising the physicalism/dualism debate by allowing that mental states are simply included in the physical.[54] His answer to this difficulty is somewhat surprising. He writes:

> I still need to explain how substantial conclusions about supervenience or token congruence could possibly follow from such a definition.
>
> My answer is that no substantial conclusions follow from the completeness of physics *per se*. But they do follow from the joint assumption that a) physics is complete and b) that it does not make use of psychological categories.[55]

This is just not satisfying. An adequate defence of the principle that physics is complete and closed cannot be based on an argument that assumes it along with the added premise that physics makes no use of psychological terms to the conclusion that substantial conclusions about nontrivial psychophysical supervenience and token congruence follow. Also, this strategy fails to characterise 'physics' in any meaningful sense. The mental may be a category whose realisation in the physical is difficult to explain, but it is by no means the only category of realised properties: consider such properties as *being a unit of money, being a river, being an atmosphere, being a hive,* or any number of similar properties. Each of these is multiply instantiable: a unit of money can be realised by printed bits of paper, specially shaped pieces of metal, an electronic process, even a promise. A river presumably could be realised by any liquid flowing through some terrain. An atmosphere is a collection of gases held to the surface of a celestial body by gravitational force: any gases will do, and any kind of celestial body, provided it is such that gases can cling to it. A hive is a collection of individuals living and working together in a particular way. This arrangement is typically realised by insects, but could just as easily be realised by spiders, birds, the Borg, or what have you. All of these kinds of properties will be picked out by a different special science: economics, geology, meteorology, xenosociobiology, and so forth.

Merely stipulating psychological categories as unemployed by physics leaves it open that categories employed by economics, geology, meteorology, xenosociobiology, and any other special science *can* be included in physics. Thus 'physics' becomes

'whatever isn't psychology', which cannot be right. Papineau's only defence of this *ad hoc* exclusion of psychological categories from physics is that, looking back on the history of science and considering the aims of physics, it is extremely unlikely that psychological categories will need to be brought in to explain anything. That may be true, but the notion grounding this unlikeliness is the idea that the psychological, like other special science categories, is realised in the physical, and thus that purely physical explanations will suffice for explanations of such phenomena. However, this is not an assumption of Papineau's argument, but rather a conclusion for which he intends to argue. This is not only circular, but it is simply to ignore arguments to the effect that the physical ontology is incomplete, some of which are rather compelling, and which are certainly deserving of far more serious attention.

In fine, Papineau presents a highly flawed formulation of physicalism, and his general defence of physicalism proves inadequate. Daly is generous to take it that this formulation provides any criteria by which physical properties could be distinguished from nonphysical properties, for it does no such thing. It fails even to provide a principled distinction between physics and special sciences. Much of the difficulty could, I believe, have been avoided had Papineau formulated the principle that physics is complete or closed as the metaphysical principle perhaps more standardly employed, the principle of the *causal closure of the physical domain,* for which principle there are good independent arguments: the difficulty involved in explaining how a nonphysical event could cause a physical event, with us since Descartes, paramount among them. The only reason I can consider that Papineau formulates his version of the principle as he does is that he wishes to keep physicalism an epistemological rather than a metaphysical doctrine, and thus to define physicalism in terms of the science of physics rather than in terms of the ontological priority of physical properties as related to nonphysical properties. I can think of no reason why such a formulation would not be possible, and many of the difficulties involved with Papineau's account could be addressed from within such a formulation. But any such formulation will have to provide a more principled distinction between the physical and the nonphysical or not strictly physical than 'all the categories needed to explain paradigmatically physical effects', where the set of 'paradigmatically physical effects' is left so vague.

Physical Properties as Properties Studied by the Physical Sciences

The final approach criticised by Daly is the identification of physical properties as those properties that feature in the physical sciences. Accompanying this approach are the notions that "what makes a science a physical science is just that it (more or less) shares its methods and principles with paradigm physical sciences"[56] and that "a physical science, such as mechanics, is a paradigm physical science simply because of its pioneering role: it is among those relatively few long-standing sciences which were originally called physical sciences".[57] Daly focuses upon the work of Geoffrey Hellman and Jeffrey Poland as advocates of this approach.

The attempt to identify physical properties as those that feature in the physical sciences faces difficulties similar to those already encountered in the Family Resemblance account: no criteria are immediately presented according to which we could say which sciences are appropriately related to whatever sciences we identify as the paradigmatically physical sciences such that they themselves count as physical sciences. Further, sciences could presumably be related to paradigmatically physical sciences via chains of sciences — say, for example, chemistry is taken to be appropriately related to basic physics via subatomic theory. Chemical properties would then count as physical. But in an alternate possible world without subatomic theory, nothing would carry the relation to more basic physics, and chemical properties would not count as physical properties.[58] Likewise, physicalists in an alternate possible world in which a putatively false science could stand in such a relation to a paradigmatic physical science (or a science unproblematically linked to basic physics) and to some paradigmatically nonphysical science, such as psychology or economics, such that its properties were mistakenly classified as physical. Thus in the absence of criteria constraining the relations of nonbasic sciences to paradigmatically physical sciences, defining physical properties as properties studied by a physical science could mistakenly identify nonphysical properties as physical, or fail to admit genuinely physical properties into the recognised physical ontology. Also, it appears that according to this system which properties would count as physical properties would in part be a matter of historical accident, dependent upon which sciences actually had been developed to carry relations through to basic physics.

The second obvious difficulty with this approach is the familiar problem of isolating what is to count as 'physics' for purposes of identifying what is to count as a physical property. Current physics is beyond reasonable doubt incomplete in some respects and incorrect in others, so is not suitable to provide an exhaustive definition of physical properties. But we cannot appeal to an unspecified but putatively complete and correct future physics, because we do not know what properties such a science will admit.

Geoffrey Hellman and Frank Thompson, in their paper *Physicalism: Ontology, Determination, and Reduction,*[59] developed an account of physicalism drawing upon current physics for its elucidation, while admitting that current physics is almost assuredly incomplete and that it very probably makes claims about the world that are false outright. This account is defended by Hellman in the paper criticised by Daly, *Determination and Logical Truth.*[60] Acknowleding that physics may be incomplete and incorrect in some of its claims, Hellman says:

> Physicalist principles (say, ontological exhaustion and the determination principle) still serve two functions: first, they may be useful in articulating conditions of adequacy that science aims to meet — that is, satisfying principles of the same logical form may serve as a useful criteria of completeness. Second, literally false principles may be highly instructive, and they may even be very nearly true (or even exactly true) in a restricted form, that is, when applied to a specific subdomain of interest, such as living or sentient beings on this planet (+satellites). Thus for example mental phenomena, if they are physically determined, are almost certainly determined by physical events at the level of neurons and synapses, and, on dimensional grounds, we expect classical (i.e. nonquantum) mechanics to be able to describe this level accurately enough to fix the

> mental (i.e., 'pin it down', not 'cure it'). Thus, a working assumption would be that, even if there are, e.g., yet to be discovered elementary particles swimming around in our brains, they hold no secrets for psychology.[61]

This claim is engineered to convey content upon the thesis of physicalism without having to be concerned with the incompleteness and incorrectness of current physics: though physics may be in a state of development, and may in the future course of its development make further claims about the intrinsic nature of such things as neurons and synapses that falsify claims about the nature of these and other things as made by current physics, that will change little about how we actually deal with neurons and synapses in neurophysiology. Insofar as neurons and synapses qualify as physical, and insofar as we describe connections between our neurophysiology and relations to our external environment and our mental lives, then we will have a meaningful physicalist thesis about the mind. Meanwhile, physics is free to develop as it will. Hellman effectively separates 'physics' from 'physicalism' — *physicalism* is for him the thesis that all macroscopic phenomena, such as mental phenomena, are comprised and determined by and explicable in terms of phenomena that can be picked out by current physics: that is, there are no (macrophysical) phenomena that do not have their bases in something describable in terms of current physics. Physics as it proceeds may identify lower-level phenomena that constitute and explain the phenomena picked out by current physics, and future physics may prove that some or all of the claims made by current physics are in fact false. What *physics* does will not falsify Hellman's restricted version of *physicalism* as long as it remains true that macroscopic physics is sufficient to provide explanations for all macroscopic phenomena and descriptions of all macroscopic entities.

Daly finds this version of physicalism inadequate. First, in the absence of a definition of what is to count as a physical property, it is difficult to make sense of the claims that all natural phenomena depend on physical phenomena and all entities are exhausted by physical entities. Secondly, Daly finds that Hellman's formulation of physicalism falls short of providing a general, unrestricted version of physicalism, as it explicitly restricts the claims of physicalism to the macroscopic. This is correct. But it seems that we only need such a general, unrestricted version of physicalism if we *require* a principled account of what is to count as a physical property. But Hellman explicitly rejects this requirement: "Hellman-Thompson physicalism appealed to existing physics in formulating ontological exhaustion ("basic, positive, physical predicate"), and it made no sense of "physical property" apart from "what is expressed [in models of relevant laws] by a predicate of physical theory."[62] Hellman accepts the fact that the current physics upon which Hellman-Thompson physicalism draws is incomplete, but they do not require a complete physics for their elucidation of the physicalist thesis. Thus, if Hellman's and Thompson's physicalism fails to provide a satisfactory account of what is to count as a physical property, that is hardly surprising: they do not intend for it to do so. What will count as a physical property, as far as restricted, macroscopic physicalism is concerned, will be those properties that are picked out by macroscopic physics. They will also be those properties picked out by quantum mechanics, by superstring theory, by whatever theories are developed in physics. The explanation of macroscopic physical

phenomena given by these branches of knowledge will be interesting to physicists, but irrelevant to restricted physicalism.

There remains the question of whether such a restricted version of physicalism can truly be satisfying. Many physicalists would prefer a stronger connection between physicalism and physics as it develops than is expressly involved in Hellman-Thompson physicalism. Further, Daly points out that Hellman-Thompson restricted physicalism is entailed by *weak physicalism* as formulated by David Armstrong, and that weak physicalism is compatible with certain notions that physicalists characteristically are concerned to reject: the idea that there exists a transcendent God is his example.[63] If indeed a stronger version of physicalism is motivated, it appears that Hellman-Thompson restricted physicalism will not address all physicalist concerns.

Jeffrey Poland describes a more general version of physicalism according to which the physicalist ontology is to be derived from the ideology of physics. He describes physicalism as "a general programme of non-eliminative, structural unification that accords a certain sort of privilege to physics."[64] The 'structural unification' involved is unification of branches of knowledge: according to physicalism as Poland describes it, all sciences have a place within a structure with physics at the foundation. He is careful to distinguish between the notions of *unitary* science and *unified* science:[65] according to the former notion, one science will be sufficient for the description and explanation of all real phenomena. This is an eliminative reductive notion, entailing that all nonbasic sciences, while they may be humanly interesting, are explanatorily superfluous. The notion of *unified* science, on the other hand, does not entail eliminativism, but is compatible with the idea that there can be real divisions between different branches of knowledge and the kinds of phenomena with which they deal, such that physics is entirely inappropriate for the explanation of some genuine phenomena that are within the domains of other sciences. However, all sciences are in some way related to physics, such that physics is privileged over all other sciences. This privilege is characterised by claims in three areas: ontology, objectivity, and explanation.

The ontological privilege of physics consists in the notion that "everything has a place in an ontological structure grounded in the ontology of physics: the physical ontology is the most basic and comprehensive relative to all that there is in the sense that all objects and attributes are dependent upon, supervenient upon, and realised by physical objects and attributes".[66] Thus there are no objects, attributes or relations independent of physical objects, attributes and relations. This does not entail the strong claim that there exists nothing apart from the purely physical, but involves only the notion that if there are genuinely nonphysical things, they are ontologically grounded in and dependent upon physical things.

The objective privilege of physics is similar to the notion that 'the physical facts determine all the facts': "The idea is that for there to be objective matters of fact and truth in some domain and for there to be objective sameness and difference in some respect between individuals, there must be certain appropriately related *physical* facts, truths, similarities and differences."[67] There is also the related epistemological notion that the physical facts are as they are independently of how any human endeavour (or any

other intellectual endeavour) describes them: after all, if physicalism is indeed true, all human endeavor is realised in the physical. Physical facts and truths are thus prior to all other facts and truths: "they are the putative objects of our knowledge, but they are in no way dependent upon that knowledge."[68]

The explanatory privilege of physics, perhaps the most interesting feature of Poland's physicalism, involves the idea of a layered hierarchy of sciences with physics as the most fundamental branch of knowledge. Descriptions of phenomena picked out by any particular science, such as biology or chemistry, can be explained and/or described in terms of phenomena picked out by sciences at a (characteristically) lower level, for example, molecular physics or chemistry, ultimately by physics. "The key idea is that of a unified explanatory system in which different branches of knowledge are organised hierarchically with physics at the foundation, and in which the generalisations and phenomena studied at each level of the hierarchy are explainable in terms of generalisations and phenomena at lower levels."[69] Poland's term for this relation is 'vertical explanation'.[70] The idea does not entail that for all phenomena at all levels, there exists an explanation *in physics,* but only that the explanation that does obtain for any given phenomenon on any particular level is related, either directly or by any number of intermediate explanatory links, to physics.[71] Physical explanations ultimately serve to demystify higher level phenomena by providing a set of basic processes and mechanisms that ground all processes on higher levels.

Poland's notion of what *physicalism* is, as well as what *the physical* is, draws upon the notion of what *physics* is, such that the physical ontology and the doctrine of physicalism are to be derived from the ideology of physics. In order to give content to his formulation, he must give a definition of physics. It is to this strategy that Daly objects. Daly says "Jeffrey Poland attempts to say which properties are physical by (1) saying what physics is, and then (2) saying what properties are physical properties in terms of saying what properties are described by physics." But this is not quite right. Poland writes:

> The central idea for an *a posteriori* approach [to formulating physicalism] is to pin down *physics* as a distinct branch of knowledge and to develop the physical bases in terms of it... the approach I favour begins with a characterisation of physics that can briefly be summarised as follows: physics is the branch of science concerned with identifying a basic class of objects and attributes and a class of principles that are sufficient for an account of space-time and of the composition, dynamics, and interactions of space-time.[72]

Let us see how Daly's criticisms do not address Poland's actual position. Daly offers several criticisms of Poland's account. Firstly he asks: in what sense of 'basic' does physics describe 'basic' classes of entities, and what kinds of dynamics and interactions does it describe? He mentions Poland's assertion that psychologically based systems, unlike physics, do not introduce properties and principles to explain the dynamics of the occupants of space-time, and comments that as human beings are unquestionably occupants of space-time and since psychology introduces properties and principles to explain the dynamics of human interactions, Poland's claim seems wrong, and physics cannot uniquely be characterised in this way. Second he asks us to consider a merchant bank, composed of a chair, shareholders, a board of deputies &c. It interacts with other

financial institutions, and engages in the dynamics of buying and selling stocks. Yet physics, presumably, is not concerned with the dynamics and interactions of this particular type of space-time occupant. The question presents itself: with what class of dynamics and interactions is physics concerned, then?

I think that these questions are answered in the passage directly following the one quoted above. Poland writes:

> The crucial feature of these classes is that they are minimal with respect to the descriptive and explanatory purposes they serve, that the magnitudes are defined for all regions of space-time, and that each occupant of space-time satisfies the principles governing these magnitudes. It is both the types of phenomena they are introduced to explain (i.e. composition, dynamics, interactions) and their complete generality that distinguishes these magnitudes and principles from others, and hence that distinguishes physics from all other branches of inquiry.[73]

Thus, though human beings are occupants of space-time and though psychology does introduce properties and principles to explain their dynamics and interactions, psychology does not introduce *minimal* properties and principles, nor does it introduce *general* ones, such as would apply to those regions of space-time not containing human beings. We never find objects or processes and dynamics such as would appropriately be explained by psychology in the absence of human beings.[74] However, wherever we find human beings, we also find masses, electrical charges, gravitational forces, &c — magnitudes shared by other regions of space-time containing no intentional agents, and which are appropriately described and explained by the properties and principles introduced by physics. The properties and principles introduced by psychology presumably will be vertically describable and explainable in terms of the science of that which is shared by all regions of space-time, and that science will be physics. The same is true for the example of the merchant bank: we never find a merchant bank or its various constituents and dynamics in the absence of objects and dynamics appropriately described by physics.

Thus physics, being concerned to identify the most basic class of objects and principles sufficient for the explanation of the dynamics and interactions of *all* occupants of space-time in *all* regions of space-time, is to be considered a universal science[75] — not, as noted above, that all phenomena in space-time have descriptions and explanations *in physics,* but that for any occupant of any region of space-time and for all interactions and dynamics involved with that occupant, there are space-time occupants comprising or constitutive of that occupant such that explanations in terms of physics are appropriate to the constitutive objects, and such explanations are vertically related to the explanations appropriate to the explanation of the interactions and dynamics of the original space-time occupant.[76] The descriptions and explanations given for those phenomena will be, if not enlightening insofar as they are applied to the higher-level space-time occupant, sufficient for the explanation of all the dynamics and interactions in the system.

For Poland, the universality of physics is not to be construed in terms of its universal explanatory power, for it does not have the power to explain all kinds of phenomena that there are or could be. Its universality consists in the kinds of questions it is characteristically directed towards answering. Poland gives a partial list of such

questions, very broadly construed:

- What are the fundamental constituents of *all* occupants of space-time?
- What are the fundamental processes that underlie *all* causation and interaction between such occupants?
- What parameters are relevant to the dynamic unfolding of *all* systems in space-time and hence to *all* change?
- What is the nature of space-time itself, its origin (if it has one), and its destiny?[77]

The answers to these highly general questions will of necessity themselves be highly general, and it is this feature that Poland finds characteristic of physics: "it is these features which establish the ontological and epistemological *relevance* of physics to the physicalist programme".[78]

In objection, Daly points out that there are branches of physics that are not concerned with questions such as those Poland lists as characteristic of physics, citing astronomy and thermodynamics. He is certainly right: such sciences as these are typically not concerned with the ultimate structure and origin of all things. However, I think it should be clear from the preceding discussion that the physics that Poland has in mind as sitting at the foundation of the explanatory hierarchy is *basic* physics. Astronomy and thermodynamics are not basic physics, so practitioners of those sciences are free to be unconcerned with the basic, fundamental features of space-time and its occupants.[79]

Surely, however, this approach falls to the difficulties Daly had earlier cited, and that Hellman had accepted in his formulation of physicalism: our current most basic physics is almost assuredly incomplete and incorrect, and we cannot simply postulate a complete and correct unspecified future physics as the most basic and universal physics, for to do so would render the doctrine of physicalism as stated vacuous. To what physics are we to appeal?

Poland is, however, perfectly aware of these difficulties. He himself raises them in objection to the idea that the physical ontology can simply be read off from the activity and production of physics — the very position he is described by Daly as occupying,[80] which position Poland describes as 'quite unacceptable'.[81] Poland objects to this position from three directions: first, 'leaving it to the physicists' to determine what is to count as physical is "unhelpful unless it is known who the physicists are and, more broadly, what physics is".[82] Without an account to answer this concern, the statement of what is to count as physical is empty or circular. Secondly, this position leaves it open that physics, and consequently what counts as physical, could be determined by "arbitrary administrative decisions or other forms of socio-historic accident"[83] in ways that would not answer to the metaphysical concerns of physicalists. Thirdly, it is open whether it is current physics, future physics, or ideal physics that the physicists to whom we are leaving it to say in what the physical ontology consists are supposed to be practising. The difficulties with all three will be familiar from the preceding discussion.

Is Poland then contradicting himself in saying that the ontology of physics is to be derived from a characterisation of physics? Not at all: the position ascribed to Poland by Daly is not in fact one that he occupies, but rather, like Hellman, Poland explicitly

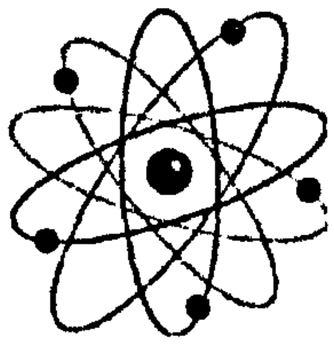
rejects the notion that any constraints at all can be set upon the physical that are not subject to revision. It is inappropriate, according to Poland, to circumscribe the physical ontology according to *any* set physical science, current, future, or ideal: rather, physicalism must recognise the changing nature of physical science, its method and content, and must accommodate and incorporate that in its formulation: "What the physical ontology consists in is, of course, *the* central question to be solved for proper development of physicalist doctrine. Recognition of the changing nature of physical theory and the consequent rejection of specific *a-priori* conceptions of the physical are critical aspects of that problem."[84]

Poland's formulation of physicalism begins with a characterisation of physics (properly basic physics) as a science concerned to answer certain kinds of questions about the fundamental nature of space-time and its occupants. Whatever properties that science says are physical are the properties *that we should take it* are physical. But this is an ideological derivation, and insofar as the ideology of physics is subject to change and revision, so too should our notion of what is to count as a physical property — or at the very least, a basic physical property — be subject to change and revision. Poland rejects the notion that constraints can be set upon the ontology of physics from outside physics. We are to take it that the properties identified by basic physics are physical properties, but for the purposes of physicalism what specific properties physics identifies — even, indeed, whether physics as characterised *exists* — is irrelevant.

Thus, neither Hellman or Poland provide an explanation of what is to count as a physical property, nor do they intend to, attempting instead to convey content upon the doctrine of physicalism while explicitly rejecting any constraints upon what is to count as physical. They do not provide accounts of what it is to be a physical property, not because the formulations of physicalism they offer fall short of doing so adequately, but because these formulations are constructed to avoid doing precisely that. Given the very serious problems already encountered for the elucidation of what is to count as a physical property, this does not appear to be a bad strategy.

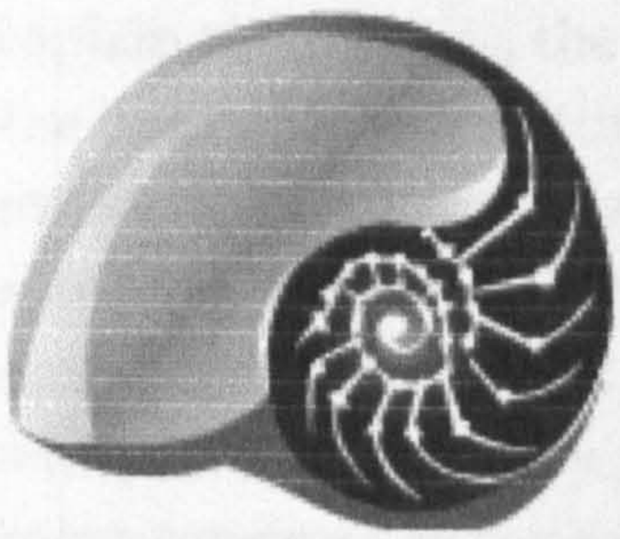
Having examined several possible ways of describing what it is to be a physical property and finding all wanting, Daly concludes that the notion of 'the physical' is not well defined, and so all metaphysical programs that assume a distinction between physical and nonphysical properties should be abandoned in favour of problems that would emerge irrespective of whether or not such a division were assumed. However, I believe that this prescription is a trifle hasty. It may be perfectly possible to construct a formulation of physicalist doctrine that is able to confer content upon the notion of physicalism without requiring *a priori* criteria by which to distinguish physical properties from nonphysical properties. Consequently, I agree with Daly that there are serious, perhaps insurmountable, problems involved in the attempt to specify what properties, amongst the great inventory of properties that we might identify, are to count as physical ones. I disagree with his conclusion that this makes the concept of physicalism itself empty. Rather, I think that important lessons can be learned from the difficulties that Daly has brought to light which indicate a direction for further development of the physicalist thesis, a direction already explored by Hellman and Poland. If we cannot say

what properties are to count as physical, then perhaps we can get somewhere by formulating a minimalist physicalism which does not draw on such a notion, but rejects it entirely, in favour of other kinds of constraints.



3
Reconsidering Reduction:

3. The Foundations of Physicalism



From the preceding discussion a picture of the character of physicalist doctrine has been constructed, along with some of the serious difficulties involved in the expression of that doctrine. Though generally presented against dualism as a thesis about the mind, to the effect that the mental is ontologically grounded in the physical, it should be clear that physicalism is a general doctrine, wherein the physicalist claims not just that *mental* things are so grounded, but that *everything* is in some way physical. The physical is the prime ontological category to which everything belongs or upon which everything depends. Further, physicalism places some constraints on the ways in which nonphysical properties can be dependent upon physical properties: it is not enough that they simply exist *alongside* physical properties, dependent upon the physical only for pure existence. Physicalists typically speak of nonphysical properties being *instantiated in* or *realised by* physical properties. I propose to differentiate the two terms, and shall henceforth use them as follows: I understand the notion of *instantiation* to be an expression of token identity, according to which each instance of a mental (or other) nonphysical property is identical with some instance of a physical property. *Realisation* I shall take to express a stronger relation of relevance, which I understand in both an explanatory and metaphysical sense.

Barry LePore and Ernest Loewer explain the notion of realisation as follows:

> The usual conception is that e's being P realises e's being F iff e is P and e is F and there is a strong connection of some sort between P and F. We propose to understand this connection as a necessary connection which is *explanatory*. The existence of an explanatory connection between two properties is stronger than the claim thatP→M is physically necessary since not every physically necessary connection is explanatory.[1]

Jaegwon Kim objects that though the idea behind LePore's and Loewer's explanatory requirement is right, the explanatory relation *simpliciter* should not be understood as

constitutive of realisation. Rather, an objective metaphysical relation should be taken as grounding the purely epistemological relation of explanation. He says:

> That *P* explains *M* cannot be a brute, fundamental fact about *P* and *M*... In the case of realisation, the key concepts, I suggest, are those of *causal mechanism* and *microstructure*. When *P* is said to "realize" *M* in some *s*, *P* must specify a microstructural property of *s* that provides a causal mechanism for the implementation of *M* in *s;* moreover, in interesting cases — in fact, if we are to speak meaningfully of "implementation" of *M* — *P* will be a member of a family of physical properties forming a network of nomologically connected microstructural states that provides a microcausal mechanism, in systems appropriately like *s*, for the nomological connections among a broad system of mental properties of which M is an element.[2]

Kim suggests that we take a realist stance towards explanation: if *P* explains *M* in some system *s*, it is because there is a specifiable causal mechanism in *s's* microstructure that is sufficient for *M* in *s*. In emphasising causal mechanism and microstructure, Kim implies that realised properties must be functional properties, to be distinguished from second-order properties generally. That is, I believe, correct: all realised properties must be construed functionally in order to understand their relation to the mechanisms that explain them. Thus, the property *water soluble* must be understood as something along the lines of *being a structure that loses internal cohesion when in water* in order to specify the microstructural mechanisms that result in the exhibition of this property of certain substances. To say, then, that the mental is *realised in* physical states is to make a much stronger claim than that mental properties are ontologically dependent upon, covariant with, or instantiated by physical properties. It is to claim that it is a fact about the microstructure of the system in which mental states occur that there are causal mechanisms that bring it about that there are mental states in that system: causal mechanisms which, had we the science to find them out, would ground a complete explanation of the obtaining of mental states in that system.[3]

This does not mean that for any type of phenomenon picked out on any level there exists a description or explanation *in physics:* there are some kinds of properties, such as *being a token of a monetary unit* or *being a desire that Ralph Nader win the US 2000 election,* for which there can be no interesting physical description,[4] and explanations involving such properties in themselves will not be interesting or illuminating on the physical level. But physical properties and physical interactions will back each instance of a property or interaction on any level of individuation.

I believe that Kim is right in advocating a realistic conception of the realisation relation. The idea that the properties of things can be explained and understood in terms of the properties of the things that comprise them is fundamental in actual scientific practice and in the ideology of physicalism. Taken realistically, the notion of physical realisation involves metaphysical and explanatory principles underlying the relation of realised properties to the physical properties doing the realising. Causal-nomic connections among appropriate microstructural relata are what underlie realised properties. Thus for the physicalist to entertain the notion of realisation, the physicalist must have some commitment to a realistic conception of an explanatory relation obtaining between properties.

The two principles which best capture these primary concerns of physicalist

doctrine without necessarily committing physicalists to potentially undesirable consequences are those of the causal closure of the physical domain and the supervenience of all phenomena upon physical phenomena. Supervenience, when applied to the physical domain the thesis that there can be no change at all without some physical change, expresses the physicalist's commitment to the dependence of all the facts upon the physical facts, while the principle of the causal closure of the physical domain is expressive of the metaphysical and explanatory constraints robust physicalism imposes upon the ideas of causation and composition.[5]

I want to say a bit about these principles, and to establish them as essentially underpinning the thesis of physicalism. Importantly, I will not attempt to establish the *truth* of these principles. I do not believe that either principle can be *proven:* there is much speculation on the nature of causation, on the idea of which any discussion of causal closure must be based, and the phenomenon is empirically opaque. Any definite answer to the question of whether the physical facts *do* determine all the facts is similarly in principle obscure. Thus, I can give no indefeasible reason why these principles should be believed. I hope to be able to say, however, why it is reasonable for serious physicalists to accept them, and how the considerations motivating the acceptance of these principles are the same ones motivating physicalism generally. My aim in this thesis is not to demonstrate the truth of physicalism to the anti-physicalist, nor is it to argue for the truth of the causal closure principle or the thesis of supervenience. Rather, in discussing the principles that I take to be at the heart of physicalism, I hope to expose some considerations which I believe physicalists ought generally to share, as well as weed out some problematic notions with which some physicalists have been concerned. Finally, I wish to make a case for a possible formulation of the physicalist thesis which may have some surprising consequences.

3.1. The Causal Closure of the Physical Domain

We have already encountered a version of the causal closure principle as presented by Papineau in his formulation of physicalism. There were some problems with his defence of that principle, some of which, I noted at the time, were resultant from his formulation of the principle. Recall that Papineau had said:

> I take it that physics, unlike the other special sciences, is *complete*, in the sense that all physical events are determined, or have their chances determined, by prior *physical* events according to *physical* laws. In other words, we never need to look beyond the realm of the physical in order to identify a set of antecedents which fixes the chances of any subsequent physical occurrence. A purely physical specification, plus physical laws, will always suffice to tell us what is physically going to happen, insofar as that can be foretold at all.[6]

The principle is formulated as an epistemological one applying to the science of physics, describing what Papineau refers to as the internal *completeness of physics.*[7] To formulate the principle in this way makes it necessary for Papineau to argue for this completeness, for there is no reason for us to assume that *physics* is complete; indeed, Papineau simply had

stipulated it to be so, and we saw how this had serious negative consequences for his argument. Further, though he states the principle in terms of the closure of the physical sciences, it is not clear that its definition is purely epistemological: the notion that all physical events are determined by prior physical events according to physical laws certainly would seem to constrain properties, entities, and states, not merely their descriptions in physics, and so to present a more metaphysical usage. Papineau perhaps could have saved himself this trouble had he formulated the principle as *the explanatory closure of physics* — all explanations of events describable in physics, he might have said, cite only causes also describable in physics, or, we need never go beyond the physical vocabulary to give a complete sufficient explanation of physical events. But had he kept the principle in a purely epistemological form, it is uncertain that it would have been able to do the work that he requires it to do in supporting the manifestability argument for supervenience and the causal argument for token congruence. I am not certain why he chooses to formulate the principle as he does: perhaps because he knows that he can describe *physics,* but not *the physical.* However, I do not think that this difficulty should stop us from understanding the principle as applying to the physical domain. Had Papineau so expressed it, he would not have had to give the arguments, which proved highly problematic, for the closure of physics, for we have good reasons to believe that *the physical domain* is complete and casually closed, even if *physics* is not.[8]

A domain is said to be *causally closed* if elements in the domain interact only with other elements in the same domain. Thus, for the physical domain, in tracing the causal history of any event falling under physical description, we need never move outside the physical domain to cite nonphysical causes. Each physical event has a complete sufficient physical cause. This implies that if a mental cause is cited in any causal chain involving an antecedent physical effect,[9] then that mental cause must somehow or other *be* a physical cause, for nothing but physical causes can result in physical effects.

Crane and Mellor had implicitly rejected this principle when they said that there is plenty of mental causation,[10] taking it that a pure mental event, qua mental, could be cited in any causal chain involving physical effects. This assertion was made in objection to the physicalist's desideratum that explanations of events should make explanatory sense. I had answered that to argue along these lines ultimately robs all the sciences of their explanatory power: if nonphysical events can straightforwardly and in themselves be causally efficacious with respect to the physical, we cannot expect that physical explanations, will be sufficient to describe physical chains of events, and the difficulty generalises over other sciences as well. But perhaps that is too strong. Indeed, the principle of the causal closure of the physical domain makes a lot of trouble for higher level properties not taken as strictly physical but assumed to be causally effective, as it seems intuitively plausible, even obvious, that so many of them are. In the face of such compelling evidence that mental events cause and are caused by physical events (and in the absence of compelling reasons to think that mental causes *are* physical causes), does it make sense to insist upon the causal closure principle? Can we not have perfectly good explanatory systems without it?

At the higher, more abstract levels of description, such as the psychological or

biological, it makes sense to regard the domain without any form of closure constraint. Indeed, it is necessary to do so, for it is an accepted truism that the generalisations employed at such levels are never exceptionless, but are subject to occasional falsification. However, when a law of the special sciences does not apply in a particular case, we take it that there is some *reason* for the exception to have occurred, or that something happened to *make* the exception occur. The laws of the special sciences are, universally, hedged or *ceteris paribus* laws, and can be interpreted as being of the form "All things being equal, C→E", or "C→E, unless, *for some reason,* not." Entities and events individuated in higher level sciences can be described in terms of entities and events in lower level sciences, and as such, elements from lower level domains can interact with elements in higher level domains in ways for which the sciences appropriate to that domain have neither terms nor descriptive apparatus. This can result in things happening on the higher level that might be completely unexpected in terms of the sciences appropriate to that level, or in higher level phenomena apparently having a somewhat unpredictable nature as far as their descriptive sciences are concerned. Hume puts it well:

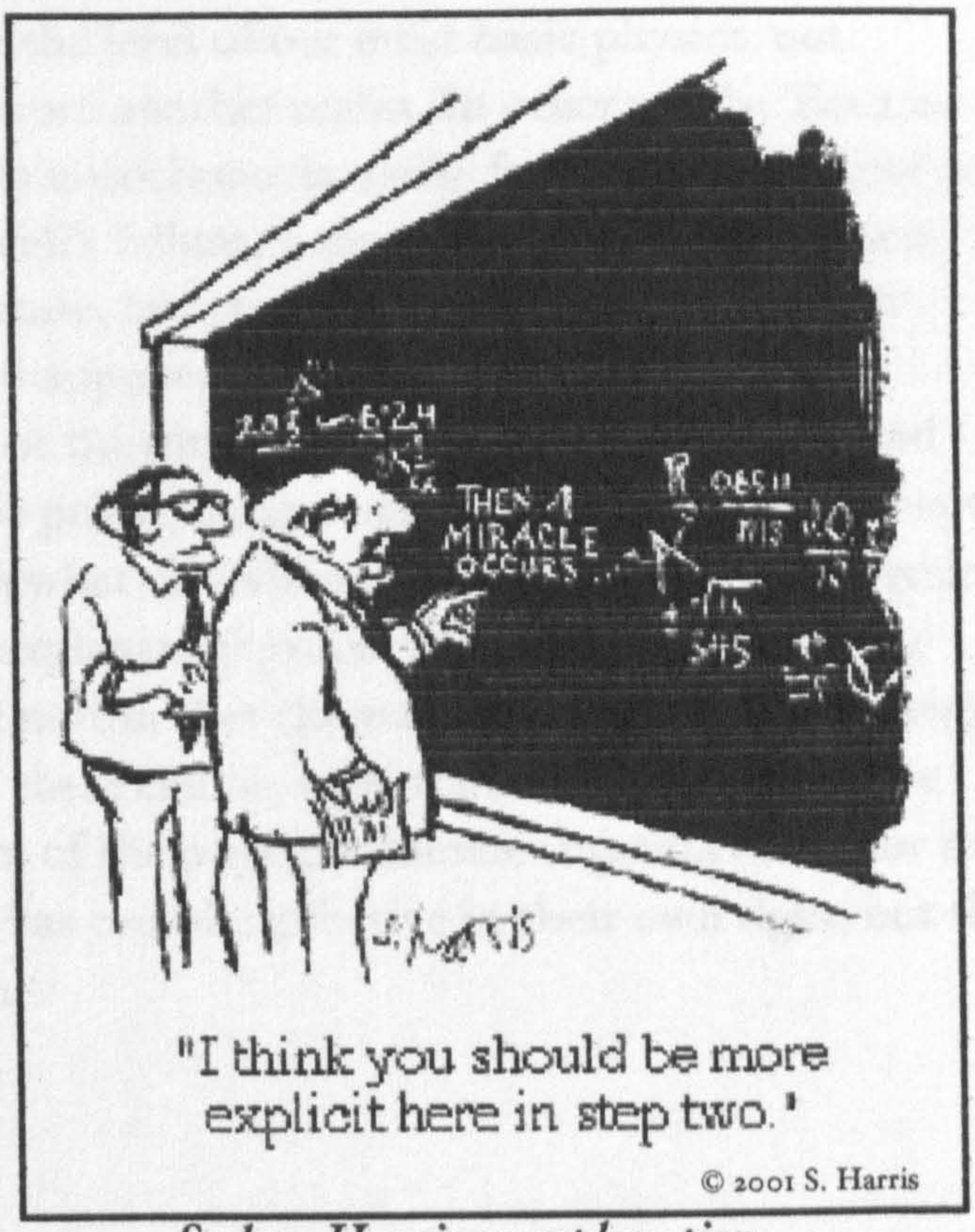
> The vulgar, who take things according to their first appearance, attribute the uncertainty of events to such an uncertainty in the causes as makes the latter often fail of their usual influence; though they meet with no impediment in their operation. But philosophers, observing that, almost in every part of nature, there is contained a vast variety of springs and principles, which are hid, by reason of their minuteness or remoteness, find that it is at least possible the contrariety of effects may not proceed from any contingency in the cause, but from the secret operation of contrary causes, and proceeds from their mutual opposition... From the observation of several parallel instances, philosophers form a maxim that the connexion between all causes and effects is equally necessary, and that its seeming uncertainty in some instances proceeds from the secret opposition of contrary causes."

The possibility of the falsification of higher level generalisations is not a brute, inexplicable fact about higher level generalisations. Uncertainty is not an intrinsic property of higher level generalisations, but is in fact perfectly explicable in terms of the lawlike operation of opposing 'springs and principles' on lower levels conspiring to bring about a effect other than what the higher level generalisation predicts.

At the level of basic physics, however, it is quite different. If a law in basic physics does not apply in any given instance, there can be *no reason* why it should have been falsified, for there are no 'opposing springs and principles' at a level lower than basic physics from which the failure of the law to hold in that instance could be understood to proceed.[12] Further, lower level laws cannot admit of falsifications by contributions from higher level phenomena, for each higher level phenomenon will be describable in lower level terms; the generalisations involved at the higher level will be backed by lower level generalisations and ultimately by physical laws. But nothing backs the laws of basic physics. Given that they are *basic,* they must simply be taken as given. There are no lower level dynamics to back falsifications of predictions in basic physics, for nothing backs basic physics. Any physcally explicable interactions at the fundamentally physical level, therefore, will take place completely within the domain of basic physics.[13] Anything interacting with the fundamentally physical domain which is not itself on that level does

so in a way that cannot be explained.

The falsifications of higher level generalisations will have reasonable explanations in lower level terms, because elements describable in lower level terms can interact with the elements described in higher level sciences. The reverse, however, is inexplicable. It is inconceivable how anything can interact with physical things on the level of basic physics except insofar as they fall under some description in basic physics. Thus, in order to meet the desideratum for explanatory sufficiency, the physical domain must be regarded as causally closed. Some may object, arguing that the inconceivability and inexplicability of the phenomenon is no reason to say that it does not happen, asserting that things outwith the physical domain could interact with the physical domain in a non-explanatory way. To maintain this position, however, would be tantamount to an assertion on interactionist dualism, and it should be clear that such a metaphysic cannot support a robust explanatory system. In stating that the rejection of the principle of the causal closure of the physical domain has serious and unacceptable consequences for the concept of sensible explanation, I intimated that the principle is derived from our intuitive rejection of the idea of substance interactionism, which problem was the downfall of Cartesian dualism. I believe that the principle of the causal closure of the physical domain is an expression of that reservation. If we did not have qualms about substance interaction, it would not be a mystery how a purely mental thing with no physical properties could affect a physical thing, and Descartes's postulation of spirits both subtle and fine whose vectors could be altered by the influence of the *Res Cogitans* would be perfectly acceptable. But it is not acceptable. The same is true for the causal powers of a thing *composed* of smaller physical things, for any mereological whole: if its causal powers are not derived from the causal powers of is components, then any effects it has upon the physical, described on the level of the parts that constitute it, are occult.



Sydney Harris on explanation.

One powerful idea behind the causal closure principle is simply that 'Weird things don't happen on the physical level', where 'weird' is to be taken in its archaic sense:[14] supernatural things don't happen, spooky things don't happen, and neither do unexplainable spontaneities of a sort not proprietary to the physical domain happen, on the physical level. We do allow in current physics for randomness, for spontaneous violations of the conservation laws, for genuine non-locatability of particles, for probabilistic outcomes of causal interactions, and for all manner of things that would have

stricken a classical physicist as downright occult. But when we do this, we take it that these are natural features of the physical at the level of our most basic physics, not strange miracles visited upon the physical from another realm. In other words: 'Because a random quantum event transported all of its mitochondria to the far side of the galaxy' is a perfectly valid explanation for a particular cell's failure to metabolise on some occasion. The chances of its being a *true* one are minute, but the point is that this event is not *miraculous:* it has a physical reason to have happened.[15]

The principle of the causal closure of the physical domain is a far-reaching and ultimately, it must be admitted, unprovable principle, and one may choose not to accept it. However, I believe that this principle is what underlies the sense in which the physical domain enjoys a certain metaphysical and explanatory primacy with regard to all other domains, and so to reject it is to reject the notion that there can be reasonable interlevel explanations as well as the possibility that there can be, even theoretically, a complete physics. The principle of the causal closure of the physical domain is problematic for the notion of higher level properties construed as causally effective in their own right, but the cost of its rejection, I think, is far too high.

3.2. Supervenience

Supervenience expresses a relation between two or more sets of properties such that there can be no variation in one set of properties without there being some variation in the other. One set of properties is said to be supervenient or to supervene upon the other set, whilst this is known as the set of subvenient properties, or as forming the supervenience base for the first set[16] . Supervenience might perspicuously be likened to the notion of composition: two mereological wholes with qualitatively identical sets of parts in all the same relations will, we expect, be qualitatively identical; if mereological wholes are different, we would expect there to be some difference in their constitution. The thesis that the properties of a mereological whole are determined by the properties and relations of its parts is, indeed, the thesis of mereological supervenience.

Three types of supervenience are generally recognised, termed *weak, strong,* and *global* supervenience. Weak supervenience nicely captures the basic sentiment of the relation given above. Jaegwon Kim has characterised weak supervenience thus: if A and B are two nonempty sets of properties, closed under standard Boolean operators, A supervenes upon B just in case:

> Necessarily, for any x and y, if x and y share all properties in B, then x and y share all properties in A — that is, indiscernibility in B entails indiscernibility in A.[17]

An equivalent definition is given as follows:

> Necessarily, for any object x and and any property F in A, if x has F, then there exists a property G in B such that x has G, and if any y has G, it has F.[18]

Note that this expression does not strictly *identify* having F with having a particular, specifiable property G. Weak supervenience only requires that if some x has some property in F in A, then it has some property in B such that if anything else had the same property in B, it would also have the same F in A. In order to identify the property F with the property G, G would have to be the unique supervenience base for F. In other words, the relation between F and G would have to be of biconditional form: for any object x and any property F in A, iff x has F, then x has G, and iff any y has G, y has F. No such biconditional is involved in the basic expression of weak supervenience given above: some z could have F in virtue of having some quite different property G* in B. But if any v had G*, v would also have F. Too, G could be a functional property, and so realisable in any number of mechanisms (the function may be supervenient upon the properties and relations of the parts of whatever performs the function — supervenience is transitive and reflexive). Thus when we read G, we would understand 'whatever performs the function that is specified by G' as being the physical[19] base. G could refer to a disjunctive property as well — suppose that some property, for example, *being a carburettor,* can be realised in a certain number of physical constructions (P1, P2, P3,...Pn). Then if any x is a carburettor, then x has G where G is some member of the set (P1, P2, P3,...Pn), and if any y has any member of that set, then y too is a carburettor. The relation of supervenience is thus much weaker than that of identity.

A distinction can also be drawn between reading the property G as a B-maximal or as a B-minimal property in B. If G is taken as a B-maximal property, then G is *every property in B possessed by some x.* Thus when we read 'if X has F, then x has some G in B, and if any y has G, y has F' we should understand the claim as being 'if any y is B-identical to some x, then y has F'. If B is the domain of physical properties and A the domain of psychological properties, as is typically taken for the psychophysical case, then the claim is that if any y is physically identical to some x, then y is psychologically identical to x.[20] This would be consistent with taking F as expressing every property in A possessed by some x. However, if we wish to take F as a particular property in A, then it is probably perspicuous to regard G as being a subset of the properties in B possessed by some x, as we might like to take it that individuals who are not in every way identical may possess some of the same properties.

To take an example borrowed from Kim[21] for the case of moral properties construed as supervenient upon nonmoral properties: suppose that St. Francis is a good man. We would naturally expect that if anyone were exactly like St. Francis, in that they possessed all of the intrinsic properties St. Francis possesses and are embedded in a sufficiently similar environment, that that person (let's call him Ed) would also be a good man.[22] However, it is certainly too much to say that one has to be exactly like St. Francis in all of these respects in order to be good, or to resemble an intractably vast and possibly infinite disjunction of properties: (*being just like St. Francis* v *being just like Mohammed* v *being just like Florence Nightingale* v... v *being just like n*). Rather, it seems that one can be good in virtue of possessing certain properties in common with St. Francis that are relevant in some important way to his being good while failing to share other properties he has

which (probably) have nothing to do with his being good: being about 5'8",[23] for example, or being disposed to talk to birds.

Let us, for the sake of brevity, construct a highly simplistic account of goodness, and say that St. Francis is good in virtue of being courageous (C), being benevolent (V), and being honest (H), such that the property *being good* in the domain A supervenes upon possession of the conjunction of properties in the domain B *being courageous* and *being benevolent* and *being honest.* Call this conjunction Gm, for G-minimal. Now, if Ed or anything else possesses Gm among any other properties it may have, it will be good. Further, it is not necessary to take just this minimal set of St. Francis's properties as the supervenience base for *being good.* Minimal properties can be disjunctive as well. For example, suppose that it is not possession of the full conjunction (C,V,H) that forms the supervenience base for *being good,* but only possession of any two of them. Then there are eight possible conjunctions of these three properties — (C,V,H), (C,V,-H), (C,-V,H), (-C,V,H), (C,-V,-H,), (-C,V,-H), (-C,-V,H), and (-C,-V,-H), and only four of them — the first four — are Gm, and form the supervenience base for *being good.*

Weak supervenience is generally considered to be too weak to meet the requirements of robust physicalism committed to metaphysical and epistemic explanatory relations obtaining between levels of properties. The reason is that under weak supervenience, nothing holds B-properties necessarily to the A-properties for which they serve as a supervenience base, such that what A-properties supervene upon which B-properties is arbitrary and world-relative. For illustration, consider an example for the psychophysical case.

Consider a a possible world W1 in which physicalism is true, and consider the kinds of properties that many physicalistically inclined philosophers would think likely to be relevant to a particular individual's being conscious. It does not particularly matter what those are, just so long as it is understood that there are some sorts of properties that could be so relevant, and other sorts of properties that simply do not fit the bill. Let us assume a fairly probable base set of properties: say that it is a certain functional organisation of physical bits and pieces relative to a certain system description **D***s* (such as a neural structure, possibly among other things, might instantiate relative to a human being), and call this organisation O. *Being conscious* (in A) then will weakly supervene upon *having O relative to* **D**s (in B) if any individual that is conscious in this world has O relative to **D***s* (remember that *being O* is a multiply instantiable minimal base property), and any individual that has O relative to **D***s* is conscious.

Now consider another possible world W2 that is physically identical to W1 except in the minor detail that it is not the property O relative to **D***s* that is the supervenience base for consciousness, but rather it is the minimal set of properties which in this world is also responsible for photosynthesis. Here is a world with parsons, periwinkles and pocket calculators, in which none of the people are conscious, but much of the plankton is. W1 and W2 are severally and together perfectly consistent with weak supervenience:

> Necessarily in W1, if any x is conscious, then x has O, and if any y has O, then y is conscious.

Necessarily is W2, if any x is conscious, then x is a photosynthesiser, and if any y is a photosynthesiser, then y is conscious.

Under weak supervenience, A-properties can be redistributed over B-properties any way you like. Whatever strongly connects A-properties to B-properties is not captured by this relation. The remedy for this is to formulate supervenience so that it applies with cross-worlds necessity. This is *Strong Supervenience.* A family of properties A strongly supervenes upon a family of properties B just in case:

Necessarily, for any object x and any property F in A, if x has F, then there exists a property G in B such that x has G, and *necessarily* if any y has G, it has F.[24]

The addition of the second necessity condition guarantees that strong supervenience holds across words, and so that if *being conscious* strongly supervenes upon *having O* in any world, it does so in every world, and if it does not strongly supervene on *being a photosynthesiser* in some world, it does not in any world. Anything that has O in any world will be conscious, and if there is a conscious photosynthesiser in any world, its consciousness will have nothing to do with its photosynthesising and everything to do with its having O.

This does capture the sense of a robust relation of relevance obtaining between properties, such as can support the expression of the relation of realisation as a strong metaphysical and explanatory relation discussed above, and as answers to our intuitive notion that there is something about properties on a lower level that makes it such that they conspire in certain relations to realize the properties that they do: how it is a matter of reason, as it were, and not just a matter of fact, that certain kinds of phenomena underlie other kinds of phenomena. But it is important to recognise that while the relation of strong supervenience does *express* that sense, it does not *entail* it. Cross-worlds necessity can be just as brute and inexplicable as world-relative necessity: instead of the properties involved with our hypothetical functional property *having O relative to Ds* being the supervenience base for consciousness, it could be *having both a heart and a kidney;* it could be *being a featherless biped;* if you're deeply skeptical about other minds, you may even entertain that it is *being (insert your name here).*

3.3. Supervenience: The Foundation of the Relation

In his Supervenience and Materialism, Mark Rowlands argues that weak supervenience is not just a weak determinative relation, but that it is not a relation of determination at all. He arrives at this conclusion through development of a problem originally raised by Simon Blackburn.[25] Under weak supervenience, as we have seen, B-indiscernibility entails A-indiscernibility within a given possible world, but allows that individuals across possible worlds may be B-identical but A-discernible. Let some

property F in A supervene upon some property G in B in some worlds, and call these worlds G/F worlds. Call those possible worlds where F does not supervene upon G G/O worlds. Following Blackburn, Rowlands asks: why there are no mixed worlds? "That is, why are there no worlds of the form G/FvO? These worlds are ruled out by weak supervenience, but it is difficult to see the justification for this. After all, weak supervenience allows both G/F and G/O situations... The difficulty is that once we have imagined a G/F world and a G/O world, it is as if we had done enough to imagine a G/FvO world, and have implicitly denied ourselves a right to forbid its existence."[26]

There are, of course, worlds where some things are G and F and some things are G but not F. In such worlds, F does not supervene upon G, for if it did, all G-things would be F-things.[27] The difficulty Rowlands identifies comes from the reading of weak supervenience as indicating that in G/F worlds, being G *determines* being F, while in other worlds, it does not. He suggests that such a determinative relation is more than a matter of brute metaphysical fact, but that there is something about G-ness that *"makes for"*[28] F-ness. If being G "makes for" being F in some world, but not in another, then it would appear that G alone is not sufficient for F. It is necessary also to make reference to some aspect of G/F worlds that plays a role in F's supervenience upon G. This indicates that F does not supervene upon G *simpliciter,* but that in G/F worlds, F supervenes upon G *and* some aspect of the world in question. Rowlands concludes:

> ...the moral of Blackburn's mixed worlds problem is not weak supervenience is a weak relation, or even an extremely weak relation. The correct inference , I think, is that weak supervenience is not a relation of determination at all. If G supposedly determines F in a G/F world, but not in a G/O world, then it cannot, after all, be G which is determining F in the G/F world. At the very least, the determining factors have to include not only G but also some aspect of the world in which G is instantiated."[29]

In order to read supervenience as expressing a genuine determinative relation, Rowlands suggests that it is necessary to adopt a realist stance towards the supervenience relation. This involves accepting the proposition that there is indeed something about G properties which "makes for" F properties, such that anything which has G has F. He defines realism about the supervenience relation as follows:

> RS: An interpretation of supervenience is *realist* if and only if it claims that:
> (i) subvenient and supervenient properties are related as the (preferred) definition of supervenience claims they are related. And
> (ii) for any given object x which instantiates both a subvenient property G and its corresponding supervenient property F, it is the possession of G by x which makes for its possession of F.30

Since weak supervenience is not strong enough to capture the sense in which G-properties "make for" F-properties, Rowlands suggests that the preferred definition of supervenience is the cross-worlds stable relation of strong supervenience. Since strong supervenience does not permit the coexistence of G/F worlds and G/O worlds, it is not vulnerable to the problem of mixed worlds. However, the essence of the difficulty he

perceives for weak supervenience is that it does not capture the sense in which it is something about G-ness that necessitates F-ness. While strong supervenience may minimalistically capture that sense, it is not fully expressive of it, and Rowlands's problem can be extended to strong supervenience as well. Given that it is the case under strong supervenience that G-things are F-things in worlds where supervenience holds, we can still ask a further question: why?

Rowlands draws a distinction between realism about the supervenience relation and the reification of the relation, and cautions that to interpret supervenience as a Real Relation obtaining between two sets of properties is to invite anew the concern that it is not G-ness *simpliciter* that makes for F-ness, but that something about the world in question — the supervenience relation itself — is a partial determinant of F-ness in G things.[31] Under this interpretation, it would appear to be possible for there to coexist two possible worlds identical in every respect save in the presence of the supervenience relation: in both possible worlds, all G-things are F-things, but in one, the relation is due to the existence of a relation of supervenience obtaining between G-properties and F-properties, while in the other, it is a matter of brute coincidence that all G-things are also F-things. That is, the coexistence of G/F and G/O worlds is possible.

Rowlands's arguments turn on the reading of supervenience as a determinative relation, which reading he suggests is the natural way to interpret supervenience.[32] However, while the reading of the supervenience relation as one of determination of one set of properties by another, or the dependence of a set of properties upon another, may be an intuitive and natural reading, it may be questioned whether it is quite the correct one. As Kim has pointed out, the relation of supervenience cannot be an explanatory relation, because it is compatible with different, mutually exclusive explanatory theories: epiphenomenalism, some forms of dualism, and type-identity theory, for example, all may assert that mental states are necessarily associated with certain physical states, but the nature of the relation between mental properties and physical properties is very different between them. The relation of supervenience can be said to hold whether it is a matter of lawlike connections, mysterious harmonies, or brute covariation between sets of properties. Supervenience by itself, then, cannot explicate what it is about G-properties which "makes for" F-ness in things which possess G. Unaccompanied by a separate account grounding the relation of supervenience in some specified determinative relation, supervenience itself only expresses the covariation between families of properties, leaving it open why it is that those families covary. The relation of supervenience cannot stand alone, but needs to be grounded in some specified relation of determination. Discussing the psychophysical case, Kim says:

> We must conclude then that mind-body supervenience itself is not an *explanatory theory;* it merely states a pattern of property covariation between the mental and the physical and points to the existence of a dependency relation between the two. Yet supervenience is silent on the nature of the dependence relation that might explain why the mental supervenes on the physical. Another way of putting the point would be this: supervenience is not a *type* of dependence relation — it is not a relation that can be placed alongside causal dependence, reductive dependence, mereological dependence, dependence grounded in definability or entailment, and the like. Rather, any of these dependence relations can generate the required covariation of properties *and thereby qualify as a supervenience relation.*[33]

Rowlands argues that weak supervenience is not a relation of determination at all. He is right: it is not. But neither is strong supervenience a relation of determination or dependence. Though it is strong enough to capture the sense in which higher level properties are genuinely determined by lower level properties, so that possession of the lower level properties in question alone suffices for possession of the higher level properties, strong supervenience needs to be grounded in some relation of determination that is a reason why it should obtain. Kim's wording indicates that any dependence relation that grounds the supervenience relation is a supervenience relation.

It is preferable to leave the interpretation of the modal operator 'necessarily' open in the formulation of supervenience, to accommodate different relations of determination that may qualify as supervenience relations. For the psychophysical case, as for the supervenience of higher level properties upon lower level properties generally, it is usual to read the first 'necessarily' in the formulation of strong supervenience as metaphysical necessity, and the second as nomological necessity. That is, if the property *being conscious* supervenes in this world upon *having some functional organisation O relative to system description* ***D***s, it does so as a matter of natural law in this world. Further, and fairly obviously, the cross-worlds stability of strong nomological supervenience only holds across worlds with the same laws of physics, as is sometimes explicitly stated.[14] There is no reason why the higher level physical properties that supervene upon lower level physical properties in this word should do so in an alternate possible world in which different physical natural laws apply. Also, we would expect that someone just like St. Francis in this world, in a relatively similar cultural climate, would be a good man, but we would not expect the goodness of such a person to be preserved were he transplanted to a culture wherein goodness were measured according to different standards. Since it is not desirable to take the laws of physics, or the conceptual systems definitive of moral or aesthetic properties, or many other such relations of dependence, as being metaphysically necessary, we should take it that the first necessity condition in the expression of strong supervenience holds only across worlds where what grounds the second necessity condition is identical.

Does this commit us to a form of what Rowlands had called the reification of the supervenience relation? For if it is necessary to ground the supervenience relation in some more explanatory determinative relation, then it would appear that again G *simpliciter* does not suffice for F in G/F worlds, but rather it is having G *and something else,* whatever relation grounds supervenience in this case — natural laws, conceptual necessity, or whatever — that are jointly sufficient for G. It would seem again to be possible for G/F and G/O worlds to coexist.

For cases of strong supervenience where the grounding relation is not metaphysically necessary, it would appear to be obligatory to specify that the worlds across which strong supervenience is to hold all be identical as to the grounding relation referred to by the second necessity condition in the formulation. However, I suggest that a far tidier alternative is to read the grounding relation as forming part of the supervenience base for supervenient properties. To do so obviates the need to specify

that whatever grounds supervenience must be the same across possible worlds in which strong supervenience holds. That is, natural laws would be part of the set of lower level properties upon which psychological properties supervene for cases of psychophysical supervenience. It is standardly taken that physical laws and physical properties are different kinds of things, but I suggest that for the purposes of formulating the abstract relation of supervenience, they should be reckoned together. This reading preserves the sense in which it is supposed to be G-properties *simpliciter* which suffice to determine F in G/F worlds, since the determinative relation grounding the supervenience of F upon G is included in G. It has the consequence, which might appear counterintuitive, that two individuals described as physically identical but embedded in different possible worlds with different natural laws were not in fact physically identical. However, I believe this is correct: if the explanations for why two apparently physically identical individuals exhibit certain higher level properties are exclusively different explanations as a result of the individuals' microstructures being subsumed by different sets of natural laws, surely we would not wish to call them identical. In any case, the relation of supervenience will be a different one, as if the natural laws are not taken as part of the supervenience base for supervenient properties, it will be necessary to specify that the relation only holds relative to a certain set of natural laws, which *ex hypothesi* it is not. According to the understanding of the determinative relation grounding and constitutive of the supervenience relation as part of the base for supervenient properties, this confusion cannot arise.

3.4. Reduction

The picture of a physically grounded world is one of a layered hierarchy, ordered such that phenomena at certain levels in the hierarchy, in appropriate relations, are constitutive of phenomena at higher levels, where the properties of those phenomena that are the parts of things at higher levels determine the properties of the wholes they constitute. Different determinative relations may obtain between the levels, but in general the relation obtaining between individuals on different levels (allowing, of course, for a certain indistinctness of boundaries) is one of parts to wholes: every individual on every level in the hierarchy except for the lowest is entirely comprised by individuals and relations on lower levels. Whatever is on the lowest level (if there is one) must be taken as brutely given, without decomposition, whose properties admit of no explanation, but which is constitutive of everything on every level.

Many kinds of mereological entities at different levels of individuation will be straightforwardly identical with the conjunctions of their parts in the appropriate relations: *being water* just is *being composed of two hydrogen atoms and one oxygen atom electrically bound.* Likewise, many properties will be straightforwardly reducible to physical properties: *being a liquid* is just having one of a number of molecular structures that cause a substance to behave in a certain fluid way. This property will be explicable in terms of

the properties and dynamics of lower level parts of molecules, and so eventually in general basic physical terms. Such properties, theoretically at least, will be reducible in this way, even if the complexity of the microphysical explanations involved would make the practical reduction of these properties a daunting of not outright impossible proposition. Some kinds of properties however, psychological properties paradigm among them, are not so reducible, because they are multiply instantiable, functionally described properties, in themselves indifferent to the physical structures that realize them. This indifference goes far beyond the indifference to their specific realisers of properties like, say, *being a liquid* that liquids *qua* liquids could be said to have. *Liquidity,* it is true, is a property that can be exhibited by any of a vast number of possible physical configurations, but at a very low level of description, there will be a kind of microphysical dynamic that is common to all things that are liquids. There will be a general physical account of liquidity that explains why liquids as such behave as they do. For some more complicated kinds of properties, however, different sorts of microphysical dynamics might perform the function characterising by the property in question, let alone different specific kinds of physical stuff. For example, one could build a carburettor out of all manner of materials, and might design it in any number of ways, so long as that structure, whatever it turned out to be, was such that it was able to perform the function that a carburettor performs. Many kinds of animals have hearts, or organs that serve the function of moving blood about the body, but these organs are wildly diverse in form across the animal kingdom. We believe that creatures of extremely varied descriptions — me, my dog, an octopus, or a silicon-based life form from Mars — would all feel pain if stimulated in certain ways, even though we all have very different biological structures. We are each able to feel pain because *pain,* like *being a heart* or *being a carburettor,* is defined not in terms of whatever physical structure realises it, but by the function that it performs in the system in which it is instantiated: a certain rather unpleasant detection of damage to the body, resulting in behaviour designed to isolate and avoid the cause of such damage, and so on. Many kinds of mechanisms could serve this function in different systems — perhaps an unlimited number of heterogeneous physical structures could fulfil this causal role. Thus *pain* is entirely indifferent to the specific physical base that realises it, and to the specific manner in which any particular physical base goes about performing that function.[35]

The multiple instantiability of mental states has led many, if not most, philosophers of mind to believe that the mental, and indeed any properties conceivably realised in wildly heterogeneous physical bases, such as economical properties, cannot be reduced to the physical in any meaningful or interesting way. Particularly, it is generally believed that the multiple instantiability of such properties prevents the special sciences in which they feature from being reducible to the physical sciences. What this irreducibility indicates is a subject of active debate: for some the irreducibility of psychological generalisations to physical laws indicates that there are, properly speaking, no psychological generalisations; for others it indicates the autonomy of psychology and other special sciences *vis-à-vis* the physical sciences. Whatever it is taken as entailing, the irreducibility of the mental based upon the heterogeneity of its possible realising

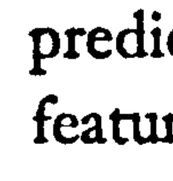
bases is largely taken to be *true*, and reductionism has fallen by the wayside as simplistic and untenable.

There is also a powerful appeal in the notion of the independence of the mental from the physical domain. For many, the ideas of the reducibility of the mental states to physical states and processes or the reduction of straightforward folk psychological explanations of human thought and action to hard neurobiology imply that the mental is *nothing but* physical states and processes, that the mind is *only* the brain, and that we in all of our loftiest aspirations are merely atoms in the void. If an elect set of properties, mental properties foremost, can be shown to be perfectly indifferent to the particular arrangement of atoms and void realising them, then the idea of the ontological dependence of the mental upon the physical loses much of its unappealing bite. The multiple instantiability of the mental is thought to open the way for a naturalistic account of mentality according to which the mental is instantiated in and supervenient upon physical events and processes, but also according to which mental state types cannot meaningfully be given analyses in physical terms, and according to which the generalisations subsuming mental dynamics are not reducible to physical laws. Such an approach can be read as preserving the natural dignity of the mental of which traditional reductionism would rob it.

The wholesale rejection of reductionism at this point in the debate is, to my mind, a little hasty. Physicalism is a doctrine with very wide appeal in philosophy as well as in other academic disciplines, and I think it is safe to say that most modern philosophers of mind in the Western tradition would categorise themselves as physicalist. Very few, however, will admit to being reductionists. But if nonreductive physicalism has its appeal, it also has its problems, chiefest among these being that of the provision of a coherent account of how the mental *qua* mental can be causally effective in the physical world in which it is grounded. It will be the burden of the next chapter to show why I think this is a problem that nonreductive physicalism is unlikely to be able to solve. Now, it will be interesting to turn to the problem generated for reductionism by the multiple instantiability of the mental (among other physically irreducible things), and to evaluate whether or not there can be a solution.

If I may be permitted to foreshadow the following chapter somewhat: I think that there can be a solution. Moreover, I think that there *must* be. It seems to me to be inescapable that *if* physicalism is true and all phenomena are grounded in physical phenomena, and *if* it is the case that mental events *qua* mental are genuinely causally effective aspects of the world, then some form of psychophysical reductionism *must* be true, at least on the level of spatiotemporally individual events. According to robust physicalism, every causal interaction on any level of description is grounded in physical interactions and is backed by physical laws. If the mental is physically realised, but for any specific instance of a mental event causing or being caused by a physical event the psychological generalisations subsuming that causal interaction float entirely free of physical laws, then surely there can be no genuinely lawlike connections linking mental phenomena to anything at all, and no nomologically necessary connections binding thought and action. That is, if mental generalisations float entirely free of physical laws,

then either physicalism, at least as I have understood it, is false, or else the mental, *qua* mental, is not a genuinely causally effective aspect of the world. Neither horn of this dilemma is terribly appealing, but perhaps the dilemma itself can be avoided by re-evaluating the notion of reduction to the physical and to physical laws.



3.5. Intertheoretic versus Functional Reductionism

The model of intertheoretic reduction standard throughout the literature, attributed to Ernest Nagel,[36] involves the derivation of the laws of a science to be reduced from the laws of a reducing science accompanied by identity statements, or 'bridge laws', linking the terms of the two sciences. A science S is reducible to a science R just in case, for any generalisation Sx→Sy, Sx and Sy can be defined in R's terms such that Sx↔Rx and Sy↔Ry, and a law exits or can be written in R such that Rx→Ry. The problem presented by the multiple instantiability thesis for such reduction is the extreme unlikelihood of the existence of such bridge principles linking physical predicates, such as could feature in physical laws, with predicates in special sciences featuring multiply instantiable properties. Jerry Fodor expresses it well for the case of economics:

> Suppose Gresham's 'law'[37] really is true... I am willing to believe that physics is general in the sense that it implies that any event which consists of a monetary exchange (hence any event which falls under Gresham's law) has a true description in the vocabulary of physics and in virtue of which it falls under a law of physics. But banal considerations suggest that a description which covers all such events will be wildly disjunctive. Some monetary exchanges involve strings of wampum. Some involve dollar bills. And some involve signing one's name to a check. What are the chances that a disjunction of physical predicates which covers all these events (i.e., a disjunctive predicate which can form the right hand side of a bridge law of the form 'x is a monetary exchange ↔...') expresses[38] a physical natural kind?[39]

The same is true for the psychological as for the economical. That there are physiological kinds coextensive with psychological kinds such that laws linking the antecedents with consequents in psychological generalisations have meaningful expressions in physical law is a highly dubious proposition. Physical descriptions might obtain for each instantiation of a mental property, and physical laws may back each instance of mental causation, but the generalisations of psychology will not be reducible to the laws of physics in any way that makes sense.

Kim gives a compelling criticism of the standard Nagelian derivational model of reduction. Beyond questions about the availability of bridge laws linking the predicates of the special sciences to physical predicates, Kim questions the appropriateness of the model from considerations of explanation in reduction and the ontology admitted by the reducing sciences. First, he points out that Nagelian derivational reduction is in essence the deductive-nomological model of scientific explanation forwarded by Hempel and

Oppenheim in 1948:[40] "Just as Hempelian explanation consists in the derivation of the statement describing the phenomenon to be explained from laws taken together with auxiliary premises describing relevant initial conditions, Nagelian reduction is accomplished in the derivation of the target theory taken in conjunction with bridge laws as auxiliary premises. It is therefore more than a little surprising that while the deductive-nomological model has had few adherents for over three decades, Nagel's model is still serving as the dominant standard in discussions of reduction and reductionism."[41] The problem faced by the deductive-nomological account of scientific explanation is, classically, that the model fails to distinguish between genuine laws of nature and accidental generalisations, a failing that easily can be seen to carry over to the Nagelian derivational model of intertheoretic reduction. Bridge laws linking physical predicates with special science predicates realised in heterogeneous bases do seem less like *laws* and more like accidental generalisations.

Kim's further questions address the degree of satisfaction we may derive from the Nagelian model of intertheoretic reduction in terms of its fulfilling the role we would expect a model of *reduction* to fulfil. Bridge laws, we have seen, express covariations between the predicates of the special sciences and reducing sciences. But these covariations are simple statements of identity: the Nagelian model does not require that anything back the correlation between a predicate in the special sciences and a physical predicate. No particular relation between a special science predicate and a physical predicate is required beyond covariation. This fact makes it the case that Nagelian reduction of psychological predicates could carry through even in a dualist, epiphenomenalist, emergentist, or dual-aspect system — if a correlation between physical predicates and psychological predicates can be identified, then the reduction can go through.[42] But this does not reflect the actual concerns shared by many philosophers over the possibility of the reduction of conscious mentality to the physical, for what is behind the reservations about this reduction — the essence of the 'Hard Problem' of consciousness, as David Chalmers has expressed it[43] — is the inexplicability of the phenomenon of consciousness in physical terms. It certainly appears to be true that consciousness constantly covaries with certain neurobiological functions, and there have been fascinating developments in neurophysiology linking specific parts of the brain and specific processes with sensory phenomena and reports of phenomenology. But nothing about the fact of such correlations explains *why* they should obtain, and the hard problem remains. The hard problem, the explanatory gap, presents no obstacle to Nagelian reduction. But surely it *should* be an obstacle to the notion of reduction. Kim phrases it quite strongly:

> When the emergentists claimed that consciousness was an emergent property that could not be explained in terms of its physical/biological "basal conditions", it was these explanatory questions that they despaired of ever answering. For them reduction was primarily, or at least importantly, an explanatory procedure: reductionism must make intelligible how certain phenomena arise out of more basic phenomena, and if that is our goal, as I believe it should be, a Nagelian derivational reduction of psychology,with bridge laws taken as unexplained auxiliary premises, will not advance our understanding of mentality by an inch. *For it is the explanation of these bridge laws, an explanation of why there are just these mind-body correlations, that is at the heart of the demand for an explanation of mentality.*[44]

Conceivably, one could impose an explanatory constraint upon the bridge laws. This measure would not address Kim's ontological question: the question of what a science must recognise. Bridge laws, it should be clear, dramatically increase the ontology of a reducing science, for in order to subsume the laws of the science it reduces, it must add the bridge laws linking the predicates of the reduced science to its own predicates as basic laws of itself. But as Kim says, "We expect our reductions to yield simpler systems — a simpler system of concepts, or simpler system of assumptions, or simpler system of entities."[45] Bridge laws, rather than simplifying, are highly complexifying. Thus the Nagelian model does not answer to the reductionist demand that, as Kim puts it, "reductions must *reduce*."[46] The addition of the bridge laws as basic laws of the reducing theory also has this possibly unsettling implication: it is a direct barrier to the possibility of a complete physics. For if physicalist reductionism in the sciences is to carry, all sciences must ultimately be reducible to basic physics. Thus, bridge laws would have to be taken as laws in basic physics. But we do not and cannot know what all the bridge laws are: a complete inventory of the bridge laws linking every actual and possible predicate of any actual or possible science to the predicates of physics is impossible. So basic physics, bridge laws intact, could never be complete. But it does not seem correct to think that the very notion of the decomposition of all phenomena to physical phenomena means that physics cannot be a complete, self-contained science. Indeed, this is quite the reverse of the sentiment behind reductionism.

These may not be hard and fast reasons to reject the Nagelian bridge law model of reduction, but they are strong motivations for such rejection. One might apply an explanatory constraint to bridge laws to address the explanatory concerns of reductionist physicalism, and one might be happy to admit the infinite and unknowable bridge laws into the ontology of physics. Then, one may not. The idea of laws linking the predicates of physics to the predicates of the special sciences may be an intuitive and appealing one, but as has been shown, it is highly problematic for the idea of reduction. Certain of these difficulties — specifically, the question of the availability of bridge laws linking predicates referring to multiply instantiable properties to physical predicates — have led to the near extinction of reductionists in the philosophical community. But perhaps it is the Nagelian model that is at fault, and not reductionism *per se*.

3.6. The Functional Model of Reductionism[47]

Kim has forwarded a model of reductionism that not only accords better with the concerns of physicalists as expressed above as well as with actual scientific practice than does the standard model, but which does not face the difficulties for reductionism presented by the fact of the multiple instantiability of certain properties. Instead of speaking of intertheoretic reduction, or reduction of one science to another, Kim focuses on reducing properties as distinguished on higher levels of individuation to properties as

distinguished on lower: a direct correlate of the notion of special science properties by proprietary properties of the lower level, ultimately physical, sciences, given the layered model of the world already indicated by the theses of supervenience and physical realisation. According to this model, the key factor determining whether or not a property described in one science can be reduced to a property described in another is whether or not the first property can be construed as a functional property, whose function can be fulfilled by a property in the reducing science.

According to Kim's model of reduction, to reduce a property one must first construe it as a second order property defined by its causal role: that is, the property must be construed functionally or extrinsically, as the property of *having whatever it is that normally defines that property*. The idea of a second order property is explained as follows:

> *F* is a *second order property* over set **B** of base (or first-order) properties iff *F* is the property of having some property *P* in **B** such that *D(P)*, where *D* specifies a condition on members of **B**.[48]

Functional second order properties are further described as "those second-order properties over **B** whose specification *D* involves the causal/nomic relation"[49] (as against, for example, the determinable/determinate relation, or the relation of conceptual necessity). Thus *being green* so construed is having the property of reflecting only frequencies of light in a certain range. That property is *reduced* when we are able to specify a first-order property that fulfils the condition *D*: certain molecular structures, in virtue of having certain structural properties, absorb most frequencies of light, allowing only the frequencies in the green range to be reflected. Thus the property *being green* can be straightforwardly identified with having such-and-such a molecular structure that absorbs most non-green frequencies in the visible spectrum and reflects green light. Reduction, Kim argues, need involve nothing more than this: a property is successfully reduced if it can be construed in functional terms, and if a causal mechanism can be specified that performs the function involved.

This kind of property reduction is suggested in the notion of the realisation of a property as explained above. A certain causal mechanism which performs a certain function in some complex whole realises a property F ascribed to that whole when the property is construed functionally, and the mechanism in question is sufficient for the instantiation of that function in that whole. Further, according to the functional model of reductionism, the mechanism is not only *sufficient* for F: it *is* F. Ontological simplification, Kim argues, can only be bought by strengthening the notion of property covariations into property identities, and the very way to do this is to construe those properties we would reduce not as intrinsic properties of the things to which we ascribe them, but extrinsically, as functions performed by some mechanism. A property is reduced if it is possible both to give a functional account of it and to find a causal mechanism in the microstructure of the object to which the property is ascribed that performs the specified function in that object. Some things appear green because the electrons in the outer shells of certain kinds of molecular structures are excited when they encounter light of certain frequencies, thus absorbing those frequencies, but not green ones; green

frequencies are reflected, so the only light that enters our eyes having bounced off of objects with these molecular structures is green light. When this can be done successfully, an explanation of the property is also given.

Because the functional model of reductionism takes the causal mechanism sufficient for the implementation of a second-order property F in some system to be identical with F in that system, it has no need for bridge laws expressing identities between special science predicates and physical predicates. Each second-order property, and each composed entity, is identical with the functional organisation of its microstructure. There need not, however, be tight connections between *kinds* of entities as picked out by the special sciences and physical kinds. There is also room for the boundaries of realised entities to be left fuzzy, compositionally as well as, to a certain extent, in time and space. This would allow for the reducibility of properties realised in highly dynamic microstructures.[50]

Where no bridge laws are involved, the problem of the availability, or the lack thereof, of bridge laws linking special science kinds with physical kinds cannot arise. On the functional model of reductionism, the desideratum that reductions be explanatory is served, without involving any expansion in the physical ontology, either in terms of entities or laws. In a sense, the functional model of reductionism can be seen as the reverse of traditional bridge law reductionism: it is a reductionism about the explicability of properties and of causal transactions, rather than about identities, with identity, and not nomic covariation, at its heart. Happily, this reversal involves a dissolution of some of the negative connotations associated with traditional reductionism.

If a Nagelian bridge law connects having a mental state M with having a physical state P, we might feel justifiably miffed when we are told that there is nothing 'over and above' our having M than our having P. We feel 'reduced' in the sense of 'lessened' — we feel that a property we had taken ourselves to have, M, is not one that in fact we possess. Instead, we are left only with P. But on Kim's account, our having M simply *is* our having P. To be M is to be P. It *is* P-ness that *is* M-ness. *And* — and this is an important 'and' — the microstructure of M that is P is a real, ontologically unproblematic aggregate with its own real, ontologically unproblematic aggregational causal potential. Nothing is P, *really,* that is not M, *really*. If we are told that we are M *because* we are P,we do not feel lessened. Instead, we feel grounded.

Another difficulty that we have seen to face reductive physicalism arises from the question of the explanatory autonomy of the special sciences. The arguments from the multiple realisability of some special science kinds indicates that the sciences subsuming those kinds are irreducible to physical sciences. It certainly seems to be the case that meaningful explanations can indeed be given in the special sciences, and that many of the special sciences involve genuinely meaningful, albeit hedged, generalisations, even when the kinds involved in the expression of such generalisations are realised in wildly heterogeneous physical bases. But if the special sciences do indeed 'float free' of physics, it is difficult to see how they can be genuinely explanatory. It is not immediately evident that traditional reductionism can allow the special sciences to be both autonomous and genuinely explanatory.

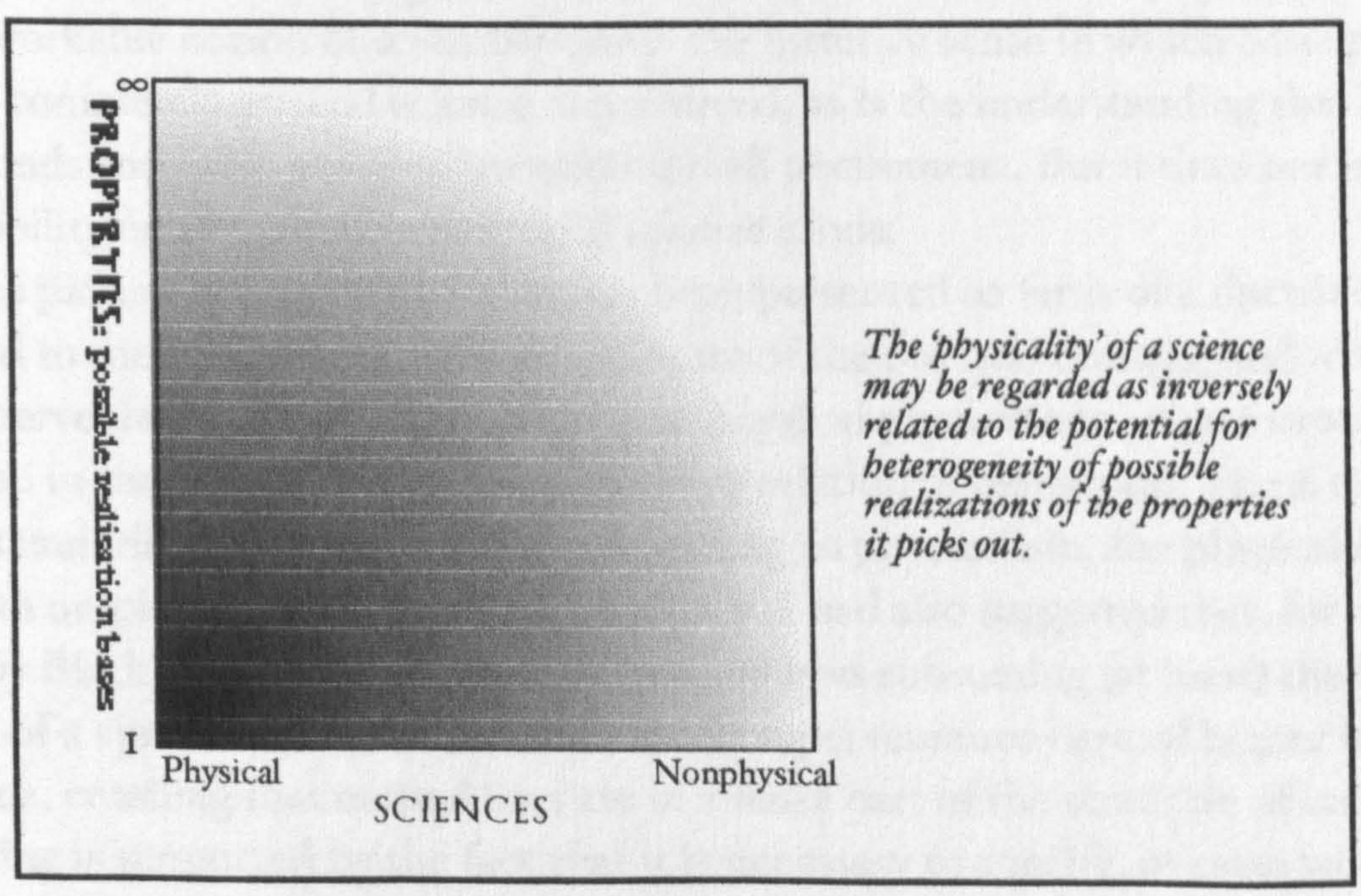
In addition to the apparent wide-ranging success of a great many of the special sciences, such as psychology, economics, evolutionary biology, meteorology, and a host of others, in spite of their irreducibility to basic physics, there is a powerful aesthetic appeal to the idea that the special sciences are autonomous with regard to the physical domain. The notion that the special sciences, most importantly psychology, can be liberated from the 'nomological net'[¤] of of deterministic physics is a very attractive one, appealing to our sense of the freedom and dignity of our particular variety of being. We hope to see the normatives of human action to be of a different order than the laws of ballistics. Traditional reductionism, it appears, may not allow us consistently to do this. Functional reductionism, however, entails that the special sciences are autonomous with regard to basic physics.

According to functional reductionism, some entity x has a mental state, M, because it has a microphysical state P that realises M in x. Some other thing, y, might also have M, because it has some microphysical structure, P2, which meets the same functional specification in y as P does in x. In the abstract sense of possessing a microstructure meeting the same functional specification, x and y are both M. But in another sense, y is *not* M, because it is not P. It is only at the level of abstraction where the causal potentials of functional organisations are construed as intrinsic properties of complex wholes that kind terms in the special sciences can be taken as referring to the same thing when they refer to a property ascribed to different individuals with diverging physical descriptions, and so only at this level of abstraction that special science laws have any coherency. Because they involve structures with the same overall functional descriptions and causal potentials, they can be taken as being genuinely predictive, and meaningful generalisations can be constructed in terms of higher level properties. But below this level of abstraction, special science generalisations are incoherent. We cannot expect laws linking causes and effects given in terms of psychology directly to be echoed in physical laws any more than we can expect ballistics to be reflected in the stochastic laws of quantum physics. But what we can expect, and what we do in fact find, is that the statistical laws of quantum physics back or ground the laws of ballistics. In the same way, we can expect the laws of physics to back or ground the laws of psychology, not in the sense that each psychological law has an expression in physical terms, but that each instantiation of a psychological law is backed by physical laws.

Functional reductionism is not a reductionism about sciences. According to this model, sciences do not necessarily reduce to other sciences about entities lower in the explanatory hierarchy. It is *properties* — and individually occurrent, dated properties at that, rather than kinds of properties — that are reducible to physical properties. The explanatory autonomy of the special sciences is a direct outcome of the functional model of reductionism, and it is not limited to the 'higher' special sciences of the wildly multiply realisable, such as psychology or economics. All of those sciences which have to do with properties susceptible to functional construal, including among them some of the 'paradigmatically physical sciences' such as mechanics and chemistry, are in a sense explanatorily autonomous with regard to basic physics. Some of these sciences, especially those we are inclined to think of as paradigmatically physical, will be closer to basic

physics, in virtue of the contingent fact that only a few, or perhaps only one, basic physical microstructure is sufficient to be the organisation behind the aggregational causal potential of their proprietary entities in this world: only a proton and an electron in a certain relation is hydrogen, and nothing else is.[52]

We can then think of the paradigmatically physical sciences as those picking out entities with fairly unique realisation bases, or, as in the case of classical mechanics, those dealing with such properties, like mass, velocity &c., which can be treated identically across realisation bases for the description of individual events. Thus the problem of how to identify which sciences are physical sciences as over and against nonphysical sciences or not-quite-physical sciences, is dissolved: strictly speaking, one science is physical (basic physics) or none of them are (if there can be no basic physics).[53] Another way of reading this would be to take it that the lines dividing the physical sciences from the nonphysical sciences are fuzzy, and they are exactly as fuzzy as the greatness of the potential (actual or possible within the actual world) for heterogeneous realisation of the properties picked out by sciences other than basic physics. We might translate this notion to fit the description of physical properties: a property is completely physical if it is a basic physical property, and the closer a property is to basic physical properties in terms of the possibility of the multiple instantiability of the property, the more 'physical' it is.[54]



What then of the dream of the unity of science? If the very formulation of reduction does not admit of the reduction of the special sciences to physics, is this a misplaced ideal? Perhaps not. As earlier noted, Jeffrey Poland has distinguished between the notions of 'unitary' science and 'unified' science. The former is a notion according to which physics is completely explanatorily sufficient, so that explanations made in terms of the special sciences can be replaced without loss of content by explanations in basic physical terms. "What is essential to this view is that science does not consist of any

branches other than physics; the special sciences serve only heuristic purposes that in principle could be served by an ideally completed physics."[55] According to the notion of unified science, by contrast, "unification consists, not in the eliminability of all branches in favour of one, but in there being one branch to which all others are related in a particular way. Such a view does not preclude the possibility that a given branch might be eliminable, but what the view affirms is that such elimination is neither the general case nor the goal of scientific activity."[56] The relation Poland sees as appropriate, which he terms 'vertical explanation', is essentially one of property explanation: "Phenomena and regularities at any given level are explainable by appeal to other phenomena and regularities at the same or lower level (not necessarily physics), but *all such appeals* are ultimately grounded in phenomena and regularities at the physical level. I shall say that a phenomenon or regularity is explanatorily grounded in the physical basis where there is a chain of 'vertical explanations'[57] that eventually links it to the phenomena and regularities in physics."[58] The notion of 'unified' science is one of grounded science, with explanations in terms of more basic sciences grounding or backing explanations given in terms of the special sciences. This does not entail that special science generalisations must have precise echoes in physical law, or that the there be identifiable coextensions in basic physical terms for the kinds picked out in the special sciences, but only that for any particular explanation given in terms of the special sciences, a chain of explanations in lower level terms eventually grounds that one explanation in basic physics. This is a perfectly workable notion of scientific unity: the intuitive sense in which basic physics must be a completely general science is preserved, as is the understanding that basic physical kinds and basic physical law underlies all phenomena. But it does not involve the reducibility or eliminability of special science kinds.

The picture of physicalism that has been presented so far is of a doctrine committed to the principle of the causal closure of the physical domain, and to the strong supervenience of all phenomena upon physical phenomena, where supervenience is grounded in the metaphysical and explanatory relation of realisation. These two principles underlie the senses in which, according to physicalism, the physical domain enjoys both ontological and explanatory primacy. I had also suggested that, for reasons exposed by Blackburn and Rowlands, the natural laws subsuming (at least) the basic dynamics of a system should be included in the supervenience base of higher level phenomena, entailing that natural laws are in a sense part of the structure of individuals. This reading is supported by the fact that it is necessary to specify, in cases where the second modal operator in the expression of strong supervenience is read as nomological necessity, that individuals possessing identical microstructures will be alike in all of the properties realised by these microstructures so long as the physical laws are the same for both of them. Counting the physical laws as part of the microphysical structure of these individuals merely obviates the need for this specification: if the basic laws are not the same for two individuals, they are not physically identical. We have seen how the functional model of reductionism provides us with a way of construing the relative physicality of properties, and of the physical sciences. But as yet, no criteria are on the offing by which physical properties *per se* can be identified as physical. Beyond the

principles of strong supervenience grounded in the metaphysical and explanatory relation of realisation and the principle of the causal closure of the physical domain, what more is needed for an adequate expression of the physicalist position?

I would like to suggest that nothing more is needed. As Daly has argued and as discussed in the previous section, there are serious difficulties involved with the attempt to set *a priori* constraints upon physical properties. For Daly, this indicates that all metaphysical programmes that assume that there is a distinction between physical properties and nonphysical properties are ill-founded, better abandoned in favour of questions that would arise independently of the assumption. I believe, rather, that the difficulty indicates that it is unreasonable to expect physicalism to provide *a priori* criteria for what is to count as a physical property. The doctrine of physicalism as I have so far presented it, encapsulated in the dual principles of strong supervenience grounded in a metaphysical explanatory relation and the principle of the causal closure of the physical domain, is a commitment to the ontological and explanatory dependence of all actual or possible properties upon physical properties and the dynamics subsuming them, and this sets strong constraints upon what can count as a physical property in a system that contains any properties that are physical. I suggest that according to this notion of physicalism, what it is to be a physical property can be explained purely by reference to to the scheme of metaphysical and explanatory realisations taken as causally closed: to be physical is just to fit into a closed causal/explanatory hierarchy wherein all properties, relations, dynamics and facts are underwritten by basic properties, relations, dynamics and facts.

According to this identification of what it is to be a physical property, 'physicality' is a world-relative property things have by virtue of their embedment within the kind of hierarchy described. This has the possibly counterintuitive connotation that 'physical' worlds are possible which, because they have very different sorts of properties subsumed by very different natural laws than those found in the actual world, are possible. That is, an alternative possible world containing only ecto-properties, which are not physical in this world, would be a physical world if ecto-properties in the ecto-world world are subsumed by ecto-laws, and if properties and laws in the ecto-world were organised into a closed causal/explanatory hierarchy such that higher level ecto-properties are realised by lower level ecto-properties and all higher level ecto-dynamics are describable in terms of lower level ecto-dynamics. So long as all phenomena in the ecto-world are functionally reducible in this way, the ecto-world is a physical world, though it contains no properties which could be classified as physical in the actual world. Further, properties which are physical in the actual world might[59] not be physical in the ecto-world, because they might not be properties that fit into its closed causal/explanatory hierarchy.

This is, to my mind, an acceptable proposition. The notion of possible physical ecto-worlds might not be so troubling when it is considered that while the properties and laws of such worlds do not resemble any properties and laws with which we are familiar, such ecto-worlds whose properties were organised in a closed causal/ explanatory hierarchy would be just like the actual world in these significant ways: higher level properties would strongly supervene upon lower level properties. A metaphysical

explanatory relation would obtain between the realisers of higher level ecto-properties and the higher level properties themselves, grounding the stroong supervenience relation, such that explanations of higher level ecto-properties would be possible in terms of lower level ecto-properties. Ecto-sciences would be possible. If an ecto-world had properties and laws that could support the existence of complex microstructures that could realize minds, and if some of these mindful structures took to the study of the general underlying features of their world, they would be doing what physicists do in the actual world, and I can see no reason to withhold the designation from these ecto-scientists, provided that their science is restricted to their world.

An objection might run as follows: suppose there is an ecto-world that is a perfect shadow of our world. All of its properties and laws exactly resemble the properties and laws of the actual world, and it has molecule-for-ecto-molecule identical duplicates of every existing thing in this world. Suppose that we transpose a pair of duplicates between worlds: I go to the ecto-world, and my ecto-twin appears in this world. Because we have the same functional microstructure, no difference should be evident, yet because the ecto-Nightshade does not fit within the closed causal/explanatory hierarchy of the actual world, and I do not fit in that of the ecto-world, neither of us in fact has any physical effect upon the worlds in which we find ourselves. We vanish.

This objection is, however, incoherent. If what it is to be a physical property is solely described in terms of the property's fitting within a closed causal/explanatory hierarchy, then the above described scenario is impossible, for the *only* thing that makes an ecto-property a nonphysical property in the actual world is that it does not fit into the closed causal/explanatory hierarchy that describes this world. The ecto-world, being a property-for-property and law-for-law shadow of our world, is not in fact an alternative possible world just like this world. By Leibniz's Law, it *is* this world.

The characterisation of the physical in terms of a closed causal/explanatory heirarchy avoids the difficulties earlier discussed for the formulation of physicalism. It sets no *a priori* constraints on what can count as a physical property, but it does place some principled restrictions on physicality: if any phenomenon is not in principle explicable in terms of the closed causal/explanatory hierarchy the base set of properties for which is physical, if it is not functionally reducible to physical properties and dynamics, then that phenomenon is not physical. As we have seen, the functional model of reductionism is not confronted with the difficulties traditional reductionism faces in terms of the multiple instantiability of many special science properties. Thus, the irreducibility of special science kinds to physical kinds and the irreducibility of the special sciences of the multiply instantiable to physical sciences is no obstacle to the functional reduction of particular instances of multiply instantiable higher level properties to physical properties. On this notion of physicalism, *kinds* can be nonphysical, while no instantiations of physical kinds can be.

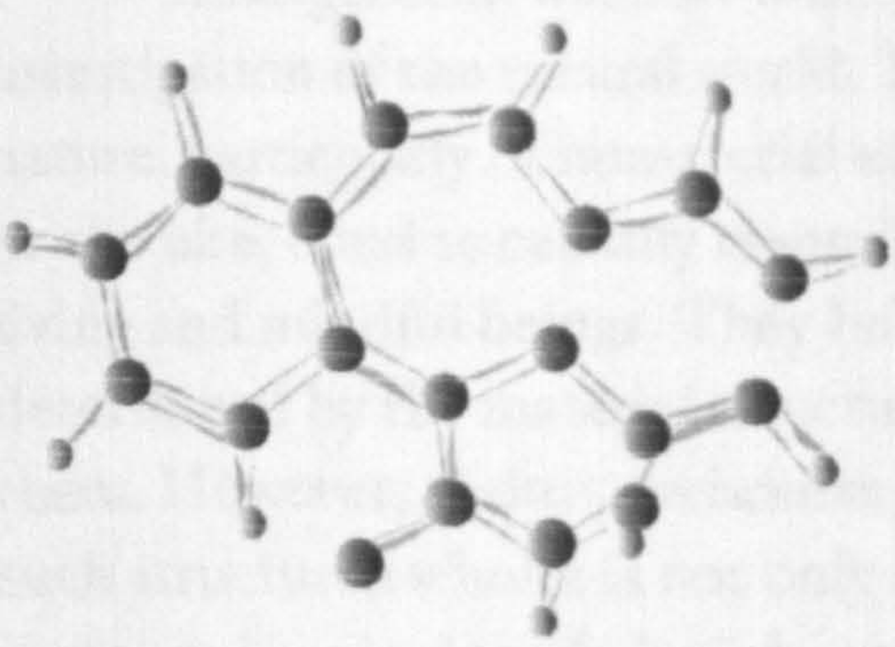
The constraint that all nonbasic physical properties be functionally reducible to basic physical properties is indifferent to existing physics, or to any possible physics we might care to describe. It therefore allows for the continuing development of physics, and to the expansion of the ontology that physics recognises. On this account of

physicalism, in the absence of a complete, correct physics, we cannot know what all of the physical properties are, nor can we in principle exclude actual properties from the physical ontology when we do not know what future physics is likely to be able to accommodate. So much is entailed by the requirement that physicalism be formulated such that future developments in physics do not render the doctrine false. Thus, physicalism, according to this characterisation, will be impossible to falsify from *a priori* argument: the truth or falsity of physicalism will only be evident when basic physics is fully developed and a full inventory of all of the properties the world contains is in. But like the principles which I have argued underlie robust physicalism, the supervenience of all phenomena upon physical phenomena grounded in a metaphysical explanatory relation and the principle of the causal closure of the physical domain, I do not believe that the truth of physicalism can be *proven*. That it be provable or falsifiable in principle, however, is an unreasonable requirement to set upon a doctrine about everything. That we do or do not have good reasons to place some faith in the truth of the doctrine, given what we actually know about the world, is enough.

4

Downward Causation

4. Nonreductive Physicalism and Emergent Evolution



IN THE WAKE OF THE COMPELLING ARGUMENTS that have been levelled against the traditional model of reductionism, which is focused around the notion of the reduction of the special sciences to the physical sciences via Nagelian 'bridge laws' linking special science kinds to physical kinds, nonreductive physicalism has taken centre stage in modern philosophy of mind. The version of nonreductive physicalism with which we are here concerned has as its motivation the multiple instantiability of the mental, among other special science kinds: if mental states are essentially functional states that can be realised in heterogeneous physical bases, then *qua* mental, they cannot be reduced to physical kinds or correlated with physical kinds via bridge laws.[1] If this is the case, then the reduction of psychology, and other special sciences that deal with multiply instantiable properties, to physics cannot go through. We have seen how the functional model of reductionism, centred around the notion of the functional construal of higher level properties and the identification of particular instantiations of higher level properties with whatever microstructure performs the specified function, does not encounter the difficulties faced by the traditional model of reductionism, and entails that the special sciences are irreducible to physics. The availability of such a model itself might motivate a reconsideration of reductionism about the relation of mental to physical properties as a viable option for the physicalist. I want further to argue that if nonreductive physicalism takes mental properties to be real, causally effective properties *qua* mental, and if it takes psychological laws to be real, albeit hedged, laws genuinely descriptive of mental dynamics, it is essentially incompatible with robust physicalism. This is because such a construal of mental properties and generalisations is directly at odds with the principle of the causal closure of the physical domain, which principle I have argued underlies robust physicalism.

In his (1992) "Downward Causation and Emergence," Jaegwon Kim has argued

that, in essence, nonreductive physicalism is a form of the doctrine of *emergentism* or *emergent evolution.* The version of emergentism with which Kim is concerned belongs to the influential body of work Brian McLaughlin (1992) has termed British Emergentism,[2] which has its origin in Mill's (1843) *System of Logic,* reaching its mature expression in Broad's (1925) *Mind and Its Place in Nature,* the last major work in the emergentist tradition. In his analysis Kim is not concerned with the differences in the expression of the doctrine of emergentism between the various writers on emergentism, but rather with the common core of the doctrine: the emergentists' notion of the relation between properties they regarded as 'emergent' and th eproperties and relations from which emergent properties were supposed to emerge, and in the causal potency assigned to emergents.[3] He argues that modern nonreductive physicalism, insofar as it is committed to a realistic construal of mental properties, is as committed to the notion of *downward causation,* essentially the notion that higher level properties can be causally effective in virtue of their natures *qua* higher level, as is emergentism.

Emergentism was forwarded as a naturalist doctrine in keeping with the empirical investigation of the natural world. The emergentists rejected talk of mysterious factors in nature, particularly of immaterial entities, such as entelechies, *élan vital*, Cartesian souls, or the like, cited as causally responsible for the presence of characteristic properties of living and mindful beings. They held that these characterising properties were fully determined by the material structure of the living and mental entities that exhibited them. However, against *mechanism,* according to which the characterising properties of such structured wholes is not only determined by the properties and organisation of the parts are in principle deducible *a priori* from a complete knowledge of the parts and their relations, the emergentists held that the whole is, quite literally, more than the sum of its parts: certain configurations of parts, they held, are sufficient for the emergence of genuinely novel properties, unpredictable from the fullest knowledge of the properties and dynamics of the parts i*n abstracta* or in other combinations.

Kim isolates the central thesis of emergentism as encapsulated by three principles: that of an Ultimate Physicalist Ontology; that of Property Emergence, and that of the Irreducibility of Emergents, outlined as follows:

> 1) **The Ultimate Physicalist Ontology:** There are basic, nonemergent entities and properties, and these are material entities and their fundamental physical properties.[4]
>
> 2) **Property Emergence:** When aggregates of basic entities attain a certain level of structural complexity (relatedness), genuinely novel properties emerge to characterize those structured aggregates. Moreover, these emergent properties emerge *only* when the appropriate basal conditions are met.
>
> 3) **The Irreducibility of Emergents:** Emergent properties are novel in that they are not reductively explainable in terms of the conditions out of which they emerge.

We can see immediately that there is a close parallel between the first two of these principles and principles generally accepted by nonreductive physicalists. It seems

commonly to be held, if not stated outright, that there is some basic level of the world's ontology at which nothing is decomposable into constituent parts, and this is generally taken to be the basic physical domain. The second principle can be read as an expression of supervenience, correlating the genuinely novel properties of certain aggregates with the structures of these aggregates: if anything has a certain aggregational structure G, then it has some characteristic property F, and if anything else has G, then it too has F. The qualification that emergent properties emerge only when their basal conditions are met is an expression of the dependency of emergents upon their emergence bases: emergent properties never float free, but are found always and only where appropriate structural conditions obtain.

There are some important clarifying remarks to be made about (3). The idea of reduction with which the emergentists were working was not the model of intertheoretic reduction common in modern literature. Rather, their notion of reduction generally corresponds more closely to property reduction. Broad speaks of the *deduction* of characteristic properties of aggregates from knowledge of the parts in isolation or in other aggregates, and terms the theory that all properties of composed objects are in principle deducible from a complete knowledge of the properties and relations of the parts of those objects *mechanism*. Illuminating the distinction between mechanism and emergentism, Broad writes:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). The mechanistic theory rejects the last clause of this assertion.[1]

Thus an emergent property is a property of a mereological whole which, whilst its obtaining is lawlike in that all mereological wholes with the same parts and having the same structure will have that property, is not explicable in terms of the properties of the parts and the laws governing the dynamics of the parts in that structure or in any other. Emergent properties, and the laws correlating them with the structured aggregates in which they are found, are ontologically *unique and ultimate*.

The notion that higher level properties of structured wholes can on the mechanistic view be *deduced,* or *predicted,* from an ideal knowledge of the properties of the parts and of the laws subsuming the dynamics of the parts in abstracta and in other combinations is the important defining characteristic of mechanism for the emergentists. However, I think that the notion of the *deduction* of higher level properties from knowledge of lower level properties is a stronger notion than is strictly required by emergentist doctrine, and the strength of the notion of deduction may cloud the idea they wish to express for those coming to their works from a more modern perspective. We are now aware that there are some systems whose behaviour, whilst being completely and wholly determined by the properties and dynamics of the components of the system, are nonetheless genuinely unpredictable. Such systems, such as, notoriously,

weather patterns, among other phenomena, are governed by chaos dynamics, an only comparatively recently developed branch of mathematics. The emergentists did not have chaos among the scientific and mathematical notions upon which they could draw, but it seems to me that if they were to have had it, they would agree that the practical or thoretical deducibility of the properties or behaviour of a system in terms of the properties and behaviour of its parts is too strong. I suggest that a notion that would be agreeable to them is that of explanation. The behaviour of chaotic systems is not predictable from a complete knowledge of the parts and their dynamics,[6] but it is so explicable. I will speak sometimes of the explanation of higher level properties in terms of lower level properties, occasionally making reference to deducibility when it appears in the emergentist text, with that caveat in mind.

It is not the case for the modern nonreductive physicalist that realised properties are irreducible to physical properties in the sense that they are inexplicable in terms of the physical systems realising them. Rather, it is generally taken that particular instances of realised properties may be fully explainable in terms of the functional organisation of the particular microstructure realising the property at that time.[7] It is, rather, *kinds* of properties which are not reducible to physical structures. In this respect, nonreductive physicalism and emergentism are *prima facie* not equivalent. But though these two claims of the nonreducibility of realised or emergent properties are different in nature, they both have the consequence that theories couched in lower level terms will be inadequate for the explanation of higher level phenomena. As Kim says,

> What is common to all these antireductionisms is the belief that phenomena and regularities at one level are *explanatorily unreachable* from another level, that the distinctively and characteristically psychological phenomena, properties, and regularities constitute *an autonomous domain,* from the point of view of scientific theory, relative to the domain of physical and biological phenomena. And this is in spite of the fact that, from an ontological point of view, mentality is dependent on physical and biological processes.[8]

So, while nonreductive physicalism and emergentism have different reasons to assert the nonreducibility of certain higher level properties and have different notions of in what that nonreducibility consists, we do see a remarkable convergence[9] between nonreductive physicalism and emergentism: both advert to a basic physical ontology underlying all phenomena, both adhere to a supervenience thesis, and both have the consequence, through different notions of irreducibility, that the laws of basic physics or of any of the lower level sciences are inadequate for the explanation of higher level phenomena that involve multiply instantiable, for the one, or emergent, for the other, higher level properties.

4.1. Realism

To construe mental properties *realistically* is to construe them as genuine causal agents in themselves and in their own right as mental properties. That is, when a mental event is cited as relevant in any causal interaction, on the realistic construal of mentality it

is the properties which are characteristic of the mental event *qua* mental that are relevant to the explanation. Suppose, for example, I tidy the kitchen. I pick up pots and pans from the stove and put them in the sink; heat water; scrub pots and pans; clean the countertops; sweep the floor; in short, by wilfully moving my body about in various ways, I cause all kinds of physical things (not just my body) to be moved and rearranged. A very reasonable explanation for this flurry of physical movings-around is that I *desired* that the kitchen should be tidy, and that I *believed* that by performing these various physical acts, the kitchen would become tidy. My actions, and the resultant state of the kitchen, is explained by reference to contentful mental states.

Kim argues that this kind of explanation, taken at face value, commits the nonreductive physicalist to the notion of downward causation. But downward causation proves highly problematic for the physicalist who is committed to the principle of the causal closure of the physical domain, which principle I have argued is essential to the physicalist doctrine. If Kim is right, then the conclusion that nonreductivism is incompatible with a formulation of physicalism that is committed to the causal closure principle is supported.

Kim defines downward causation as the idea that "*higher level mental events and processes cause lower level physical laws to be violated*, that the molecules that are part of your body behave, at least sometimes, in ways different from the way they would behave if they weren't part of a body animated by mental processes."[10] If any piece of physical behaviour is explicable *only* in terms of a mental cause — if, as the quotation says, the physical parts of my body as I tidy the kitchen behave in ways that they would not or could not in the absence of a mind, in that the mental state has to be invoked to explain the physical permutations my body undergoes — no complete physical description could be given for that piece of physical behaviour. The mental states attributed to me are essential elements in the complete explanation of this physical activity. This, however, is to violate the principle of the causal closure of the physical domain, which asserts that in tracing the causal history of any event falling under physical description, we need never move outside the physical domain to cite nonphysical causes, and that each physical event has a complete sufficient physical cause. When the emergentist maintains that certain structured aggregates are enabled to behave in ways that the parts of the aggregate could not *in abstracta* or in other combinations, and that they are so enabled in virtue of a genuinely novel property that cannot be deduced from or explained in terms of the properties and relations of the aggregate's parts, they assert that causal interactions involving these aggregates involve the participation of a factor for which no complete physical description exists, but which must be invoked on its own level, in its own terms. The emergentist is thus committed to downward causation.

Kim's contention is that the nonreductive physicalist who construes the mental realistically is as committed to the occurrence of downward causation as is the emergentist. His elegant argument is as follows:

> At this point we know that, on emergentism, mental properties must have causal powers. Now, these powers must manifest themselves by causing either physical properties or other mental properties. If the former, that already is downward causation. Assume then that mental property M

causes another mental property M*. I shall show that this is possible only if M causes some physical property. Notice first that M is an emergent; this means that M* is instantiated on a given occasion only because a certain physical property, P*, its emergence base, is instantiated on that occasion. In view of M*s emergent dependence on P*, then, what are we to think of its causal dependence on M? I believe that these two claims concerning why M* is present on this occasion must be reconciled, and that the only viable way of accomplishing it is to suppose that M caused M* by causing its emergence base P*. In general, the principle involved is this: *the only way to cause an emergent property to be instantiated is to cause its emergence base property to be instantiated.* And this means that same level causation of an emergent property presupposes the downward causation of its emergence base. That briefly is why emergentism is committed to downward causation. I believe that the argument remains plausible when emergence is replaced by physical realization in appropriate places."

I should like to return to this argument later. At this point it will be interesting to look more closely at the doctrine of emergentism itself, so as better to appreciate its subtleties as well as its similarities and differences in respect of nonreductive physicalism. In what follows I shall focus on C. D. Broad's Mind and Its Place in Nature, as Broad's is a particularly clear expression of developed emergentist thinking.

4.2. British Emergentism

The context of Broad's work, and of emergentism generally, is the debate between vitalism, which asserts that the characteristically lifelike properties of living entities could not be explained purely in terms of processes of the physical body, but that appeal has to be made to a novel and irreducible *component* present in the living body: an entelechy or *élan vital* to explain the liveliness of the living. Mechanism, on the other hand, holds that all living processes, and all other processes as well, are fully explicable in terms of the properties and laws which describe the dynamics of basic non-living systems. The emergentists rejected appeal to a mysterious, generally nonphysical component taken to be present in aggregates exhibiting the property *being alive,* while acknowledging apparently inexplicable properties of aggregates, and were concerned to account for these properties in a naturalistically acceptable way.

One of the factors motivating the emergentists was the apparently inexplicable nature of the properties of chemical compounds in relation to the properties their component elements could be observed to exhibit in isolation, and the then inscrutable laws of chemical bonding. Though at the the time that he was writing the Tarner lectures, Broad was aware that atoms were not understood to be fundamental aspects of nature, but that they were possessed of a certain microstructure of protons and electrons (neutrons, as McLaughlin notes, were not discovered until 1932[12]), the theory that could explain chemical bonding — quantum mechanics — had not been developed. Though Broad accepts that a future science might show chemical as well as vital phenomena to be fully mechanistically explicable, he had no way to anticipate the sciences that would do so.[13] For all he knew or could expect, the facts having to do with chemical bonding had simply to be taken as brute. Thus for Broad, as for the British Emergentists generally, the scope of the problem was wider that the debate between vitalism and mechanism. For Broad, the question is: "Are the differences between merely physical,

chemical, and vital behaviour ultimate and irreducible or not? And are the differences in chemical behaviour between Oxygen and Hydrogen, or the differences in vital behaviour between trees and oysters and cats, ultimate and irreducible or not?"[14] Or: are there unique and ultimate differences between certain aggregates and their components, or between individuals at higher and at lower levels, or are there not?

To illuminate the dilemma, Broad gives the contrasting, now classic, examples of a clock and a chemical compound. Even if a person had never seen a clock, if he is knowledgeable about the general rules describing mechanical behaviour, which he could learn through the observation of other mechanical systems, and if he is told how a clock is constructed, he could predict how it would behave. At the time Broad was writing this was not even theoretically possible for the case of chemical properties. Broad writes:

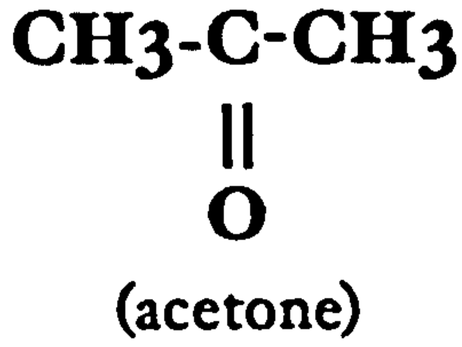
> Oxygen has certain properties and Hydrogen has certain other properties. They combine to form water, and the proportions in which they do this are fixed. Nothing that we know about Oxygen by itself or in its combinations with anything but Hydrogen would give us the least reason to suppose that it would combine with Hydrogen at all. Nothing that we know about Hydrogen by itself or in its combinations with anything but Oxygen would give us the least reason to expect that it would combine with Oxygen at all. And most of the chemical and physical properties of water have no known connexion, either quantitative or qualitative, with those of Oxygen and Hydrogen. Here we have a clear instance of a case where, so far as we can tell, the properties of a whole composed of two constituents could not have been predicted from a knowledge of the properties of these constituents taken separately, or from this combined with a knowledge of the properties of other wholes which contain these constituents."[15]

The doctrine of property emergence was forwarded precisely to accommodate this apparent gap in our understanding naturalistically, without invoking mysterious entities or powers.

Having dismissed (rightly, I think) as logically possible, but empirically unrealistic and philosophically unsatisfying theories of property dependence according to which the characteristic behaviours of mereological wholes are not in any way dependent upon their microstructural organisation,[16] Broad distinguishes two types of the complementary theory, according to which the structures of aggregates are key factors in determining which characteristic properties these structures will exhibit: *component theories,* which hold that the characteristic behaviour of a certain object or class of objects is in part dependent on the presence of a peculiar component which does not occur in anything that does not behave in this way,[17] and those to which I shall refer as *structure theories.* This kind of theory denies that there need be any peculiar component which is present in all things that behave in a certain way and is absent from all things which do not behave in this way. Rather, such theories hold that the components may be exactly alike in both cases, and they explain the difference of behaviour wholly in terms of difference of structure.[18]

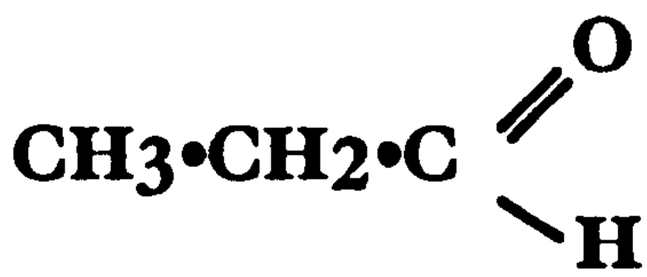
Component theories, Broad notes, are the kind standardly employed in chemistry: silver chloride has certain properties, and common salt has others, because in addition to chlorine one compound contains silver, and the other sodium. Differences in the structural arrangement of chemical compounds are also cited as explaining the different properties they may exhibit. Thus, the reason that the two chemical compounds acetone and propion aldehyde have such very different properties though

they are composed of the same chemical elements is ascribed to the fact that they have the different structures symbolised by



(acetone)

and



(propion aldehyde)[19]

This blend of structure and component theories will be appropriate if the properties of the chemical elements are brute, irreducible properties. This has turned out not to be the case for the chemical elements, but in any case, it should be noted that whatever the state of future scientific practice, whatever the properties of whatever is taken to be the most basic and fundamental element of nature will have to be taken as brute, not further decomposable. It is reasonable to expect that any fundamental theory will take the form of a blend of structure and component theories.

The doctrine Broad identifies as *Substantial Vitalism,* which holds that a necessary factor in explaining the characteristic behaviour of living bodies is the presence in them of a peculiar component, often called an "Entelechy", which does not occur in inorganic matter or in bodies which were formerly alive but have now died,[20] has the form of a component theory about mereological wholes. Given that we do accept component theories on the basic levels of explanation, why should we reject them for aggregates?

The reason why we feel better about accepting ultimate and irreducible properties of chemical components in explanations about why certain chemical compounds have certain properties, and not so good about accepting entelechies or *élan vital* in explanations about why certain mereological wholes exhibit lifelike properties has to do with the nature of the proposed entelechy itself: a purely hypothetical entity of a typically immaterial nature. Some chemical elements, Broad notes, have been hypothesised and used in chemical explanations long before they ever actually were isolated. But prior to that, other chemical elements *had* been isolated, with greater or lesser degrees of difficulty. There is no reason to suppose that hypothetical chemical elements *could not* be isolated, or if the degree of difficulty involved in the isolation should be too great, that they could not be isolable in theory. No entelechy, however, has ever been isolated, and it is difficult to see, given the supposed immaterial nature of the thing, how such a feat could even be accomplished. An entelechy is a *purely* hypothetical entity in a way that

hypothesised but as yet unisolated chemical elements are not. To the objection that an entelechy is in principle not something that could empirically be detected, Broad perspicuously answers

> If it be said that an isolated entelechy is from the nature of the case something which could not be perceived, and that this objection is therefore unreasonable, I can only answer (as I should to the similar assertion that the physical phenomena of mediumship can happen only in darkness and in the presence of sympathetic spectators) that it may well be true but is certainly very unfortunate.[21]

Another objection might be advanced to the effect that some chemical elements cannot exist in isolation, and that therefore the objection that entelechies cannot be isolated is inappropriate. Broad responds:

> It is true that some groups which cannot exist in isolation play a most important part in chemical explanations. But they are *groups* of known composition, not mysterious simple entities; and their inability to exist by themselves is not an isolated fact but is part of the more general, though imperfectly understood, fact of valency. Moreover, we can at least pass these groups from one compound to another, and can note how the chemical properties change as one compound loses such a group and another gains it. There is no known analogy to this with entelechies. You cannot pass an entelechy from a living man into a corpse and note that the former ceases and the latter begins to behave vitally.[22]

Further, since an entelechy is not a material entity, and so is not necessarily in space, it is very difficult to make sense of the notion that a living body is a compound of physical stuff and an entelechy. Thus Broad rejects substantial vitalism, acknowledging it to be logically possible, but certainly less than appealing from a naturalistic standpoint. Since his arguments are general, at the same time he rejects any component theory about mereological wholes.

Broad identifies two types of structure theories: mechanistic and emergent. Within this distinction two versions of mechanistic theories are distinguished: *pure mechanism,* a very strong reductive mechanism with an eliminativist character, and plain mechanism, a weaker version. According to the ideal of pure mechanism, there is only one fundamental kind of particle, and only one law of composition governing the possibilities for its combinations. All mereological entities and all of the sciences that describe them would be reducible to descriptions in terms of the fundamental stuff and its fundamental law. All the apparently different kinds of stuff, under this ideal, are just differently arranged groups of different numbers of the one kind of elementary particle; and all the apparently peculiar laws of behaviour are simply special cases which could be deduced in theory from the structure of the whole under consideration, the one elementary law of behaviour for isolated particles, and the one universal law of composition. On such a view the external world has the greatest amount of unity which is conceivable. "There is really only one science, and the various "special sciences" are just particular cases of it."[23] This kind of mechanism offers the highest degree of tidiness possible, but it is not necessary for the mechanist to take so strong a stance. A mechanist could admit that there were different types of element and a group of laws describing their combinations and dynamics, as the electronic theory of matter at the time did.[24]

One could also be a mechanist about particular levels of explanation: *E.g.*, if a biologist held that all the characteristic behaviour of living beings could be deduced from an adequate knowledge of the physical and chemical laws which its components would obey in isolation or in non-living complexes, he would be called a "Biological Mechanist" even though he believed that the different chemical elements are ultimately different kinds of stuff and that the laws of chemical composition are not of the type demanded by Pure Mechanism.[25]

As we have seen in Kim's description of the fundamental tenets of emergentism, the crucial difference between emergence theories and mechanistic theories lies in the manner in which the properties of structured aggregates are held to be related to the structures of the aggregates themselves. Mechanistic theories and emergent theories differ, says Broad, "according to the view that we take about the laws which connect the properties of the components with the characteristic behaviour of the complex wholes which they make up".[26] According to mechanism, "the characteristic behaviour of the whole is not only completely *determined* by the nature and arrangement of its components; in addition to this it is held that the behaviour of the whole could, in theory at least, be *deduced* from a sufficient knowledge of how the components behave in isolation or in other wholes of a simpler kind."[27] According to emergentism, on the other hand, "the characteristic behaviour of the whole *could* not, even in theory, be deduced from the most complete knowledge of the behaviour of its components, taken separately or in other combinations, and of their proportions and arrangements in this whole."[28]

Mechanism asserts that the properties and laws that subsume the dynamics of the parts of a particular aggregate are sufficient to determine the characteristic properties of that aggregate. Emergentism denies this part of the mechanist thesis. For the emergentist there is indeed something over and above the sum of the parts of certain structured aggregates. In holding that emergent properties are determined by the structure of the entity in which they inhere, but which are in principle inexplicable and nondeducible in terms of the properties and laws appropriate to the parts, the emergentist introduces two fundamental elements to the world: a unique and ultimate emergent *law,*[29] which has no reflection in lower level laws, and the emergent *property* itself, a new, ultimate and irreducible property.

There are two ways of thinking about the emergent nature of chemical properties and the laws binding them to certain structured aggregates. We can consider it to be among the "properties" (Broads scare quotes) of a particular element, that when mixed with some different element in a certain proportion yields a compound with such and such characteristic properties; another of its "properties" that when mixed with some other element in some proportion, yields a compound with such and such other characteristic properties; and so on. The complementary properties can be assigned to the other elements involved. In this case, we can never know all of the properties of any element until we have observed it in all possible compounds and all possible combinations a potentially impossible proposition, even if merely possible properties of elements generated by their combination with merely possible but nonactual elements with merely possible elemental properties are overlooked.[30] In the other case, we can

take it that the properties of an element are those which it has intrinsically, and which it exhibits when it is not behaving chemically in any way. In this case, we can never know all of the laws having to do with a particular elements forming compounds with certain characteristic properties when in combination with other elements in different structural organisations. Broad takes the second approach, but he notes that these divergent views are the same in practice, and they both have the same consequence: "that the behaviour of an as yet unexamined compound cannot be predicted from a knowledge of the properties of its elements... it matters little whether we ascribe this to the existence of innumerable latent properties in each element, or to the lack of a general principle of composition... by which the behaviour of any chemical compound could be deduced from its structure and from the behaviour of each of its elements in isolation from the rest."[31]

Emergentist doctrine describes the world as a hierarchy of levels of structural complexity, from the purely physical, to the chemical, to the vital, to the psychological. Each level is distinguished by the characteristic properties unique to entities on that level. Such entities are mereological wholes, structured aggregates of lower level components, and the characteristic properties exhibited by these wholes are taken to be fully determined by the structure that exhibits them. However, these characterising properties are irreducible, ultimate properties without echo or explanation on the level of the parts which in that specific organisation determine them. Nonetheless according to emergentist doctrine the relation between the structure of a mereological whole and emergent properties of that whole was more than a of matter brute covariation, but is a nomological determinative relation.

Broad's emergentism distinguishes two different kinds of laws: laws governing the interactions of elements within any particular level of the hierarchy, and laws that bind emergent properties to the mereological structures in which they are found, and which characterize them as higher level in respect of their parts. These are termed Intra-ordinal Laws and Trans-ordinal Laws respectively. Just as emergent properties are unique and ultimate properties of aggregates not explicable from even the mot complete knowledge of the properties of the parts, a trans-ordinal law is a unique and ultimate law. It is "a statement of the irreducible fact that an aggregate composed of aggregates of the next lower order in such and such proportions and arrangements has such and such characteristic and non-deducible properties."[32] Intra-ordinal laws subsume dynamics within any particular level, whether or not there are emergent properties on that level, and in company with the characteristic properties on that level are those to which merely resultant properties and laws on higher levels could conceivably be reduced. Recall the example of the biological mechanist: it is consistent with emergentism that some higher level properties be reducible, albeit not necessarily all the way to the physical level, as will be the case when entities with emergent properties are components of wholes exhibiting merely resultant properties. A school of fish or a flock of birds might be examples: the behaviour of the school or of the flock is reducible to the behaviours of the individual members, but on emergentist doctrine, the vital properties of the individuals themselves are not further reducible.

Intra-ordinal laws proprietary to higher levels are special science laws. As expressed, some of these laws will be reducible to laws of lower level sciences, when those laws have to do only with resultant properties of the individuals on the level in question. When intra-ordinal laws generalise over emergent properties on their own level, reduction will not be possible. Thus, while Broad himself does not characterize them this way,[33] it is fair to take such intra-ordinal laws as generalise over emergent properties as a species of emergent law, and we might call them emergent intra-ordinal laws. What is important to notice here is that since these emergent intra-ordinal laws have to do with the dynamics of unique and ultimate, irreducible emergent properties of structured wholes, the behaviour of the wholes as subsumed by emergent intra-ordinal laws is *unique and ultimate behaviour,* without echo in physical laws. This is downward causation exactly as Kim described it: higher level emergent processes are according to their own laws, causing lower level laws to be violated.

4.3. Downward Causation in Nonreductive Physicalism

We now have enough to reconsider Kim's argument that modern nonreductive physicalism is in essence a species of emergentism, and is committed to downward causation in the same way that emergentism is. If Kim is right, then his is a formidable case against nonreductive physicalism, for as we have seen downward causation is directly at odds with the principle of the causal closure of the physical domain. Remember that Kim's argument runs as follows: emergent mental properties have novel causal powers which must manifest themselves by causing some other properties to come about, either physical properties or other mental properties. If emergent mental properties cause physical properties to come about, this already is downward causation: a property which cannot be given a purely physical description is invoked to explain a physical chain of events. If a mental property causes another mental property to obtain, then that too is downward causation, for "the only way to cause an emergent property is to cause its emergence base property to be instantiated."[34] A particular mental property, *M,* obtains in a system because its physical emergence base property, *P,* obtains in that system. For *M* to cause another mental property, *M*, M* must somehow bring it about that *M*s* physical emergence base property, *P*,* occurs in the system. Thus if *M* causes any event, *M* causes a physical event, which is downward causation. Kim's contention is that this argument can be run just as well with nonreducible property in the place of emergent property.

I believe that while this argument *can* be run for nonreductive physicalism, it cannot be applied to British emergentism, because it is not true according to emergentist doctrine that the only way to cause an emergent property to come about is to cause its emergence base property to come about in the sense that the argument requires. This is because of the very different ways that nonreductive physicalism and emergentism construe the relation of higher level properties to their lower level base properties. As

already mentioned above, according to nonreductive physicalism, it is not individual realised properties that are irreducible to their physical bases, but *kinds* of realised properties that are irreducible in virtue of their multiple instantiability. This is not the case for the emergentist, for whom it is *particular instances* of emergent properties are inexplicable in terms of the base properties with which they are associated via unique and ultimate trans-ordinal laws.

Kinds of multiply realisable properties for the nonreductive physicalist are irreducible to physical kinds, but typically for the nonreductive physicalist, any one specific higher level property will be explicable in terms of its realization in a particular physical base. Higher level properties as nonreductive physicalism describes them are functional properties whose function can be fulfilled by one of a number of possible physical systems, adhering to the physical laws subsuming those physical systems. The realization of such higher level properties is therefore explicable in terms of the properties, relations, and subsuming generalisations of the relevant parts. At the level of the special sciences, where multiply instantiable properties are construed as intrinsic properties, no general physical story can be told about their obtaining in physical systems or about the causal relations into which they may enter. Hence the special sciences of multiply instantiable properties are irreducible to physics, while specific, individually occurrent realised properties are reducible to, and explicable in terms of, physical properties.

The relation that obtains between emergent properties and their emergence bases is quite different from that which obtains between functional, realised properties and the physical bases that realise them as construed by modern nonreductive physicalists, and so is the notion of the actual properties themselves. The nonreductive physicalist typically speaks of mental properties in terms of specific mental *states:* of a certain belief that p, say, as a property of a certain organism. It does not appear that the emergentists in general thought of emergent properties as such individually occurring states, but rather in terms of novel *capacities to behave* emerging from physical aggregates meeting certain structural specifications. For example, Broad speaks of the power of reproduction as an example of an ultimate characteristic of [the vital] order.[35] Structured aggregates with mental properties have "the power of cognizing, the power of being affected by past experiences, the power of association, and so on."[36] These are emergent powers within which individual states, or particular properties, such as a particular cognitive exercise, or a particular memory or association, occur. Once a structure is so constituted as to exhibit emergent properties considered as powers or capacities to behave, such aspects of its total behaviour as are characteristic of that emergent property are mediated by the property itself, on the level that it characterises. In other words, it is not a particular mental state M that emerges from a particular physical structure, P, but *mentality* that emerges from structures like P. This kind of thinking is particularly evident in Morgan's words. He writes:

> IN the foregoing lecture the notion of a pyramid with ascending levels was put forward. Near its base is a swarm of atoms with relational structure and the quality we may call atomicity. Above

> this level, atoms combine to form new units, the distinguishing quality of which is molecularity ; higher up, on one line of advance, are, let us say, crystals wherein atoms and molecules are grouped in new relations of which the expression is crystalline form ; on another line of advance are organisms with a different kind of natural relations which give the quality of vitality ; yet higher, a new kind of natural relatedness supervenes and to its expression the word " mentality " may, under safeguard from journalistic abuse, be applied.[37]

The difference can be illuminated somewhat by reflection on the different roles played by the intra-ordinal and the trans-ordinal laws in Broad's system. Recall that intra-ordinal laws are those governing the dynamics of entities within any level of emergence, whereas trans-ordinal laws connect emergent properties to their bases, and are an indicator of layer adjacency. The emergentists were were perfectly aware that many of the higher level properties they considered to be emergent were multiply instantiable: many kinds of structures of highly diverse descriptions, for example, exhibit vital properties. They were also sensitive to the fact that isolable substructures within a structure were often responsible for specific aspects of that systems overall emergent behaviour: different parts of the living body handle different aspects of vital behaviour, and a certain subsystem is correlated with the higher mental functions in those structures that exhibit mental properties. In Broad's emergentism, different trans-ordinal laws connect various aspects of the characteristic behaviour of a structure to different substructures in that system. Once the subsystems are in place and working as they ought to form the base for the emergent aspect, it is intra-ordinal laws that subsume the interaction of the emergent properties with other properties on their own or on other levels. Broad writes: "The law which asserts that all aggregates composed of such and such chemical substances in such and such proportions and relations have the power of reproduction would be a instance of a Trans-ordinal Law. The laws connecting the reproduction of living bodies with other ultimate characteristics of living bodies would be instances of Intra-ordinal Laws."[38]

So it does not appear to be the case that, as Kim's argument requires for the case of emergentism, the only way to cause an emergent property to be instantiated is to cause its emergence base property to be instantiated in the sense that, for some mental state M to cause some other mental state M*, M must somehow cause some new physical emergence base property P* to occur to serve as M*s emergence base. Rather, on emergentism, as a consequence of an entity's being composed of such and such chemical substances in such and such proportions and arrangements, that entity exhibits certain vital and mental behavioural capacities. M's causing M* is characteristic of that behaviour and is subsumed by the intra-ordinal laws on the mental level. This does not mean that mental changes can occur without there being concomitant physical changes, or that the mental is somehow abstracted from the physical in its functioning. The emergentists would have rejected such a mysterious notion. They considered the functioning of the body and brain to be constitutive of vital and mental behaviour, via nonreducible, ultimate trans-ordinal laws. Rather than associating particular mental states with particular physicochemical states, which might be cited as causally bring about other physicochemical states bound to other mental properties via special trans-ordinal laws, the vital and mental capacities to behave conferred upon an entity by virtue of its

aggregational structure orchestrate a unique and ultimate, novel arena of behaviour according to their own emergent laws. But as earlier discussed, this means that the lower level laws that pertain to levels lower than the mental, and ultimately physical laws, in which mental generalisations have no echo, will be violated. This is downward causation exactly as Kim had described it.

I suspect that the problem of downward causation is not one that troubled the British Emergentists.[39] It should, however, concern modern physicalists, to whom the the principle of the causal closure of the physical domain is important, and Kim's assertion that nonreductive physicalism is committed to downward causation if it is committed to a realistic construal of realised mental properties. As earlier described, to construe a higher level property realistically is to construe it as causally potent, and mental realism is the claim that mental properties, in virtue of their characteristic features *qua* mental, are causally effective properties. Individual realised properties as nonreductive physicalism construes them are taken to be fully realised in physical states, and therefore to have complete physical descriptions, unlike as is the case in emergentist doctrine. Psychological properties *as such,* however, do not admit of any particular physical description. But if it is the characteristic features of properties construed intrinsically on the psychological level psychological properties *as such* that are described as being causally relevant, *and not the physical properties that realise them,* then the claim is that a kind of property that does not admit of physical description is capable of entering into causal transactions involving the physical, violating the principle of the causal closure of the physical domain.

If it makes sense to construe psychological properties realistically, then surely it makes sense to construe psychological laws realistically as well. If it is taken to be the case that a mental property say, for example, *the desire that p* is a real property with causal potential in its own right, then surely the law that says that *the desire that p,* taken in conjunction with *the belief that -p unless q* results, all things being equal, in *the intention to bring it about that q,* is a *real law,* albeit a hedged one. But what kind of a law would it be? It certainly cannot be a law grounded in physical laws, for the properties it subsumes are all multiply instantiable, irreducible functional properties. If it is taken that psychological laws are real laws, that they somehow supersede physical laws such that they are appropriate to the description of causal interaction on the psychological level, *and not physical laws,* then psychological generalisations begin to look very much like emergent intra-ordinal laws. If psychological laws are construed realistically and as irreducible to physical laws, then it is difficult to see in what other way they could be taken. But, as we have seen, emergent intra-ordinal laws involve the violation of lower level laws, and are incompatible with the principle of the causal closure of the physical domain.

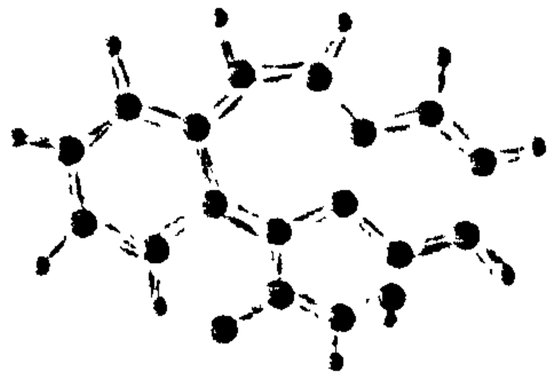
Consider Kim's argument again: suppose some mental property *M* brings about some physical event. If it does so according to generalisations that are not grounded in physical law and in virtue of features that have no physical base, then it is a clear instance of downward causation. Because the nonreductive physicalist describes mental properties as realised at any point in time in a particular physical base, then for *M* to cause a mental event *M**, *M* must bring it about that *M**s realization base, *P**, is

instantiated. To cause any event, M must cause a physical event, which again involves downward causation. It seems then that if the nonreductive physicalist is committed to a realistic construal of mental properties as causally potent, then they are equally committed to downward causation.

If, on the other hand, the nonreductive physicalist rejects downward causation, claiming instead that it is the physical properties, the properties of the physical realization base of a particular mental property that are genuinely causally responsible for whatever effect the mental property is said to have, it is difficult to see in what sense the mental property could be construed realistically as described above, as causally potent *qua* mental. This difficulty is what Kim has elsewhere termed the problem of causal exclusion in the context of the compatibility of physical and mental explanations for the same piece of behaviour: suppose some mental event M is cited as a cause of some physical effect event E. M is a higher level property realised in a particular physical base, P. Now, to deny that E has a complete physical cause would be to violate the principle of the causal closure of the physical domain. Thus, it seem that the physicalist committed to that principle should acknowledge that P, the physical realization base of M, is in fact a complete cause of E. But if it is so, then what is left for M, *qua* M, to contribute? It would seem that to consider M as causally effective, one must not construe it as an intrinsic property in its own right, but in terms of its physical realization base, and this would imply that, in order not to violate the principle of the causal closure of the physical domain, one cannot construe mental properties realistically. Further, if it is impossible to construe mental events realistically *qua* mental as causes of physical (or other mental) events, then it is difficult to see how the science of the mental, or the sciences of any multiply instantiable higher level properties, can count as genuinely explanatory sciences.

The nonreductive physicalist would seem to be at an impasse. It appears that the notion that mental states *qua* mental are causally potent, connectedly that the psychological sciences are genuinely explanatory, the notion of the multiple instantiability of the mental, and the principle of the causal closure of the physical domain are not mutually compatible. Yet none of these principles is easily rejected. That psychological generalisations are not genuinely explanatory and meaningfully predictive is not only an unappealing proposition, but if the idea behind the rejection that they are so is generalised to encompass all of the special sciences that involve multiply instantiable properties, it begins to look wildly improbable: whatever one may think about their ultimate explanatory status, it is difficult to deny that many of the special sciences very simply *work*, and that the phenomena they describe economies, migratory behaviour, evolutionary chains, human interactions are *real phenomena*. The proposition that mental states can be realised in systems with very different physical descriptions seems very likely to be true each individual human brain is different from every other, with different connections and neural pathways formed over time by different experiences. Yet we communicate; we share thoughts about our experiences and use words (often) to refer to the same objects. And the principle of the causal closure of the physical domain, I have argued, is what underlies the idea of interlevel explanation. To reject it would be tantamount to rejecting physicalism in any recognisable form.

We have seen how the functional model of reductionism solves the problem of causal exclusion for realised mental properties. According to this model, mental properties, when construed extrinsically as second order properties specified functionally, are reducible to and identical with whatever performs the specified function. Where the higher and lower level properties involved are identified, there can be no question of which one of them is the real causal factor. But the question of the status of the special sciences, and of the generalisations that they employ, remains open.



5
Dynamic Emergence
5. Downward Causation in the Leptothoracine Ants

It is uncontestable that mereological wholes exhibit properties and causal powers, derived from the properties and organisations of their parts, that the parts do not and cannot have *in abstracta*. It is likewise beyond reasonable suspicion on the level of common sense that the special sciences, and most notably psychology, employ a descriptive and predictive apparatus that is genuinely workable. It is reasonable to think that this is because psychology, among others of the special sciences, describes *real properties* that can be ascribed to certain kinds of wholes, that these properties characterise the entities that exhibit them in ways that can be described in general terms, and that the generalisations that describe the dynamics of entities exhibiting these properties describe *real laws*. In other words, it seems reasonable to think that what is behind the descriptive and predictive success of the special sciences is the reality and causal efficacy of the properties and generalisations to which their terms refer. However, we have seen that there is a conflict involved in the realistic construal of such generalisations as the special sciences employ and a commitment to robust physicalism, for according to robust physicalism, all causal interactions must take place according to basic physical laws, and basic physical laws must back all higher level causal transactions. But laws involving properties that may be instantiated in wildly heterogeneous physical bases cannot be expressed in physical terms. Thus, the realistic construal of such laws appears to be in direct conflict with the principle of the causal closure of the physical domain.

According to the traditional model of reductionism, special sciences whose terms are applied to multiply instantiable properties cannot be reduced to physical sciences, and we have seen how this irreducibility leaves the nonreductive physicalist to choose between rejection of the principle of the causal closure of the physical domain (if they retain their commitment to a realist construal of psychological generalisations) or

instrumentalism about psychological generalisations (if they retain their commitment to a physicalism that has the principle of the causal closure of the physical domain as a tenet). The functional model of reductionism described above avoids many of the problems and shortcomings of the traditional bridge-law model of intertheoretic reduction, and entails that the special sciences do not reduce to physics, but in itself the functional model of reductionism leaves open the problem of the status of special science generalisations as descriptive of real dynamics and along with that the causal efficaciousness of mental states *qua* mental. If we wish both to commit to robust physicalism and to maintain a realistic attitude towards mental properties, among other properties described by the special sciences, we must find a way to describe the relation of the dynamics described by the special sciences to those described by physics in a way that respects both the integrity of special science properties as causally potent described intrinsically and the principle of the causal closure of the physical domain. We must find a way to characterise multiply instantiable properties and the generalisations that subsume them in a way that is unproblematic from a physicalist perspective.

5.1 Emergence Revisited

It might be tempting to describe higher level properties as a kind of 'emergent simplicity', or as a kind of identifiable pattern characterising the causal potential of complex wholes that 'rides upon' the complex properties and interrelations of the whole's components. One modern writer who has discussed the notion of emergence in scientific explanation in such terms is David Deutsch. He defines 'emergence' as follows: "An emergent phenomenon is one... about which there are comprehensible facts or explanations that are not simply deducible from lower-level theories, but which may be explicable or predictable by higher-level theories referring directly to that phenomenon."[1] However, unlike as is the case in British Emergentist doctrine, he describes higher-level explanations as not incompatible with lower-level explanations.[2] For him, the core idea of emergence has to do with the explanation of higher-level phenomena in terms appropriate to basic physics as opposed to the special sciences to which those phenomena are appropriate. Basic physics can provide descriptions of any phenomenon in so far as it can provide descriptions of the properties and configurations of the basic physical constituents of that phenomenon, and in so far as basic physics can describe the basic physical state of a system at any time, it may be possible (in principle) to predict the basic physical configuration of the system at some later point in time.[3] But, he argues, this is not to provide an explanation of the phenomenon in question.

For example, Deutsch asks us to consider the position of one particular copper atom in the tip of the nose of the statue of Winston Churchill that is in Parliament Square in London. Why is that copper atom in that particular place? It may be the case that a very sophisticated physics "...would in principle make a low-level *prediction* of the probability that such a statue will exist, given the condition of (say) the solar system at some earlier date. It would also in principle describe how the the statue presumably got

there. But such descriptions and predictions (wildly infeasible, of course) would explain nothing."[4] This is because the description of the trajectory of our particular copper atom through space and time, and the concomitant description of the trajectories of vast numbers of other atoms, will leave out the higher level reason for that particular copper atom's being there: "It is because Churchill served as prime minister in the House of Commons nearby, and because his ideas and leadership contributed to the Allied victory in the Second World War; and because it is customary to honour such people by putting up statues of them; and because bronze, a traditional material for such statues, contains copper, and so on."[5]

The understanding of such higher-level explanations involves the resolution of very complex descriptions of low-level phenomena into overlying patterns. Recognition of such patterns enables us to understand the trajectories of vast accumulations of basic physical entities, incredibly complicated in itself, in simpler terms. These overlying patterns are what Deutsch terms emergent phenomena. He says

> The reason why higher level phenomena can be studied at all is that under special circumstances the stupendously complex behaviour of vast numbers of particles resolves itself into a measure of simplicity and comprehensibility. High-level phenomena about which there are comprehensible facts that are simply not deducible from lower-level theories are called *emergent phenomena.*[6]

The notion of emergence presented by Deutsch is one of higher level patterns of simplicity overlying lower-level complexity, such that the ordering of physical systems can be understood in terms of the interactions of physical organisations over which such patterns exist. Further, explanations that do not advert to such higher-level systems, he thinks, are not *explanations,* but are merely *descriptions* and *predictions.*

It should be clear, however, that this characterisation of emergent phenomena is not sufficient to set higher-level phenomena described as emergent apart from their component parts such that generalisations involving higher level phenomena can be construed realistically, for the problem of explanatory exclusion has not been addressed. If, as Deutsch admits is the case, there exist complete descriptions in terms of basic physics for all higher level phenomena, then for those phenomena that are emergent in Deutsch's sense and are multiply instantiable, a conflict can be seen to exist between explanations in terms of the generalisations subsuming the phenomena as described in higher-level terms and the explanations in lower-level terms subsuming the interactions of the parts. For while Deutsch asserts that descriptions of higher level phenomena in lower level terms are not *explanatory* in that they fail to capture the inherent qualities of the higher-level phenomena that they describe, this is primarily because higher level terms do not exist in the lower level sciences. Basic physics cannot capture the higher level phenomena involved in the position of our particular copper atom because basic physics does not employ concepts such as 'Prime Minister', 'War', or 'Honour'. That certain structures of elements or certain microphysical dynamics can be resolved into such things as Prime Ministers, wars, honourings, and so forth, involves a 'creative process of discovery', through which the intrinsic nature of such systems as they are described on the higher level is creatively uncovered:

> ...even if you had the superhuman capacity to follow such lengthy predictions of the copper atom's being there, you would still not be able to say, 'Ah yes, now I understand why it is there.' You would merely know that its arrival there in that way was inevitable (or likely, or whatever) given all the atoms' initial configurations and the laws of physics. If you wanted to understand why, you would still have no option but to take a further step. You would have to inquire into what it was about that configuration of atoms, and those trajectories, that gave them the propensity to deposit a copper atom at this location. Pursuing this inquiry would be a creative task, as discovering new explanations always is. You would have to discover that certain atomic configurations support emergent phenomena such as leadership and war, which are related to one another by high-level explanatory theories. Only when you knew those theories could you fully understand why that copper atom is where it is.[7]

But exactly how is the understanding of the position of the copper atom in terms of social policy a privileged explanation? If the understanding of a certain copper atom's being in a certain geospatial location is said to require the understanding of its placement there in terms of such phenomena as prime ministers, wars, honour, and such, we might rightly ask what justifies this explanation over other, compatible varieties of higher-level explanatory theories. Proponents of the selfish gene theory, for example, might argue that the positioning of that particular copper atom had far more to do with the propagation of Winston Churchill's genetic code than with the social notion of honouring prime ministers with statues. Too, it could be pointed out that while different explanatory theories may offer different explanations for the same phenomena, relative to each other, they will all 'leave something out', relative to each other: the selfish gene theorist may complain that the social explanation makes no reference to the fitness of Winston Churchill's genes. The basic physicist may argue that while the social explanation may give a reason why *a* copper atom is in that particular geospatial location, it does nothing to tell us why *that one* is there: surely any old copper atom would suffice to help conspire in a bronze representation of Winston Churchill's nose? If there is no particular privilege that one explanatory theory can assert over another, then how many explanatory theories must we discover before we can be said to *understand* why that copper atom is there?

Certainly, there is a sense in which there is a 'something else' that we may ascribe to aggregates insofar as we individuate them on the macroscopic level, for whatever reason we may do so. Things are particular dynamic configurations of subatomic particles, fields, and forces, or whatever our most basic physics describes things as being. They are also cubic centilitres of heavy water, gold doorknobs, coffeepots, zebra finches, dual carriageways, and the like. We recognise such higher level individuals as higher level individuals because we pick them out as defined aggregates, and we group them into kinds insofar as examples of them have the same kinds of properties, properties which they have as a result of possessing certain microstructures with certain causal potentials. But the reading of these higher level properties as patterns of simplicity overlying the complexity of basic physical components and dynamics is not enough to make real sense of the generalisations subsuming higher level properties. Though it is beyond question that large collections of atoms organised into human beings in societies with Prime

Ministers, wars, commemorative statues and the like are capable of producing behaviour which those atoms could not if not so organised, the description of such capabilities in terms of a pattern of simplicity riding upon the complex behaviours of the atoms is not enough to save society from a basically Hobbesian analysis. Similarly any such interpretation of higher level properties as properties that can be 'read into' the complexity of the microstructure of the whole instantiating the properties is not enough to privilege the laws subsuming the dynamics of higher level properties with the status of *real laws* and explanations on their terms as *genuinely explanatory*. If we wish to think realistically of higher level generalisations and also respect the principle of the causal closure of the physical domain, it must be possible to characterise at least some higher level properties as robust properties, in the sense that these properties are best understood as intrinsic properties of a complex system, from whatever vantage point from which we wish to view a complex whole and its components, and we must find a way to describe such properties without compromising their dependency upon their lower-level microstructures.

Though in much philosophical literature the term 'emergence' generally refers to the sort of phenomenon described by Broad and the other British emergentists, *vide* a higher level property of a system that cannot be explained by reference to the microstructure of that system, it has another use more common in scientific discourse, and which is gaining philosophical currency particularly in the fields of the philosophy of artificial intelligence and artificial life. In this sense, 'emergent' is applied to a property of a mereological system which arises out of the dynamic behaviours of its elements, but which is neither 'coded for' by any of those elements nor 'directed' by some central organising component or arrangement. Such properties are perhaps best illustrated by example.

In an article published in *New Scientist*,[8] Brian Goodwin describes a rhythmic pattern of oscillation between activity and inactivity that is observable in colonies of some species of ants in the genus *Leptothorax*[9] observed by entomologist Nigel Franks and further studied and described by Blaine Cole. In general, the Leptothoracine ants are active for a certain amount of time, then they lapse into quiescence for a while. Across entire *Leptothorax* colonies, the activity patterns of the ants are oscillatory: ants tend to oscillate between activity and inactivity such that most ants are active or inactive at the same times, taking about 30 minutes to complete one cycle. Individual ants, however, do not display this behaviour. Observing isolated ants and ants in very small groups, Cole found that the ants exhibit no intrinsic rhythmic or co-ordinated activity, but rather that their activity/inactivity behaviour is chaotic. But when the density of the ant population reaches a certain level at which *Leptothorax* colonies do in fact tend to be found naturally, the oscillatory patterns in the activity of the ants appeared.

Together with Richard Solé and Octavio Miramontes, Goodwin developed a computer simulation to show how such a collective dynamic could appear at a special population density given individuals whose behaviour in isolation does not resemble that of individuals in the dynamic colony or of the colony as a whole, when the behaviour of the individuals is constrained by certain simple rules. Solé, Miramontes and Goodwin

designed computer simulated 'ants' whose movements about the grid in which they were situated were directed by neural networks producing behaviours described by deterministic chaos, as was their activity/inactivity behaviour patterns, so that their individual behaviours were like those of real *Leptothorax* individuals. Their simulated ants could interact with each other in some ways similar to those in which real *Leptothorax* interact: whenever an active sim-ant encountered an inactive sim-ant on the grid, the inactive sim-ant would be stimulated into activity. They found that, when the sensitivity of inactive sim-ants to be stimulated into activity by active ants was within a certain (comparatively wide) range, and when the population density was at the right level (about eighty individuals), the 'colony' exhibited a well-defined oscillating pattern of activity/inactivity. Goodwin describes this orderly behaviour as an emergent property of the modeled colony: "...model ants, behaving chaotically and interacting "socially" by stimulation, can generate a collective rhythm throughout a colony. This is a clear case of emergent behaviour: it was impossible to predict the collective, rhythmic pattern just from a knowledge of the chaotic behaviour of the individuals."[10]

What is important to notice about this rhythmic oscillation of activity across *Leptothorax* colonies is that it is not just a pattern that can be recognised as overlying the collective activity of many individuals, not merely a simple pattern into which complexity on the level of individual ants can be resolved when looking at the colony from a higher perspective. It is a robust property of the colony as a whole, and one which may be[11] of benefit to the colony as a whole. But the behaviour has, it must be stressed, no expression in the behaviour of the individual ants, much though it is determined by the properties of multiple individuals in certain concentrations. Such well-defined, robust organisation arising in this way from the collective actions and interactions of individuals with no several intrinsic directing towards such organisation has been termed 'order for free',[12] or self-organization.

Perhaps some of the most obvious and compelling examples of emergent organised behaviours occur in the social insects, such as the building of complex nests, a behaviour that is of obvious advantage to the insects concerned. Andy Clark describes the mechanism by which termites build arched passageways: individual termites are disposed to make mud balls, which at first they will drop in random locations, adding as they do so a chemical trace to the mudballs. Termites prefer to drop mudballs in places that carry the chemical signal, so that a termite carrying a mudball after a few mudballs have been placed will prefer to place her mudball on top of another mudball, which carries the trace. Now there are two mudballs in one location, increasing the attracting power of the chemical trace in that location. Other mudball-carrying termites will be drawn to place theirs in these locations bearing stronger chemical signals. The more mudballs, the stronger the signals, until eventually columns of attractive mudballs are formed. If two columns are within a certain range of each other, termites will tend to try to place their mudballs where the chemical signal is strongest, on the side of a column facing a neighbouring column. This increases the attractiveness of the site, and eventually an archway is formed. Thus, blind insects with tiny, simple brains are able to accomplish a veritable architectural feat, literally without having any idea what they are

doing. All they 'know' is that mudballs are to be placed where the chemical signal is strongest. By following this and a few other simple rules, inevitably, a complex structure emerges, with ordered tunnels, chambers and cells, and one which is suited to the natural features of the environment in which the nest is built: an apparently designed artefact appears, in the complete absence of designers. This kind of behaviour involves the added dimension of *stigmergy,* the process by which the results of one termite's actions, the placing down of a mudball bearing a chemical trace, stimulates another termite to perform another action, the placing of a mudball on top of that one, which Clark uses the arch-building behaviour to illustrate.[13]

Photograph by John Mullen, ©John Mullen 2001

A termite mound in Matopos National Park, Zimbabwe. No design or blueprint exists for this, anywhere. [14]

While properties like these are fully determined properties explicable in terms of the properties and relations of the individuals in the colony (simulated and natural), as Goodwin, Solè and Miramontes show in their simulation and as Clark describes, they are robust and interesting properties of the collectives as a wholes. Because these kinds of properties arise from the collective activities and interactions of dynamic components of complex systems, and to distinguish the sense in which robust properties of whole systems arising from the collective dynamics of the components of the system are 'emergent properties' of the system considered as a whole from the traditional sense of 'emergence' as employed by the British emergentists, an inexplicable property of an aggregate necessarily linked to aggregates of appropriate description, I propose to term such properties 'dynamically emergent'.

Dynamically emergent properties are by no means confined to the eusocial hymenoptera. Clark cites Hutchins (1995), who provides another example, interesting in that it features not the behaviour of social insects, but the very human example of the successful navigation of a ship. Hutchins writes:

In fact, it is possible for the [navigation] team to organize its behaviour in an appropriate sequence without there being a global script or plan anywhere in the system.[15] Each crew member only needs to know what to do when certain conditions are produced in the environment. An examination of the duties of members of the navigation team shows that many of the specified duties are given in the form "Do X when Y" Here are some examples from the procedures:

A. Take soundings and send them to the bridge on request.

B. Record the time and sounding every time a sounding is sent to the bridge.

C. Take and report bearings to the objects ordered by the recorder and when ordered by the recorder.[16]

While it is possible for the human agents responsible for the various tasks they perform aboard the ship to conceive of the ship, of the navigation of the ship, and of the role that their activity plays in that navigation, it is not strictly necessary that they do so. No one member of the crew needs to understand *how the ship gets navigated,* and though it is surely the case that this property of the ship and its crew— that it be successfully navigated from one point to another, adapting its behaviour to deal with whatever hazards it may encounter— is a designed one, represented in somebody's mind, its actual execution can be unrepresented and undirected. It might be carried out by computers and robots programmed only for very specific tasks, just as, for example, Rodney Brooks's six-legged walking robot 'Attila' (now retired) is able to navigate over sloped and cluttered terrains without a 'control centre' of any description, but through the collective activity of several finite-state motors and processors which together yield skilful walking behaviour."

Clark distinguishes two ways in which robust properties of collectives of individuals can arise without the need for the property to be 'coded for' or directed by any property exhibited by the individuals themselves, which he terms *direct emergence* and *indirect emergence.* The key notion distinguishing the two forms is the role of the environment in producing the emergent, or 'target', property. Direct emergence "relies largely on the properties of (and relations between) the individual elements, with environmental conditions playing only a background role." The oscillation of activity in *Leptothorax* colonies, arising out of nothing but certain properties of the individuals and their relations to each other (e.g., population density) would be an example of direct emergence. Indirect emergence "relies on the interactions of individual elements but requires that these interactions be mediated by active and often quite complex environmental structures".[18] The arch-building behaviour of termites, with its dependence upon stigmergy, would be an example of indirect emergence. The use of the environment as cues for the actions of individual members of the collective allows for extremely complex robust properties of the collective to arise whilst keeping the properties of the individuals in the collective simple and economical. It would be a very complicated termite indeed that had an on-board representation of a complete termite mound in its head, which had the sensory and cognitive capacity to evaluate the proposed nest site, plan a mound for that site, evaluate termite mounds in various stages

of completion, and which knew how to organize its own actions or the actions of other, less complex 'follower termites' towards the completion of a functioning termite mound. But in order for strong, appropriately formed, functional termite mounds to emerge from the collective activity of termites, no such complexity of the individuals is necessary, and no 'foreman-termite' directing the behaviours of its underlings need exist. Termites need only be able to recognise and react to specific stimuli in certain simple ways, Again, Clark:

> At no point in this extended process [constructing a mound] is a plan of the nest represented or followed. No termite "knows" anything beyond how to respond when confronted with a specific patterning of its local environment. The termites do not talk to one another in any way, except through environmental products of their own activity. Such environment-based coordination requires no linguistic encoding or decoding and places no load on memory, and the "signals" persist even if the originating individual goes on to do something else.[9]

Complex properties of collective systems can thus be 'coded for' in simple rules governing the actions and interactions of members of the collective with one another and with their environment, with no particular component or subset of components acting as a director or control centre organising the system towards the 'target property' (Clark's term). Robust order can be bought very cheaply in collective systems through exploitation of simple rules describing how individual members of the collective interact with each other and, in the most interesting cases of emergence, with the environment in which the collective is situated. Such order can be (and generally is, when it is found in natural systems) of positive value for the collective as a whole: nest-building, food-storing, brood-raising, and such properties are of clear benefit to the social insects, and examples such as the robot walkers show how clever humans can design systems to exhibit certain desirable target properties without specifically coding for them, and thus avoiding the cost in complexity that such coding would necessitate. Further, this kind of order is not simply a resolution of the complex behaviours of the individuals into patterns of simplicity enabling us to understand the dynamic of the system in new ways. Quite the opposite, it is complexity arising from the simple behaviours of individual components of the collective. No single element of the system has an on-board representation of the target property it and its cohorts conspire to realise. None of them is individually capable of the complex relations between the collective and the environment into which the whole is able to enter: no single processor/motor subsystem of Attila 'knows' how to step over an obstacle, and no termite 'knows' how to build an arch. Out of whole systems exploiting simple parameters of their components arise complex behaviours which, if they were to be accomplished by a single devoted system representing and designing the target behaviours, would require immense powers of representation, computation and manipulation. Such robust properties of collective systems can, in a sense, be interpreted as characterising those systems: Attila is a collection of motors and finite state processors, but in virtue of the way those processors interact, it is also a *walking robot;* a ship with its crew is a *seafaring vessel.* An ant colony is a bunch of ants of different descriptions searching out food, moving soil about, laying eggs, and so forth, but to

consider the behaviour of individual ants is really to miss the point of the colony, which is a thriving thing and a vehicle for the survival of the species, which individual workers, being by and large sterile, cannot be. As myrmecologists E. O. Wilson and Bert Holldöbler put it, "One ant alone is a disappointment; it is really no ant at all".[20]

Another interesting and important feature of dynamically emergent properties which is particularly compelling of their construal in realistic terms is that once a system exhibits a dynamically emergent property, the property becomes self-propagating, in that the behaviours of the individual components of the system instantiating the property are constrained by the presence of the property. For example, an individual representative of *Leptothorax,* abstracted from its colony, is about as likely to be active as inactive at any given time. Within the colony, however, the individual is determined to exhibit a routine, ordered activity/inactivity cycle co-ordinated with that of its nestmates, which are similarly constrained by the order that arises from their behaviour at appropriate densities. Collective activity thus constrains the behaviours of individuals within the collective. This is a feature of all self-organising systems. As Clark says, "["self-organising" systems] are such that it is simultaneously true to say that the actions of the parts cause the overall behaviour and that the overall behaviour guides the action of the parts."[21] This feature, he tells us, is sometimes termed 'circular causation'.[22]

This feature is especially interesting upon consideration of Kim's definition of 'downward causation'. Recall that Kim had characterised the notion of downward causation as the idea that "*higher level mental events and processes cause lower level physical laws to be violated*, that the molecules that are part of your body behave, at least sometimes, in ways different from the way they would behave if they weren't part of a body animated by mental processes".[23] Within a self-organised collective instantiating a dynamically emergent property, the individuals do indeed behave in ways in which they would not if they were not within a collective 'animated' by the property. However, they do so in a way that violates no lower-level laws. As Goodwin, Solè and Miramontes have shown through their computer model of the oscillation of activity patterns in simulated ants and as Clark has amply illustrated in various examples, ordered collective behaviour arises naturally out of the individually directed behaviour of components interacting in specific ways. Thus, even better than 'order for free', dynamic emergence can be understood as giving us a kind of 'downward causation for free': the constraint of the behaviour of individual components within a dynamic collective by the overall behaviour of the collective is a natural, nonmysterious phenomenon, in that it is fully explicable in terms of the properties and relations of the interacting parts of the system.

This feature is also illustrative of the way in which understanding the collective dynamic of a system exhibiting a dynamically emergent property in terms of that property is motivated over consideration of systems in terms of patterns of simplicity overlying collective complexity. As I expressed earlier, the privilege of a certain explanatory framework forwarded to understand certain kinds of phenomena is questionable when different compatible explanatory frameworks are possible, such as the explanation of the position of a particular atom in basic physical terms, or in terms of the social construct that built the statue in whose nose the atom is, or in any one of many possible

explanations in terms of alternative theories. The properties referred to in these several systems will not directly be represented in other, compatible explanatory theories or in theories that pick out properties on different levels of description: there is no correlate for 'prime minister' in basic physics, and social theory does not include the predicates of basic physics among its explanatory apparatus. But because dynamically emergent properties are self-propagating, they can be seen as pervasive through the system as far down as the lowest complete dynamic components:[24] in constraining the behaviour of individual components, dynamically emergent properties are robustly present at the level of the component. Thus, there is a higher degree of objectivity in the assignation of such properties to wholes, and concomitantly a higher degree of motivation for the understanding of such wholes in terms of dynamically emergent properties that they instantiate.

5.2. Realism

Dynamically emergent properties are special kinds of properties of mereological wholes, and are especially interesting in a number of respects. They are completely determined by and explicable in terms of the properties and relations of their active and interactive components, and in the most interesting cases, features of the environment upon which the components of the system act in various ways and from which they receive various behavioural cues. As such, they are unproblematically realised properties from the point of view of robust physicalism. Though they are perfectly physically realised properties susceptible to physical explanation, however, there are provocative reasons for the consideration of dynamically emergent properties in their own terms, as genuinely furthering understanding of the systems instantiating them. The phenomenon of 'circular causation' is particularly compelling: dynamically emergent properties are, once realised, also dynamically self-perpetuating, constraining the behaviours of component elements in ways for which the components have no intrinsic leanings for such constraint. Such a phenomenon certainly invites the understanding of particular behaviours of individual components within a system in terms of the higher level dynamically emergent property that component has a part in realising. Systems exhibiting dynamically emergent properties are also able to act in and interact with the environment and with other similar systems in highly complex and even apparently intelligent ways, without requiring that the components themselves or any subset thereof be complex or intelligent, or that they receive direction from something that is. By exploiting simple parameters limiting the behaviour of its components the whole system generates for itself what Goodwin called 'order for free'. Dynamically emergent properties are particularly robust properties, and the causal potential of mereological wholes exhibiting dynamically emergent properties can be seen as on another order from 'merely resultant' properties of complex wholes, though without compromising the explanatory concerns of robust physicalism.

Inasmuch as dynamically emergent properties are fully realised by their

components and the relations of those components to each other and in some cases to aspects of their environment, phenomena involving dynamically emergent properties *can* be explained in terms of the properties of the components realising them. But Clark argues that explanation of emergent phenomena in terms of the components (Clark terms this kind of explanation 'componential explanation'[25]) may not be completely informative in terms of a full understanding of such phenomena. Clark gives two primary reasons for this, the first motivated by the fact that dynamically emergent phenomena often involve aspects of the environment in which the components of the system are situated, as well as the properties of the components themselves. Explanations that track the behaviours of the individual components rather than that of the whole will in general fail to do justice to dynamically emergent phenomena so rooted. Rather, Clark says, an ideal explanatory framework for the understanding of such phenomena will have two main features: it will be "...well suited to modeling both organismic and environmental parameters" and it will model both kinds of features "...in a uniform vocabulary and framework, thus facilitating an understanding of the complex interactions between the two".[26]

The second reason Clark gives has to do with the possible nature of the components realising dynamically emergent properties themselves. The analysis of a system in terms of the components of the system is perfectly appropriate and informative when those components contribute to the behaviour of the whole system in well defined ways characteristic of the kinds of components that they are and the particular function they perform for the system. For example, *how a car works* could interestingly be explained by reference to what all of the various car parts do in a functioning car. But for systems where there may be very little or no difference between components which all perform much the same function, or for systems the behaviours of whose components are highly interdependent such that the states of some are determinants of the states of others, such explanation may not be appropriate. Clark says:

> When each of the components makes a distinctive contribution to the ability of a system to display some target property, componential analysis is a powerful tool. But some systems are highly homogenous at the component level, with most of the interesting properties dependent solely upon the aggregate effects of simple interactions among the parts... A more complex case occurs when a system is highly nonhomogenous, yet the contributions of the parts are highly inter-defined— that is, the role of a component C at time t_1 is determined by (and helps determine) the roles of other components at t_1, and may even contribute quite differently at a time t_2, courtesy of complex (and often nonlinear— see note 8[7]) feedback and feedforward links to other subsystems."[28]

Analysis of systems like this in terms of what individual parts contribute is likely to be uninformative as to how the whole behaves, and in the absence of understanding of the collective dynamic and how it affects the individual components of a system, the behaviour of the individual components may be difficult to interpret. It seems that though explanation of the properties and behaviour of complex systems exhibiting dynamically emergent properties and of phenomena involving them in terms of the microphysical structure of such systems might be possible, explanations that attend to the collective dynamic are interesting and informative not just for understanding of the

systems considered as wholes, but for understanding of the behaviours of the components of the systems as well. Further, since dynamically emergent properties are so very robust, enabling the wholes exhibiting them to behave in extremely different and sometimes vastly more complex ways than anything of which the dynamic components are capable, and since dynamically emergent properties pervade the system on the component level, explanations that do not track the collective dynamic will miss a very real aspect of the system instantiating it.

Clark suggests that phenomena involving what I have termed dynamically emergent properties are best understood in terms of dynamical systems theory, a branch of mathematics describing systems in essentially geometrical terms, tracking the changing states of dynamic systems as the progression of the system through various states in a space defined as all possible states in which the system might be, including attractive features of the topology of the state space which tend to draw the system towards certain states, all of this defined by a set of governing equations. Considering the theory as a model for explaining the behaviour of complex systems, these equations can be taken as expressing *laws* about the behaviour of the system: it illuminates counterfactual conditionals about the behaviour of the system in possible situations. Dynamical systems theory is a well-established and successful explanatory and predictive scientific tool, which treats the changing states of whole systems irrespective of the microphysical properties of the systems concerned. It is thus able to deal meaningfully with systems with homogenous or highly interdependent components. It is also able to accommodate phenomena involving systems whose behaviour involves reacting to or manipulating features of the environment by treating the system and the environment as a coupled system, described by a set of interlocking equations.[39] Clark is, however, at pains to emphasise that the dynamical systems approach to modelling the behaviour of real systems should be regarded as complementary to explanatory systems treating the properties and functions of the parts of the system. In the absence of information about the parts and how they contribute to the dynamic of the whole system, we lack understanding of how dynamic systems are grounded. According to Clark, both kinds of understanding are required for a full, complete explanation of dynamic systems: understanding of the collective behaviour of the system, which constrains the behaviour of the component and may not in the slightest resemble the behaviours of the components, and information about the properties of the components that determines how their collective activity realises the behaviour of the whole system.

Dynamically emergent properties are very real, robust properties of complex systems, enabling the systems instantiating them to act in and interact with their environment in highly complex ways, even as the actions of the components of complex systems may be governed by simple, highly specific rules. Thus the properties of the whole system can be radically different from the properties of the components which fully realise the system. Dynamically emergent properties of whole systems also affect individual components of the system, such that the activity of the components is constrained by the whole in ways for which the components themselves have no intrinsic leanings. For the reasons detailed above, the understanding of phenomena involving

dynamically emergent properties partially in terms of the dynamically emergent properties themselves, not just in terms of the microphysical structure of the systems exhibiting dynamically emergent properties, is motivated. If the understanding of complex phenomena involving dynamically emergent properties essentially involves explanation in higher level terms, as Clark has argued, then realism about higher level generalisations subsuming the dynamics of the whole system, such as might be described in dynamical systems theory, is motivated. These laws deal with the unfolding of the states of whole systems through time, not with the components of the system. Dynamical systems theory is blind to the physical realisers of real phenomena it may describe, and the laws expressed in its terms are abstracted from physical law. However, when a dynamical systems account is accompanied by an account of components of the system detailing how the complex dynamic is determined by the properties and relations of the components, the laws expressed in the dynamical systems account can be seen to be grounded in the physical laws governing the behaviours of the components.[19] Of course, since dynamically emergent properties are fully realised in physical properties and the generalisations subsuming them are fully grounded in physical law, explanations in purely physical terms will exist for causal interactions involving complex systems exhibiting them. However, these explanations are best interpreted as grounding real generalisations subsuming robust, albeit irreducible in themselves, properties of whole systems, just as descriptions in microphysical terms can be seen as grounding any multiply instantiable property. The notion of dynamic emergence and the model for the understanding of dynamically emergent properties in terms of both the collective dynamic of the whole system and the physical realisers of the system provides us with a framework in which we can make sense of higher level generalisations as grounded in physical laws, and a way of construing higher level generalisations realistically without compromising the requirements of robust physicalism.

The notion of dynamicism is enjoying attention in neurophysiological research, as well, and some neuroscientists are beginning to think that the brain, with its population of some thirty billion neurons, may work on similar principles. Following John McCrone's[*] exposition of some recent thinking and work in neurophysiology, I would like to consider how one of the brain's abilities – the ability to recognise objects – can be thought about as a dynamically emergent property arising out of the actions and interactions of a great many specialised neurons in different areas of the visual cortex. The account I present will be, of course, highly simplified, but it is not so important that the picture presented be exactly right, as that it provide some insight into the way that brains might see.

In 1972, C. Gross, C. Rocha-Miranda and D. Bender were performing single-cell recordings from the inferotemporal cortex, an area of the brain now understood to be involved in the recognition of objects, qualities of objects and in reaching and grasping actions, of an anaesthetised[*] Macaque monkey. Using a projector, they presented different kinds of visual stimuli to the monkey in an effort to determine what kinds of stimuli would induce the cells in the inferotemporal cortex to fire. None of the stimuli they had prepared was having any interesting effect, until quite by accident, one of the

researchers caught his hand in the projector beam. A cell immediately responded, most enthusiastically, and would do so whenever a hand shape was present in the visual field. It appears that there are cells in the cortex of the Macaque that are specialised to signal the presence of hands.[33]

That different areas of the brain are specialised to handle different tasks has been common knowledge for some time, but it was not until the development of the technique for recording the responses of single cells in the brain to different kinds of stimuli that researchers came into a full appreciation of how specialised individual cells can be, from cells that code for such specific stimuli as a dark line against a light background angled at 11:00 and moving from left to right presented to a certain tiny corner of the visual field, to cells that code for entire objects, such as a hand. With much painstaking labour, neurophysiologists are beginning to be able to piece together the complex hierarchy of processing that is involved in the chain of events proceeding from light entering the eyeballs, to nerve signals entering the primary visual cortex at the back of the brain through the lateral geniculate nucleus, through many layers of processing culminating in the firing of cells coded to signal the presence of whole objects or significant parts of objects. Over thirty richly interconnected areas have been identified in the primate visual cortex, but it is standardly divided into five major areas: the primary or striate cortex, or V1; V2, V3, V4 — the 'colour centre' of the brain, and V5, or MT, the middle temporal cortex.[34]

Information about stimuli presented to the eyes enters the visual cortex via V1, a sizeable area, with six distinct layers, in the occipital lobe. As David Hubel and Torsten Wiesel discovered in 1958, cells in the V1 area are very picky about what will induce them to fire. Groups of cells responsive to certain frequencies of light presented to specific parts of the visual field occur at regular intervals throughout, but most of the cells in V1 are coded to respond to the presence of lines or edges in specific areas, only a few degrees across, in the visual field. Moreover, they respond only to lines appropriately angled. Some prefer dark lines against light backgrounds, whereas some prefer light lines against dark; some respond to an appropriately angled boundary between light and dark. Some neurons in V1 are also sensitive to motion: they will respond to the presence of an appropriately angled line in their few degrees of visual field, or to an appropriately angled line that is moving; most respond only to motion in a particular direction. Hubel and Wiesel classified cells in V1 as 'simple' or 'complex' depending on the type of stimulus to which they respond. 'Simple' cells will respond to just the right kind of line appearing in just the right spot, and have sub-regions in which the right kind of stimulus will have an excitatory effect, inducing the cell to fire enthusiastically, and sub-regions in which the presence of stimuli will tend to inhibit the cell's response. Simple cells are not sensitive to motion, whereas complex cells are, though complex cells are less choosy about where in their particular few degrees of visual field a stimulus is presented: so long as it is at the right angle and moving in the right direction (though some will fire if the stimulus is stationary), the cell will fire.

Thus there is in V1 "the machinery for a rich representation of line":[35] "With thousands of neurons to code for the same point in visual space, there would always be

the right kind of cell, or combination of cells, to specify exactly what the eyes were seeing."[36] Cells in V1 are also very neatly arranged: throughout the six layers of V1, cells preferring lines of the same kind and at the same angle are clustered together, whereas across the surface of the cortex, a cell preferring lines at a certain angle will sit neatly between cells preferring lines at an angle a few degrees less to one side, a few degrees more to the other, so that the arrangement of the whole cortex is a very ordered map. This order appears to pervade the sensory and motor cortices."

V1 maps fine-grained information, with cells reporting the presence by their firing or the absence by their silence of very particular kinds of stimuli in precise locations on the visual field. The other components of the visual cortex appear to take the data reported by V1 and refine it in various ways. V2 receives information from both line-sensitive and colour-sensitive cells in V1. Cells in V2, according to McCrone, have "a more powerful reaction to boundaries and wavelength changes, as if they were emphasising the principle shapes and blocks of colour present in the visual scene".[37] V2 in turn passes information on to V3, most of whose cells are sensitive to orientation, motion and depth, and to V4, which, as previously noted, has been described as the 'colour centre' of the brain. This is because it is in V4 that the rich array of colours we are able to perceive is added to our experience. The eye itself can only detect three discrete colours, but by refining the information feeding in from V1 and V2 about the distribution of those colours and general illumination, V4 maps the rich spectrum of our experience onto what our eyes detect. McCrone explains the process as follows:

> The process starts with the mapping of wavelength information in V1 and V2 using cells tuned to represent a red/green contrast or a blue/yellow contrast. So, for example, a green-coding cell will be switched on by predominantly green wavelength light hitting an area, and turned off if the mix of the three wavelengths is mainly red. But the response of each cell is too unrefined to say that it really sees a colour as such. It is only in V4 that the firing of cells begins to match the subjective experience of witnessing individual hues. Each v4 cell appears to combine the the response of many 'red', 'green', 'yellow', and 'blue' neurons, and to mix it up with some general information about illumination levels — the amount of 'black' or 'white' in the picture — to produce the complex reaction of seeing an actual shade such as turquoise, brown or pink. So what started out as a crude physical measurement — the wavelength response of a retinal pigment — could be turned by a chain of remapping and condensing into a highly specific psychological response."

The same kind of process of finer and finer refinement appears to be the case for the recognition of objects as well. On V1, there are representations of basic lines, basic motions, and colours, but very little analysis of the scene presented to the eyes. As the data is processed through the hierarchy of the visual cortex, the data becomes further and further analysed and refined, parts emphasised and parts filtered away into the background, until at the higher levels of processing in such regions as IT, cells light up to signal the presence of complex objects or recognisable properties of objects. Cells coding for more complex patterns, such as the hand-coding cell in the IT cortex of the Macaque monkey, behave differently from cells on lower levels of processing. They tend not to be concerned with where in the visual field their kind of stimulus is presented: the macaque's hand-coding cell would fire whenever a hand was visible to the monkey. Moreover, whereas when a stimulus stops being presented to a cell on the lower levels of processing such as V1 or V2, the cell will immediately stop firing, cells in higher areas of

processing will continue to fire, sometimes long after the stimulus has been removed.

The picture of the processing of information in the visual cortex presented so far is fairly straightforward: very specific maps of visual information made out of the firing of cells in the primary visual cortex are 'fed upwards' through higher and higher maps, becoming more and more refined, as if V1 were a kind of screen, where pixels of information are lit up, with patterns in the pixels stimulating cells in higher maps to light up as they resolve patterns in the lower-level maps. This is, as above noted, an extremely oversimplified account: there are multiple subdivisions in the different visual cortices, each connected in different ways with themselves and with other layers in other cortices. Too, information does not flow just one way, from the primary visual cortex 'upwards': the higher levels of processing communicate with the lower levels as well, and there are complex feedback and feedforward loops on all levels. Nonetheless, for current purposes, this highly simplified account of the processing of information from the eyes in the brain will suffice to illustrate the general point, and the general problems with this account.

A particularly troublesome problem for an account of the workings of the visual cortex presented above has been termed the 'Grandmother cell' problem. Consider again the hand-coding cell in the IT cortex of the Macaque. In the experimental conditions, the cell would fire when, and only when, a hand was present in the visual field. It would not fire in response to any other kind of stimulus. As McCrone says, "It had become semantically specific. By extension, this suggested there might be a high-level cell to stand for every level of experience, and the firing of that cell would somehow fill our consciousness with its presence. So every time we saw or thought about our grandmothers, somewhere in our brain a grandmother-coding cell would have to burst into action."[40] But this cannot be the case. One reason is that, notwithstanding the vast number of cells in the higher processing regions of the cerebral cortex, our abilities to experience complex phenomena far outreach the number of cells we would require on this model. Consider: we are able to recognise our grandmothers by numerous criteria. We may recognise her by sight, by the sound of her voice, or by other means. We recognise her from the back and in profile as well as when she is facing us. Too, we recognise her when she appears to us in guises in which we have never seen her: if she has always had long hair, for example, and one day has it cut very short, or if we recognise her in a picture taken when she was in her twenties. In each of these cases, the presentation of our grandmother to our senses produces a very different lower-level map. To recognise our grandmothers in as many guises as we actually can, we would have to have a vast number of Grandmother-coding cells, one to respond to each of the many significantly different grandmotherly experiences we may have, and this would have to be true for all of the higher-level phenomena we encounter, not just our grandmothers. We would thus require an impossible proliferation of cells in higher-level cortical areas to stand for all the phenomena we are able to experience. Even though we do have quite a lot, we simply do not, and cannot, have that many brain cells.

Further, it is the case that even if our ability to map the world is specifically compromised, we continue to be able to identify complex objects. People who have

suffered lesions in the V4 area, for example, have a condition called *achromatopsia,* in which they are neither able to perceive colours, nor to remember colours they had once perceived, if the lesion was suffered after infancy.[41] Achromatopsic individuals experience the world in shades of grey, because the area of the brain that deals with colour information is not functioning. Thus, the maps presented to the higher level processing areas in their cortices are incomplete. Yet they still recognise their grandmothers, just as would a person with a healthy V4 if presented with a black and white photograph of Grandmother. Another consideration is that brain cells die at a regular rate, or are killed by injuries. On a 'one cell, one function' account, we would expect very specific losses in the ability to experience to occur concomitantly with the death of cells in the higher processing regions of the cerebral cortex: we might expect the death of a certain cell or group of cells in IT, for example, to result in the failure properly to recognise very specific objects, such as Grandma or hands. But while we do find that damage to specific cortical regions can result in very specific cognitive deficiencies, we never find such gaps as, say, the failure to be able to experience Morbier cheese, and nothing else.

Another problem with the kind of interpretation given above is that the maps presented by lower-level processing areas to higher-level ones cannot be as crisp and sharply defined as they would have to be, for the simple reason that the response of cells to stimuli is not sharply defined. A cell in V1 coding for a dark line falling across a light background at a 45 degree angle from lower-left to upper-right in a small area of the visual field, for example, will fire enthusiastically when just that is present. It will also fire, albeit less enthusiastically, if the degree of the line is just slightly off in either direction, or if the line does not go all the way across the part of the visual field it is attending, or if the line isn't as dark as all that really or the background isn't as light. A cell in V4 that fires most happily in response to light at 640 nanometers, a colour McCrone describes as "a cheerful tomato red",[42] will also fire more weakly to 620 – a kind of orangy colour – and to 660, a more crimson shade, and indeed its response to orange and to crimson will be identical. The response of cells to the stimuli that fall within their receptive fields is better represented by a narrow bell curve than a sharp, well-defined peak. But how could higher-level processing areas of the brain build precise maps out of such ambiguous data?

What is ambiguous when reported by one cell, however, can be resolved into crisp precision when information across an entire group of cells is considered. To elucidate this notion, McCrone asks us again to consider the identification of that cheerful tomato red. Even if no cell responds to light with a frequency of 640 nanometers, tomato red can still be 'coded' in the responses of cells responsive to different shades across a population of cells:

> For example, say a red spot of light fell across the receptive fields of three V4 cells, each with a slightly different frequency response. One neuron might be provoked to fire at 80 per cent of its capacity, while another fired at 40 per cent, and the third at 10 per cent. None of the cells would be claiming to be 100 per cent sure about anything. None would be a tomato red cell, as such, but would merely be reporting how close the spot happened to come to hitting their own personal peak frequency. However, the particular combination of firing rates could only stand for a single

> wavelength. Only a hue of 640 nanometers could provoke a response of an 80, 40, and 10 per cent response within this small group of cells. From a population reaction, a pattern of voting by a spread of cells, would emerge an exceptionally precise trajectory of firing.[43] (94)

On such an account, there need be no cell or group of cells dedicated to alerting the brain of the presence of cheerful tomato red in the visual field. The group activity of populations of cells can 'code for' cheerful tomato red, even though the individual cells involved are dedicated to reporting something quite different. This kind of 'group coding' scheme works for the recognition of objects or specific motions as well. Cells in V1 through V5 produce maps of the raw data that is presented to the eyes, within a certain degree of precision. Across an entire population of cells and their activity and interactivity throughout the visual cortex, an unmistakable pattern can emerge that may be picked up by a cell in IT, which fires to signal the presence of a hand. 'Population coding' also solves the 'Grandmother Cell' problem. While it seems reasonable that we might have cells or groups of cells that fire in response to such commonly perceived outlines as hands in general, or faces in general, the notion that we have individual cells that code for each specific kind of higher level experience – Grandma's face, as against Ralph Nader's face – is not reasonable. But coding for Grandmother can exist in the complex action and interactions of a population of cells none of which is devoted to representing Grandmotherly experiences *per se*.

The idea of group coding allows us to interpret the inherent 'sloppiness' of the responses of cells, their tendency to fire within various degrees of the presentation of the right sort of stimulus, into a positive virtue. Nor is it an unlikely theory: as McCrone says, there is support for the notion that the brain actually does make use of the individually ambiguous firings of cells across whole populations. By recording from many different cells in the motor cortex of a monkey as it carried out some specific motor task for which it had been trained, such as keeping its eyes focused upon a particular spot on a screen or reaching out to touch a spot, researchers found that individually, the firing of cells was approximate, but the overall firing across the populations measured matched the actions the monkey was performing.[44] Population coding also can explain how it can be possible that our actual perceptual abilities are more precise than can be accounted for by the properties of individual cells. McCrone says:

> ...astonishingly, the brain does much better at detecting sensations that the receptive field properties of its neurons would seem to allow. In standard psychophysical tests, the human eye can pick out very tiny details. We are sensitive to the faintest bend in a straight line, or the slightest gap between two almost touching balls. It would seem that to be able to code for such a gap, we would need to use at least three retinal cells: one to represent the gap, and one to represent the space in the middle. Yet experiments show we can see a gap five times thinner than the shadow it casts across a single receptor on the back of our eyeballs. Our vision far out-performs the physical dimensions of our light detecting equipment.

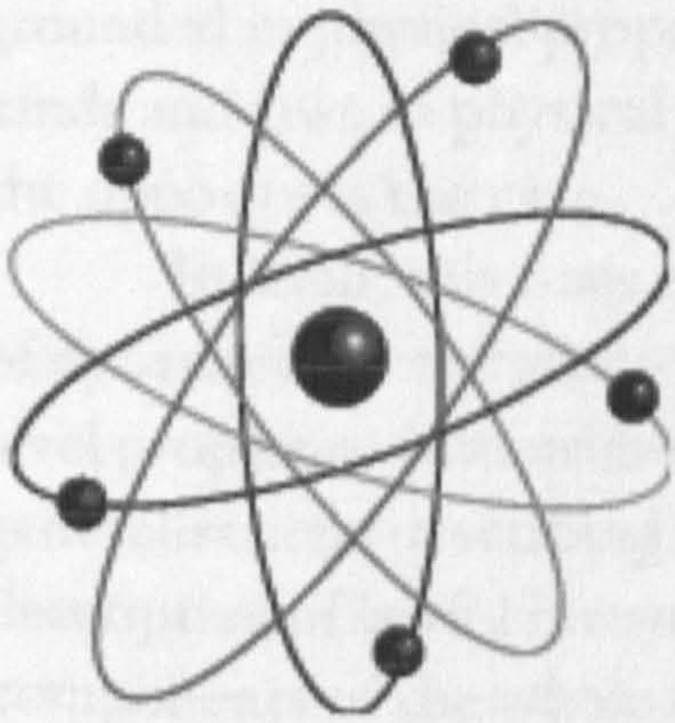
On the group coding account, it is not necessary that individual cells be as precise and as accurate as they would have to be to account for human perceptual abilities as described above, for this level of accuracy emerges 'for free' in the interactions of entire populations

of cells.

Population coding thus enables the whole brain to achieve feats of precision well beyond the abilities of its component neurons. It also generates robustness in the representational capacities of the brain: because patterns in whole populations of cells are involved in the representation of a complex object, the whole brain can still be able to represent the object even if parts of the brain that would normally be involved are damaged or missing. People who are achromatopsic because of a lesion in V4, though none of the cells in V4 contributes to their representation of Grandmother, will still recognise her, albeit without the rich colour detail that cells in V4 contribute, because the remaining population is sufficient to realise the pattern that is associated with Grandmother. It also allows for degrees of certainty: the greater or lesser degree to which a population coding for a complex object did so, as against the degree to which another pattern or no pattern in particular obtained, could account for the uncertainty we can experience in identifying objects we encounter in situations where they are hard to pick out: Grandmother encountered in a dark room with her face turned away; or Grandmother in a taxi going by at a fair clip.

Population coding is a property that inheres in the dynamic actions and interactions of cells in the brain each individually following their own very specialised instructions. It enables the brain to represent possible infinitude of different complex objects or states of affairs; to detect subtleties far beyond the capacities of individual neurons; to plan and command complex motor actions far outreaching the ability of individual fuzzy nerve firings. Such functions as the representation of an object or the planning of a movement are dynamically emergent properties, arising out of the complex actions and interactions of individual simple components. Thus the observations earlier made for emergent properties generally can be carried over to the brain's capacity to represent and to act: representational properties are very real, robust properties that enable the whole brain and body to represent and recognise aspects of the world, and to act upon and interact with the environment in ways that the individual components realising it never could. The abilities of the brain, and its structure and dynamics, are vastly beyond the properties and organisation of ant and termite societies, but the same observations that can be made for these much simpler, easier understood examples carry over to the case of the brain: if the brain is a complex dynamic system instantiating dynamically emergent properties, then these robust and real properties merit treatment in their own terms and under the generalisations subsuming them at the higher level at which they exist. Further, the presence of a dynamically emergent property demonstrably constrains the behaviour of the individual components realising it. If a representation in the brain is understood as a dynamically emergent property, we can begin to see how the representation, *qua* higher level property, can be a genuinely causally effective aspect of the whole system in which it is instantiated, in a way that is unproblematic from a physicalist perspective. Mental causation and the autonomy of psychology might be possible under a robust physicalism after all.

Conclusion



IF WE WISH SERIOUSLY TO ENTERTAIN A PHYSICALIST view about the relation of mental properties, among other kinds of properties, to basic physical properties, we must seriously accept the consequences of this view. In assessing the central tenets of the physicalist thesis, I hope to have exposed what those consequences are, and also to have identified some associated notions from which serious physicalists need not suffer. Physicalists do not have to provide an *a priori* account of what properties are physical, or hard and fast criteria for the identification of physical properties. 'Physicality' can be understood as a contingent feature something can possess in virtue of its fitting within a closed causal/explanatory hierarchy. Such an account allows that the development of the physical sciences in ways that we are unable to anticipate will not falsify physicalism by identifying some properties we currently cannot explain in physical terms as physically explicable after all. Neither does the physicalist have to be constrained by any particular elucidation of what *physics* is: we do think, after all, that physicalism could be true even if there were no physicists. This is, I believe, a perfectly workable account of what it is to be physical, which does not involve physicalists in some of the difficulties that have been identified for the formulation of the doctrine.

The requirements that all properties be related to physical properties in a metaphysically explanatory way, and the principle of the causal closure of the physical domain, do entail that physical laws underlie all causal transactions on any level of description. For any actions we take, any decisions we make, and any thoughts we have about anything, there exists a complete sufficient description in physical terms. But serious physicalists do not have to accept that physical properties are the only kinds of properties that there are, or that physical laws are the only kinds of laws. Physicalism can recognise and tolerate properties that are not straightforwardly reducible to physical kinds, and physicalists need not require that laws in the special sciences be reducible to physical laws. On the functional model of reductionism, what is central to the reducibility of any particular nonbasic property is that it can be construed extrinsically as a functional property, and that some physical system can be identified in the

microstructure of whatever instantiates the property that is sufficient to perform that function. More basic laws will describe the behaviour of the microstructure involved, and eventually, basic physical laws will describe the dynamics of the basic physical structure, or as close to that as can be identified. At the level of abstraction at which higher level properties are considered as intrinsic properties of entities, meaningful generalisations can be constructed describing the behaviour of higher level multiply instantiable properties, but below that level, these generalisations are incoherent. Nonetheless, physical laws can be seen to ground special science laws, by being descriptive of the low level dynamics that occur in any given higher level causal interaction. All that physicalism taking the functional model of reductionism as providing the criteria according to which any given higher level property is physically reducible requires is that particular instantiations of higher level properties and special science generalisations should be grounded in physical properties and physical laws. The reducibility of special science kinds and laws to physical kinds and laws is neither required nor desirable; indeed quite the opposite is the case.

In itself, this does not solve the problem of causal exclusion. However, the notion of dynamically emergence provides us with a way of thinking about some kinds of higher level properties instantiated in dynamic systems as robust, very real properties, and of the generalisations describing the behaviour of whole dynamic systems as genuinely descriptive of lawful interactions: because the actions and interactions of the simple components of the whole system enable the whole to act in and on its environment in complex ways for which no coding directly exists, basic physical explanations of causal transactions involving such systems, though they may sufficiently describe the transactions, can be seen to miss a very important and very real feature of dynamic systems. Since dynamically emergent properties pervade the system down to the lowest complete components, constraining the activity of the components in ways for which the components alone have no intrinsic leanings, and because dynamically emergent properties can enable whole systems to interact with their environments in apparently intelligent ways, we can think of the robust higher-level dynamically emergent property as having a real causal influence upon the world *qua* higher-level property, without invoking downward causation. It also shows us how the very complex abilities of a system to act in and interact with its environment can be bought very cheaply in the collective actions and interactions of very simple components. Robust, functional abilities of whole systems can emerge unproblematically in the collective dynamics of simple components, any single one of which could not manifest the capability the whole system exhibits, and without requiring that complex coding for higher level complex abilities be 'coded for' anywhere. Dynamic emergence is thus very interesting from a philosophical perspective, providing us with a framework in which we can think of some kinds of higher level properties as real features in the physical world, and really causally effective, considered in their own right. If, as many researchers believe, the brain utilises the collective dynamics of entire populations of cells to code for complex abilities such as the representation of particular objects or the planning of very precise movements, or indeed any of the great many things that brains do, we can begin to appreciate how higher-level

properties realised in the brain can drive the actions of the whole system in which those properties inhere. We can begin to see how the complex capabilities of the dynamic brain – mental states – can enable the brain and body to carve trajectories through the world in ways unanticipated by the capacities of its simple components. We can begin to see how happy intellection, or indeed graceful error, may enable the whole brain mindfully to correct its cave.

Functional reductionism and the notion of dynamic emergence may help us to see how mental states, *qua* mental, can be construed realistically as causally effective in themselves, and how psychological generalisations, though irreducible to physical laws, are meaningfully expressive of real dynamics. But what to my mind is the greatest challenge to physicalism remains unaddressed. The human organ of thought is a wonderfully complex and amazing thing, so much so that we are only beginning properly to understand just how wonderful and amazing it is, how subtle and powerful its functioning is. But nothing that has been said can provide any insight into why any of the subtle functionings of the brain should be associated with conscious, qualitative experience. It is impossible to see how a quale could be construed extrinsically as a functional property, and so functional reductionism is silent on whether or not qualia are reducible to the physical or not; no mechanism can be identified that is sufficient for the realisation of a quale. The phenomenon of consciousness thus remains just as much a challenge to physicalism as ever it was. Though the notion of dynamic emergence provides us with a fascinating insight into the ways mental states, insofar as they can be considered representational higher level states in the brain, can be genuinely causally effective aspects of the world, it cannot illuminate why so many of them feel quite the way they do. While the cave may be corrected, it remains in darkness. Much work is yet to be done.

Notes to the Introduction

[1] T. Crane and D. H. Mellor, "There Is No Question of Physicalism," *Mind* 99 (1990)
[2] Crane and Mellor (1990) p. 186.
[3] C. Daly, "What Are Physical Properties?," *Pacific Philosophical Quarterly* 78 (1998)
[4] Daly (1998): p. 196.
[5] J. Kim, Mind in a Physical World (Cambridge, Massachusetts: The MIT Press, 1998)
[6] J. Kim, "'Downward Causation' in Emergentism and Nonreductive Physicalism" in Beckermann, Flohr, and Kim (eds.) Emergence or Reduction?: Essays on the Prospects of Nonreductive Physicalism (Berlin and New York: De Gruyter, 1992)

Notes to Chapter 1

[1] Tim Crane and D. H. Mellor, "There Is No Question of Physicalism," *Mind* 99 (1990): 186.
[2] Crane and Mellor (1990) op. cit. p.186
[3] Crane and Mellor (1990) op. cit. p. 206
[4] Crane and Mellor (1990) op. cit. p. 185
[5] Crane and Mellor (1990) op. cit. p. 185
[6] Crane and Mellor (1990) op. cit. p. 189
[7] Crane and Mellor (1990) op. cit. p. 185
[8] Crane and Mellor 1990, op. cit. p. 185
[9] I expressly do not wish to say that the physicalist will claim that for all special science terms there are coextensive terms in the physical sciences, for this is precisely a point on which eliminativist and noneliminative physicalists will diverge in opinion. Hence the weaker claim.
[10] Crane and Mellor (1990) op. cit. p. 187
[11] Crane and Mellor (1990) op. cit. p. 186
[12] Here is a further problem:: conceivably, it could turn out to be the case that what we currently think of as our most basic physics is, from the point of view of a more advanced physics, a special science. It is further conceivable that what we currently think of as our most basic physics is irreducible to this putative more advanced physics. Would that indicate that those properties and entities we now consider to be quintessentially physical and those things to which terms in our most basic physical sciences refer are nonphysical and, under eliminativist interpretation, not really *real*? In that case, how could we know of anything at all whether or not it is really real? To my mind, this thinking shows that eliminativism, as an epistemological doctrine, is essentially sceptical, though this may not bother some eliminativists.
[13] ...which is not to say that all physically possible sciences are physical: there may be perfectly reasonable and informative sciences that aren't, which fact would not necessarily imply that physicalism is false, but only that on a certain level of description, observable phenomena and any subsuming generalisations are not *reducible* to physical phenomena and laws, while they may still be *backed* by them. More on this later.
[14] Arguably, some entities and phenomena may count as nonphysical because there is nothing physically interesting about them, or the generalisations that subsume them may not be backed by predictively interesting physical laws.
[15] Crane and Mellor (1990) op. cit. p. 188
[16] Shu-Yuan Chu of the University of California theorised that gravity and quantum physics could be unified within superstring theory, an idea which he was developing at the time of his death in 1998. Superstring theory is, as far as I am aware, hardly orthodox in basic physics, but it does have a substantial following.
[17] "...even if [physics] is not [unified], physicists will still accept gravity, quantum and electromagnetic phenomena as physical, to be identified and described in their own terms by independent physical sciences." Crane and Mellor 1990, op. cit. pg. 188. The wording here would seem to preclude the reconciliation of quantum mechanics with gravity as described by Einstein's relativity theory, as well as the subsumption of quantum, gravitational and electromagnetic phenomena under some more comprehensive development in basic physics. But it is not obvious that this cannot be done (see previous note). It is undoubtedly possible, if not probable, that were a more comprehensive physics to be developed, the sciences currently employed would not be abandoned: they are useful. But in that case I should think that the sciences subsumed would be used and interpreted instrumentally by physicists, the results they obtained through the practice of these sciences explainable and describable in terms of the more comprehensive physics, and in that sense, unified *enough*.
[18] Crane and Mellor (1990) op. cit. p. 189
[19] Crane and Mellor (1990) op. cit. p. 189
[20] In fact it seems to me that here Crane and Mellor are conflating two slightly different species of reductionism, microreductionism and methodological reductionism. Methodological reductionism is indeed the doctrine that, in the sciences, the smaller the entities postulated in meaningful explanations of phenomena on any level, the better. Unlike microreductionism, the methodological reductionist may stop short of saying that everything must be reducible to compounds of one sort of extremely small type of object — she may only wish to claim that, within a particular domain — the sociobiological, say — everything can be understood in terms of a certain small component factor — the gene, say. She may not wish to carry the reduction farther down, for the very good reason that to do so would be unenlightening vis à vis sociobiology. The microreductionist, on the other hand, is concerned to show that all macrophysical objects, properties, and states are reducible to microphysical, ultimately basic, objects, properties and states. As such the microreductionist may well have to give up 'small is beautiful' is empirical science or other considerations indicate that phenomena with more generous boundaries are component determinants of apparently tiny phenomena.
[21] T. Horgan, *Blackwell: A Companion to Metaphysics*, 1995 K. Kim and E. Sosa, eds., s.v. "Reduction, reductionism," 438-439.
[22] 'Space' springs to mind.
[23] Crane and Mellor (1990) op. cit. p. 190.
[24] Crane and Mellor (1990) op. cit. p. 190.
[25] John Gribbin is an astrophysicist and prolific popular writer on cutting edge physics. The work to which I referred is his encyclopaedic Q is for Quantum: Particle Physics from A to Z (London: Phoenix (a division of Orion Books), 1998), entries on Superposition, pp 477-478, and Mach's Principle, pp 265-268.
[26] The relation is as far as I understand taken to be causal: just as the first quanta is genuinely *neither* up nor down until it is measured, but is in a superposition of up-ness and down-ness, so neither is its partner. The measurement of one particle causes the collapse of both into their respective states.
[27] The standard interpretation of such ensembles is, according to Crane and Mellor, that they are complete, and completeness generally does imply closure. This interpretation might be suspect anyway in terms of quantum mechanics, for quanta are supposed to be able to wink in and out of existence; they have a certain probability of spontaneous and instantaneous transport to the other side of the galaxy, &c. With such properties as these I do not see how it would be possible once and for all to declare any quantum ensemble closed within a universe which featured other such ensembles (or potentialities for such). Thus, if the standard interpretation is indeed that such ensembles or closed, I cannot help but think that it must be with some caveats.
[28] Gribbin 1998, op cit. pp 265-268
[29] Crane and Mellor (1990) op. cit. p. 191
[30] C. G. Hempel, Philosophy of Natural Science, NJ: Prentice-Hall, 1966

[39] E. Nagel, 1961. The Structure of Science. New York: Harcourt, 1961
[40] Jaegwon Kim has developed such a model (Kim 1998), which will be discussed in detail later on.
[41] See Chapter 3, especially subsections *vi* and *vii*.
[42] F. Brentano, Psychology from an Empirical Standpoint 1874
[43] Crane and Mellor (1990) op. cit. p. 193
[44] Crane and Mellor (1990) op. cit. p. 194
[45] Hilary Putnam, (1975), "The Meaning of 'Meaning'", in Keith Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis, University of Minnesota Press, 1975
[46] Jaegwon Kim, Mind in a Physical World (Cambridge, Massachusetts: The MIT Press, 1998) 37.
[47] Putnam (1975)
[48] Crane and Mellor (1990) op. cit. p. 194. I cannot help but think that there is a little bit of circularity here: aren't *models* and *images* contentful in themselves?
[49] Crane and Mellor (1990) op. cit. pp. 191-192.
[50] Crane and Mellor (1990) op. cit. p. 192
[51] Crane and Mellor (1990) op. cit. p. 195.
[52] Crane and Mellor (1990) op. cit. p. 195.
[53] Not all animals have both: Dr. Jonathan Sheps informs me that there is a nematode worm that has a kidney, "At least if by kidney [is meant] the various organs which at one time or another are employed by the various animal phyla to cleanse the blood of nitrogenous wastes and to direct these same effluvia outwith the body with minimal loss of water" (Dr. Sheps, via e-mail, 8th February 2001), but which has no organ performing the function that a heart performs in other animals. Presumably, this beastie is so constituted that it does not need a heart, where that is meant to indicate any organ responsible for moving blood about the body.
[54] Perhaps the arggument could be more clearly put as follows: I agree that if these two statements are laws:
(H--> Heart)
(K--> Kidney)
and that if it is true and just simply a fact, not a law, that all mammals have both hearts and kidneys, neither of these
(H--> Kidney)
(K-->Heart)
are implied. However, if it is not a law that all mammals have both hearts and kidneys, then there is no law to create a nonextensional context, as Cranes' and Mellor's arugument requires that there be. However, if it is a law that all mammals have both hearts and kidneys, so that all these are true:
(H-<->Heart)
(K<--Kidney)
(H<-->K)
then no nonextensional context is created: if something has H, it has K, so if something has H, it has a kidney. Thus the substitution 'Mammal with a heart' can truthfully be substituted for 'Mammal with a kidney' in extensional contexts. I use the biconditional in all three statements to discount the possibility that something other than H could be responsible for hearts and something other than K for kidneys; if some H' could produce hearts in mammals, then obviously the example wouldn't work unless (H'<-->K) were a law too.
[55] Crane and Mellor (1990) op. cit. p. 195
[56] Crane and Mellor (1990) op. cit. p. 196
[57] Crane and Mellor 1990: op. cit. p 197. The embedded quotation is from C. McGinn, "Mental States, Natural Kinds, and Psychophysical Laws ", *Proceedings of the Aristotelian Society Supplementary Volume*, 1978, p. 197.(by interesting coincidence).
[58] Crane and Mellor (1990) op. cit. p. 203
[59] Hilary Putnam, (1975), "The Meaning of 'Meaning'", in Keith Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis, University of Minnesota Press, 1975
[60] Crane and Mellor (1990) op. cit. p. 205
[61] I recall an exercise in quantum mechanics given to students by Dr. Jim Mahoney of Marlboro College, Marlboro, Vt, described to me by Brooks Walsh: students were asked to calculate the likelihood of a co-ordination of random quantum events transporting them entire from the Science Building to the Dining Hall. The answer indicated that this might probably happen once in many, many times the estimated age of the universe, some 15 or 16 billion years or so: the chances of it occurring were vanishingly small.

Notes to Chapter 2

[1] C. Daly, "What Are Physical Properties?," *Pacific Philosophical Quarterly* 79 (1998)
[2] Daly (1998) op. cit. p. 196-199.
[3] Daly (1998): abstracted from pg. 198.
[4] I've replaced Daly's term 'individual' with 'objects, properties, and states' so as to be able to accommodate functional properties and states of affairs not easily characterised as individuals such as *orbits*, *flocking behaviour* and such.
[5] Daly (1998) op. cit. p. 199
[6] Thanks to Steven Hanlon for helpful clarification of this point.
[7] Frank Jackson made the suggestion in his John Locke lectures at Oxford University, Trinity term 1995, crediting the idea to Papineau, according to Daly's footnote #8 pg. 214.
[8] David Papineau, Philosophical Naturalism (Oxford: Basil Blackwell, 1993)
[9] G. Hellman, "Determination and Logical Truth," *Journal of Philosophy* 82 (1985): 607-618.
[10] J. Poland, Physicalism: The Philosophical Foundations (Oxford: Oxford University Press, 1994)
[11] Daly (1998) op. cit. p. 200.
[12] Daly (1998) op. cit. p. 200, referring to T. Kuhn's The Essential Tension, 1977, ppg. 309-313.
[13] Daly (1998) op. cit. p. 200
[14] J. Levine, "Materialism and Qualia: The Explanatory Gap," Pacific Philosophical Quarterly 64 (1983), and "On Leaving Out What It's Like", in M. Davies and G. Humphreys, eds., Consciousness (Cambridge, Massachusetts: Blackwell, 1993).
[15] T. Nagel, "What Is It Like to Be a Bat?," The Philosophical Review 83 (1974).
[16] Snowdon, 'On Formulating Materialism and Dualism', in Cause, Mind, and Reality: Essays in Honour of C. B. Martin, Heil, ed., Kluwer Academic Press, Dordrecht 1989, pg. 153
[17] Snowdon, 153. I have quoted slightly farther than Daly: Daly omits the sentence "The discovery (if such is possible) of the essence of this kind. is a posteriori".
[18] Daly (1998) op. cit. p. 204, raises specific difficulties for the notion of a property's being physical if it exists in physical space. Physical space, he notes, must be distinguished from other kinds of space (the quality space of the colour circle and the 'space' formed by the one-dimensional ordering of the determinates of a determinable quantity such as *temperature* are his examples, though I think that there could be some argument about whether such 'spaces' are really *space*), in which case a principled definition of what *space* is needs to be given. If Snowdon takes a relational notion of space, then part of what it is to be a physical object is to exist in a physical space which is itself defined in terms of relations between physical objects. If on the contrary an absolute notion of space is taken, then part of what it will be to be a physical object will involve being located in another physical object. Moreover, if physical space is a physical object, then there is at least one physical object which is not in physical space.
[19] The attempt to set appropriate limitations is faced with the same difficulties encountered above, for without a preconceived notion of which basic and shared essential features can count as physical, it would be very difficult to say which of the properties ghosts and apparitions might exhibit count as the sorts of properties shared with all other things, and which separate ghosts and apparitions from physical things
[20] Snowdon 1989, op. cit. p. 152
[21] See especially his *section 10: Varieties of Materialism*, ppg. 154-156, in which he speaks of nonreductive physicalism affirming that "mental states of affairs are physical in the sense that they are such as can obtain in a purely physical domain..."
[22] Daly, *What Are Physical Properties?*, op. cit. p. 205.
[23] Snowdon (1989) op. cit. p. 138
[24] Snowdon (1989) op. cit. p. 140
[25] Snowdon (1989) op. cit. p. 152.
[26] Snowdon (1989) op. cit. p. 152.
[27] Snowdon (1989) op. cit. p. 152.
[28] Daly (1998) op. cit. p. 205
[29] He does: Snowdon (1989) op. cit. p. 141.
[30] Snowdon does discuss the constitution relation at some length, but in my opinion he does little to meet this objection: pp. 140-143.
[31] David Papineau, Philosophical Naturalism (Oxford: Blackwell, 1993)
[32] Papineau (1993) op. cit. p. 10. He says: "Supervenience on the physical means that two systems cannot differ chemically, or biologically, or psychologically, or whatever, without differing physically; or, to put it the other way round, if two systems are physically identical, then they must also be chemically identical, biologically identical, psychologically identical, and so on."
[33] See especially page Papineau (1993) p. 24.
[34] Papineau (1993) op. cit. p. 13
[35] See especially his section on Realisation for discussion of causal overdetermination : Papineau (1993) pp. 23-27.
[36] Papineau (1993) op. cit. p. 16
[37] Papineau, op. cit., ppg. 17-18. Note that this principle is vulnerable to arguments such as those David Chalmers has forwarded about the possibility of absent, inverted, or dancing qualia. If such arguments are consistent, then there is at least one type of mental occurrence — qualitative occurrences — whose differences would not manifest themselves in any physical context.
[38] J. Kim, "The Nonreductivist's Troubles with Mental Causation" in J. Kim: Supervenience and Mind: Selected Philosophical Essays (Cambridge: Cambridge University Press, 1993) 348.
[39] S. Alexander, Space, Time, and Deity, volume 2, p 8, quoted in Kim 1993. op cit. 348
[40] Except, perhaps, for whatever dynamics there may be internal to the electron, but on the level of electrons, no causal transactions take place.
[41] Papineau considers the plausibility of token congruence without supervenience: Papineau (1993) pp. 28-29.
[42] This is not uncontroversial: see Kim (1989): *Mechanism, Purpose and Explanatory Exclusion*, reprinted in Kim, Supervenience and Mind: Selected Philosophical Essays, Cambridge: Cambridge University Press, 1993.
[43] Papineau (1993) op. cit. p. 24. Papineau uses the term 'state' to be a 'stylistic variation' on 'properties': footnote #22, pg. 25.

[44] Papineau (1993) op. cit. pp. 29-30
[45] Papineau (1993) op. cit. p. 30
[46] Daly notes that such an approach has been taken before by Fiegl (1958), who differentiated between two uses of the term 'physical': physical1, having something to do with space, time and causality; and physical2, which is applied to 'the type of concepts and laws which suffice in principle for inorganic processes'.
[47] Such a move would also trivialise the physicalism/dualism debate for the actual world, as I shall explain below.
[48] The relations of determinate to determinable is transitive, and so the implication of a new physical determinable would imply that the determinable 'physical' would then have to include this new determinable, from which the newly discovered determinates emanate.
[49] Someone has termed such possible physical properties 'alien', but unfortunately I cannot remember who it was.
[50] Papineau (1993) op. cit. p. 29-30
[51] Daly (1998) op. cit. p. 208
[52] Daly (1998) op. cit. p. 208
[53] Papineau (1993) op. cit. p. 30
[54] See especially Papineau op. cit. ppg 29-32.
[55] Papineau (1993) op. cit. p. 30
[56] Daly (1998) op. cit. p. 208
[57] Daly (1998) op. cit. p. 204
[58] Before the development of subatomic physics, many chemical properties were taken to be brutely emergent and inexplicable in terms of physics.
[59] Hellman and Thompson, "Physicalism: Ontology, Determination, and Reduction," *Journal of Philosophy* 72 (1975)
[60] Hellman, "Determination and Logical Truth," Journal of Philosophy 82 (1985)
[61] Hellman (1989) op. cit. p. 610
[62] Hellman (1989) op. cit. p. 609
[63] Daly (1998) op. cit. p. 210
[64] J. Poland, Physicalism: The Philosophical Foundations (New York: Oxford University Press, 1994) 13.
[65] Poland (1994) op. cit. p. 11
[66] Poland (1994) op. cit. p. 18
[67] Poland (1994) op. cit. p. 19
[68] Poland (1994) op. cit. p. 20
[69] Poland (1994) op. cit. p. 20
[70] See Poland (1994) p 21, pp 22-23
[71] It might be assumed that these relations must always be to lower level branches of science, but this is not necessarily the case. Vertical explanation can proceed via sciences on the same level or even via sciences at higher levels: it might be the case, for example, that there are explanations in folk psychology for the behaviour of some individual human being. Relating folk psychology to physics, however, might have to proceed partly by way of sociology. Though the point is contestable, it seems that the science of individuals is lower in the scientific hierarchy than the science of social groups.
[72] Poland (1994) op. cit. p. 124
[73] Poland (1994) op. cit. p. 124
[74] With deference to 'the intentional stance' as described by D. Dennet, which can be applied to almost anything: an electron's being attracted to a proton might be describes as the electron's wanting to be where the proton is. However, such a stance cannot be taken by an agent which is not itself intentional, so again, we never find phenomena apt for psychological descriptions in the absence of human beings (or other intentional creatures).
[75] Poland denies that this is uniquely a feature of physics, and so physics cannot simply be characterised as 'the most universal or general science': see ppg. 120-121.
[76] The objection could be raised that, as explored earlier in this paper, microscopic systems are often not closed within the same spatiotemporal boundaries as the macroscopic objects they are taken as constituting. However there is nothing in the notion of vertical explanation that suggests that they must be so closed, and I do not see that this objection carries for the idea of vertical explanation in Poland's sense.
[77] Poland (1994) op. cit. p. 125
[78] Poland (1994) op. cit. p. 125
[79] Whether or not astronomy and thermodynamics do count as physical sciences, why they do, and why other sciences do not, is a further question. If all sciences are vertically related to physics, then how do some sciences, such as chemistry or ballistics, qualify as physical sciences whereas other sciences, such as psychology or virology, do not? I think there is certainly room for a principled answer, and that such an answer is most likely to come from the notion of vertical explanation itself. A world-specific method of identifying some sciences as physical and excluding others from the qualification could be given as follows: a science is a physical science just in case its proprietary objects and principles are such that they are directly related to basic physics in ways that are interesting from the perspective of basic physics. For example, chemistry will count as a physical science according to this view, because its proprietary entities and principles — molecules, atoms, electrical forces and their dynamics, &c — are uniquely realised by particular arrangements of subatomic and quantum phenomena, and by no other things. Chemical properties are not multiply realisable. Nothing is a water molecule that is not also a collection of hydrogen and oxygen in a particular relation, and also a collection of protons, neutrons and electrons in a particular relation, and so on down to the most basic elements physics can describe. If a science has proprietary elements or principles that have some features or aspects that are not interesting from a physical perspective — as, presumably, virology has, for some of the properties of viruses could be instantiated by one or another physical structure — then that science is not a physical science, though it is vertically related to physics. Such a method for identifying physical sciences does, incidentally, draw upon an important distinction between kinds of properties that could usefully elucidate the difference between physical and nonphysical properties without committing the system to any *a priori* constraints upon the notion of a physical property and while allowing that notion to be subject to revision as physics progresses. These themes will be explored in the final section of this paper in more detail.
[80] I refer again to the above cited passage: "Jeffrey Poland attempt to say which properties are physical by (1) saying what physics is, and then (2) saying what properties are physical properties in terms of saying what properties are described by physics."
[81] Poland (1994) op. cit. p. 118
[82] Poland (1994) op. cit. p. 118
[83] Poland (1994) op. cit. p. 118
[84] Poland (1994) op. cit. p. 41

Notes to Chapter 3

[1] Ernest LePore and Barry Loewer, "More on Making Mind Matter," *Philosophical Topics* 17 (1989): 177-178, quoted in Kim, below.
[2] Jaegwon Kim, "The Nonreductivist's Troubles with Mental Causation", reprinted in Kim (1993): Supervenience and Mind: Selected Philosophical Essays, Cambridge: Cambridge University Press, . 343.
[3] This does not mean that the physicalist must accept the realisation relation for mental states, or indeed any nonphysical properties. They may be happy with the realisation relation obtaining between higher level, but still fairly uncontentionably physical properties, such as the realisation of chemical properties by atomic. Yet they may still deny that a relation of this type obtains between the mental and the physical. Those who adhere to nonreductive physicalism would forward theses in this direction, as the realisation relation as above expressed implies reduction. The degree to which such theses coherently can be maintained will be a topic of discussion later in this work.
[4] This is because the physical instantiators of properties of this kind can be, and probably characteristically are, wildly heterogeneous. This will be discussed in more detail later.
[5] I include composition here because I believe that the notion that composed entities that are somehow 'more than the sum of their parts', when the relation between the parts is included in that sum, is inconsistent with the principle of the causal closure of the physical domain; this will be argued in a later section.
[6] Papineau (1993) op. cit. p. 16
[7] Papineau (1993) op. cit. p. 13. His italics.
[8] *Physics*, of course, can't be causally *anything*; physics is an explanatory system, or a *theory*, and beyond the entailment of some statements in a theory by others, I can think of no causation that goes on within *theories*.
[9] The principle of the causal closure of the physical domain specifies only that the causes of physical events must be physical. It does not imply the converse, that ox events can only have physical effect. It is consustent with the closure prnciple as I interpret it that there be epiphenomena caused by, or at least consistenly associated with, physical phenomena, so long as phenomena proprietary to substantial realms outside the bounds of the physical have no effect upon the physical domain.
[10] Crane and Mellor (1990) op. cit. p. 192
[11] David Hume, Enquiries, Reprinted from the 1777 edition with introduction and analytical index by L. A. Selby, 3rd ed. (Oxford: Oxford University Press, 1994) 86-87.
[12] It may be objected that quantum randomness falsifies predictions in basic physics. That, however, would be incorrect. The stochastic nature of quantum dynamics is a perfectly well-behaved feature of the quantum domain, which feature is reflected in the laws of quantum physics. If a quantum prediction goes 'x will y with 99.998% probability, and then if x fails to y, there is a perfectly good reason for this: this particular occasion was one of the .002% of all x-y cases where x does not y.
[13] It is conceivable that there may be no ultimate basic level of physics, or, that all entities are infinitely decomposable. This does not affect the central point being made, however. If it is the case that there is no basic ontological level, then it will be possible that there are no levels on which laws are exceptionless, but it will still be the case that in order to make sense of explanations on any level, falsifications of generalisations by inputs from higher levels must be ruled out.
[14] The author believes that the English language lost a wonderfully expressive term, which does not have a suitable modern equivalent, when 'weird' lost this sense.
[15] Thanks to Sydney Harris for the kind permission for the use of his cartoon, which so aptly illustrates this very point.
[16] In general, supervenience is taken as expressing a relation of determination of the supervenient properties by the subvenient properties and of dependence of the supervenient properties upon the subvenient properties, though as we have seen, the supervenience relation *simpliciter* is not expressive of such asymmetry. Nonetheless, it is useful in expressing asymmetric property relations, and so I shall here follow the convention of taking it as being used to do precisely that.
[17] J. Kim (1987) "'Strong' and 'global' supervenience revisited" in J. Kim (1993): Supervenience and Mind: Selected Philosophical Essays, Cambridge: Cambridge University Press, p 79.
[18] Kim (1987) op. cit. p. 80
[19] It should be pointed out that, though functionalism and supervenience are typically advocated by physicalists, and that though discussions of functionalism and supervenience typically occur in the context of discussions about physicalism, neither functionalism nor supervenience are physicalist theses in essence.
[20] I refer here to a relation between physical properties and psychological rather than mental properties, because it is standardly taken that in descriptions of supervenience like this one, a) the properties F and G are taken to be intrinsic properties and b) *mental* states, such as are picked out by content, do not supervene on intrinsic properties alone, but involve relational properties, *vide* the relation of an individual to factors in that individual's environment. *Psychological* properties, in the sense I wish to use the term, are those properties of an individual's mentality that do not depend on content: they might be normatives, such as 'if you believe p, do not at the same time believe that -p', or other non-content dependent properties of mind, as the syntactical elements of the language of thought would have to be, or the concepts of space-time and cause would have to be for Kant. As mentioned in previous discussion, it is possible to formulate supervenience to accommodate relational properties, and there is no immediate reason beyond convention to assume that G is always an intrinsic property, but I abide by this convention for clarity.
[21] J. Kim (1984) "Concepts of Supervenience" in J. Kim (1993): Supervenience and Mind: Selected Philosophical Essays, Cambridge: Cambridge University Press, pp. 58-60.
[22] I disallow the possibility that this person could be female, because they are supposed to possess all of St. Francis's intrinsic properties, including *being male*.
[23] The author actually has no idea how tall St. Francis was.
[24] Kim (1987), op cit. p 80. The italics are his.
[25] S. Blackburn, "Supervenience Revisited," in I. Hacking (ed), Exercises in Analysis (Cambridge: Cambridge University Press, 1985)
[26] Mark Rowlands, Supervenience and Materialism (Hants, England and Brookfield, VT: Ashgate Publishing Company, 1995) p. 10.
[27] There are worlds of the form G/FvO where G is a B-minimal property realising F within a system relative to that system. In such cases, the expression of the supervenience of F upon G will have to specify the system description in which it does so.

[44] Rowlands (1995) op. cit. p. 11. His emphasis
[45] Rowlands (1995) p. 11
[46] Rowlands (1995) op. cit. p.12
[47] See especially Rowlands (1995) op. cit. pp. 18-20.
[48] Rowlands (1995) op. cit. p. 11
[49] J. Kim, Mind in a Physical World (Cambridge, Massachusetts: A Bradford Book, The MIT Press) 1998 p. 14. Emphasis on the last clause mine.
[50] See for example Kim (1998), Chalmers (1993), Papineau (1993)
[51] The multiple realisability of mental properties, among other kinds of properties, now a generally accepted truism in the philosophy of mind, was originally brought to the attention of the community at large by Putnam , (1975)"The Nature of Mental States" in his Mind, language, and Reality (Cambridge: Cambridge University Press) and Fodor (Special Sciences) 1974 "Special sciences (or the disunity of science as a working hypothesis)", *Synthèse*,28, pp. 97-115
[52] Ernest Nagel, The Structure of Science (New York: Harcourt, Brace, 1961)
[53] Gresham's law: the principle that "bad money drives out good, i.e., when depreciated, mutilated, or debased coinage (or currency) is in concurrent circulation with money of high value in terms of precious metals, the good money is withdrawn from circulation by hoarders." —from the entry on Gresham, Sir Thomas, at http://www.encyclopedia.com. Another internet source gives it this way: "Basically, the law is that bad money drives out good. When more than one kind of money is in circulation (for example, gold and silver coins), the money which is overvalued at the official price will tend to remain in circulation because it is worth more as a medium of exchange than as bullion in the market. The money which is undervalued will disappear from circulation to be hoarded or to be melted down because it has higher value as metal than as legal tender." at xrefer, http://www.xrefer.com/entry/344050. The law is not Gresham's originally, but has had various statements throughout history. For a detailed exposition, see Robert Mundell's "Uses and Abuses of Gresham's Law in the History of Money" at http://www.columbia.edu/~ram15/grash.html. All sites functioning as of 15/2/01.
[54] J. Fodor, "Special Sciences," *Synthèse* 28 (1974): 103.
[55] J. Fodor, "Special Sciences," *Synthèse* 28 (1974): 103. A physical natural kind, according to Fodor, is one that features in laws of physical science: with caveats, "P is a natural kind predicate relative to S iff S contains proper laws of the form $Px \rightarrow \alpha x$ or $\alpha x \rightarrow Px$" -Fodor 1971, op cit. p 102.
[56] C.G.Hempel, P. Oppenheim. "Studies in the Logic of Explanation", 1948. In: *Readings in the Philosophy of Science*. Herbert Fiegl, u.a., Hg. New York 1953.
[57] Kim (1998) op. cit. p. 26
[58] Kim (1998), op. cit. especially pp. 91 and 97.
[59] D. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* Vol.2 Issue 1 (1995): pp. 200-219
[60] Kim (1998) op. cit. pp. 95-96
[61] Kim (1998) op. cit. p. 96
[62] Kim (1998) op. cit. p. 96. Italics mine.
[63] The functional model of reductionism is Kim's. I do, however, take what I perceive to be entailments of this model further than he describes, and acknowledge that he may not agree with everything I am about to say.
[64] Kim (1998) op. cit. p. 20
[65] Kim (1998) op. cit. p. 20
[66] Nagelian reductionism does not disallow this, but it would add another level of complexity to the required bridge laws, even further lessening the appeal of the model according to Kim's criteria.
[67] The phrase is Davidson's, from the opening sentence of his *Mental Events*, in: D. Davidson, Essays on Actions and Events, (Oxford: Clarendon Press, 1980)
[68] As far as I know, this is true.
[69] One might propose that those sciences are physical that pick out entities with only one possible realisation base in this world. This might be a viable proposition, but it must be taken on board that it will be subject to revision, if future microphysics reveals that there are entities lower than we currently suspect, and two or more arrangements of those entities turn out to be sufficient for the realisation of properties that we currently take to have a unique realisation base.
[70] I suspect that according to this notion of the relative physicality of properties, some kinds of properties, for example, economical properties, will appear to be less physical than psychological properties, by virtue of the fact that it is probable that such properties are more multiply instantiable than psychological properties as a matter of nomological necessity. It seems to me highly unlikely that, in the actual world, a functional brain could be constructed of beer cans and bubble gum , even in principle. Functionalism says that multiply instantiable higher level properties are to be defined not in terms of whatever bits and bobs constitute the system instantiating them, but in terms of whatever function those bits and bobs perform in that system. But this does not mean that just any old bits and bobs are of the right sort to perform that function. There may be, and there probably are, only a limited number of appropriate bits and structures realisable by those bits that are adequate to the task of performing certain functions in the actual world. The living body is a wonderfully complex and dynamic thing, the brain in particular being dazzling in its complexity, so much so that we are only just beginning properly to appreciate it for the natural wonder it is, or to understand most, let alone all, of the factors involved in the function it performs. Given that the laws of physics are as they are in the actual world, I think it highly unlikely that a system constructed of beer cans and bubble gum would be able to do what a brain can do. This is not to imply that there *can* be no other system in the actual world that can do what a brain can do — that 'ugly bags of mostly water', as a crystalline native of the planet designated by the Federation as Velera III termed humanity in a discussion with Captain Kirk over the destruction of a human terraforming operation on its home planet (Start Trek, 'Home Soil'), are the only possible mindful systems in the actual world. There might be other possible systems. If there are, that too is a contingent fact about this world and its population of properties and laws.

By contrast, anything can be money. Following the total undermining of the US economy and the abolishment of all currently recognised forms of American currency, some elected official could stick two beer cans together with a piece of gum, call it US$5, and if everyone agrees, it is. Does this mean that *being US $5* is less physical than *being the belief that Ralph Nader should have won the US 2000 election?* Probably not. Though brain states may be more limited in their instantiators than economical states by virtue of there being more stringent limitations on what can be a brain than on what can be money, *intentional states* will almost certainly remain more multiply instantiable than economical states, for there are myriad — probably unlimited — ways for dynamics in the brain to become associated with particular propositions. So rest assured, true-believers: if economical states are pretty nonphysical, your beliefs are even more so.

[11] J. Poland 1994. Physicalism: the Philosophical Foundations. (New York; Oxford: Clarendon Press; Oxford University Press, 1994) p. 11

[12] Poland (1994) op. cit. p.11

[13] Poland's footnote: "a 'vertical explanation' of some phenomenon is, roughly, an explanation of why it occurs in terms of lower level phenomena."

[14] Poland (1994) op. cit. pp. 12-13

[15] I anticipate that there might be some overlap between the physical or nonphysical status of properties cross-worlds: some or indeed all of the properties that are physical in this world might fit into the closed causal/explanatory hierarchy of another possible world, while that world might contain properties which are not physical in this world.

Notes to Chapter 4

[1] Another prominent form is Donald Davidson's Anomalous Monism, a full discussion of which is outwith the scope of this thesis.
[2] McLaughlin, "The Rise and Fall of British Emergentism" in Beckermann, Flohr, and Kim (eds.) Emergence or Reduction? Berlin and New York: De Gruyter 1992 pp. 49-93
[3] Kim, "Downward Causation and Emergence" in Beckermann, Flohr, and Kim (eds.) Emergence or Reduction? Berlin and New York: De Gruyter, 1992, p. 122
[4] It is not immediately inconsistent with emergentism that there be no ultimate ontology of any sort: it might be, as it were, emergents all the way down. Neither is it the case that the ultimate ontology has to be physical: Alexander, Kim notes, held that matter is emergent from space and time. However, the British Emergentists generally did hold that there were some basic, nonemergent entities and properties, and generally these were held to be physical entities and properties.
[5] Broad (1925), ch. 2 par. 35. Unfortunately, at the time of writing (May 2001) I did not have available a paper copy of Broad's *Mind and Its Place in Nature*. Fortunately, however, the bulk of the relevant parts of the text had been made available online at http://www.ditext.com/broad/mpn.html, which version I used. This version, of course, does not have *pages*, so I cannot give page references. For citations of this work, I adopt the convention of giving paragraph references. For those wishing to follow my references, the mention of some tricks are in order: 1) Chapter subheadings count as paragraphs. 2) Broad occasionally makes use of chemical diagrams. In the online version, these are accomplished with formatted text, and therefore count as several paragraphs. I have replaced the formatted text with single graphics, thus each chemical diagram counts as a single paragraph. 3) The beginning of the text of each chapter, if downloaded directly, contains a number of chapter and subchapter headings for browser navigational purposes. I have omitted these from the paragraph count, but included his quotation from Dickens. 4) For those importing the text into a word processor for automatic paragraph count, a false paragraph (a paragraph consisting of a single space) tends to appear following subheaders. I have deleted these.
[6] This nonpredictability is, it should be noted, not a merely practical impossibility, unless you include within the domain of practical considerations the amount of information it is possible to contain in the universe. Dr. Michael Charleston of Oxford tells me that, in chaotic systems, "...you have a completely deterministic system (like fractals, the weather, population models etc.) but because of its Sensitivity to Initial Conditions it becomes impossible to contain the complete state of the system in a finite universe. The precision required is infinite. This means that you can't predict the future state, though it is completely deterministic, more than a short time into the future." (Dr. Charleston, via ICQ, 24th April 2001 [ICQ, by ICQ Inc., is an application for communicating with other people over a network or the Internet])
[7] Qualitative properties are a notable exception.
[8] Kim (1992) op cit. p 131. Italics Kim's.
[9] Kim (1992) op cit. p 131
[10] Kim (1992) op cit. p 120. Italics Kim's.
[11] Kim (1992) op cit. p. 136
[12] McLaughlin (1992), op cit. p 57
[13] Phenomenological mental properties, Broad held, were emergent *a priori*.
[14] Broad (1925) ch. 2 par. 22
[15] Broad (1925) ch. 2 par. 38
[16] Broad (1925) ch. 2 par. 24
[17] Broad (1925) ch. 2 par. 25
[18] Broad (1925) ch. 2 par. 31
[19] Broad (1925) ch. 2 ppar. 25-28
[20] Broad (1925) ch. 2. par 29
[21] Broad (1925) ch. 2 par 30
[22] Broad (1925) ch. 2 par. 30
[23] Broad (1925) ch. 2 par. 58
[24] See Broad (1925) ch. 2 par. 10
[25] Broad (1925) ch. 2 par. 32
[26] Broad (1925) ch. 2 par. 32
[27] Broad (1925) ch. 2 par. 32
[28] Broad (1925) ch. 2 par. 32
[29] Broad (1925) ch. 2 par. 40. Broad says "No doubt the properties of silver-chloride are completely *determined* by those of silver and of chlorine; in the sense that whenever you have a whole composed of these two elements in certain proportions and relations you have something with the characteristic properties of silver-chloride, and that nothing has these properties except a whole composed in this way. But the law connecting the properties silver-chloride with those of silver and of chlorine with the structure of the compound is, so far as we know, a *unique* and *ultimate* law." Italics Broads.
[30] This is not, I believe, an issue with which Broad was directly concerned.
[31] Broad (1925) ch. 2 par. 42
[32] Broad (1925) ch.2 par. 59
[33] According to Brian McLaughlin, Broad never speaks of intra-ordinal laws as emergent: McLaughlin (1992), op. cit. p 81
[34] Kim (1992) op. cit. p 136
[35] Broad (1925) Ch. 2 par. 60
[36] Broad (1925) p. 436, quoted in McLaughlin (1992) p. 51
[37] C. Lloyd Morgan. *Emergent Evolution*. London: Williams and Norgate (1923), p. 35
[38] Broad (1925) Ch. 2 par. 60
[39] The reader might be interested to peruse C. Lloyd Morgan's "Lecture X: Causation and Causality" in his (1923) Emergent Evolution or Broad's Ch. 3, "The Traditional Problem of Body and Mind", in his (1925) Mind and Its Place in Nature.

Notes to Chapter 5

[1] David Deutsch, The Fabric of Reality, Penguin Books, 1998) p 30.
[2] Deutsch (1998) op.. cit. p. 21
[3] The caveat that chaotic or probabilistic laws may be involved is appreciated.
[4] Deutsch (1998) op. cit. p. 22
[5] Deutsch (1998) op. cit. p. 22
[6] Deutsch (1998) op. cit. p. 20
[7] Deutsch (1998) op. cit. pp. 22-23
[8] Brian Goodwin, "All for one, one for all", *New Scientist* no. 2138, 13th June 1998
[9] It is by no means restricted to this genus of ants. According to Octavio Miramontes, the oscillating activity of *Leptothorax* colonies "appears to be universal to the genus and possibly to to other social genuses as well". (Miramontes: "Order-Disorder Transitions in the Behaviour of Ant Societies" *Complexity* 1995 vol. 1 no. 3)
[10] Goodwin (1998), p 34. Actually this characterisation of emergence is slightly misleading: there is no principled reason why this behaviour could not theoretically at least be predicted, as will be clarified below.
[11] In all probability, it is of benefit to the colony, but as far as I am aware, precisely how it is so is currently imperfectly understood. Goodwin speculates that it may contribute to the successful raising of the brood. Ants will seek out unattended larvae and pupae to feed and groom, and if the activity of the worker ants is co-ordinated, the probability is greater that care will be distributed evenly over the brood, ensuring that a greater number of young reach adulthood. Miramontes further speculates that the maintaining of the population density at the boundary at which order emerges from the chaotic behaviour of individuals may enable the colony as a whole to adapt more readily to changes in its environment (Miramontes: "Order-Disorder Transitions in the Behaviour of Ant Societies" *Complexity* 1995 vol. 1 no. 3); the self-maintenance of ant colonies at this level is widely observed.
[12] Goodwin (1998) op. cit. p. 34
[13] Andy Clark, Being There: Putting Brain, Body and World Together Again (Cambridge, Massachusetts and London, England: The MIT Press, 1999) 75.
[14] Thanks to John Mullen for his kind permission for the use of his photography.
[15] Clark's note: "On many vessels there is in fact a formal plan. But crew members do not explicitly use it to structure their actions; indeed, Hutchins (1995, p. 178) suggests, the plan would not work even if they did." —Clark (1999), note 8, p.234
[16] Clark (1999) op. cit. p. 76
[17] See Clark (1999) ch. 1, 'Autonomous Agents', for a description of Attila and other similarly designed robots.
[18] Clark (1999) op. cit. pp. 73-74
[19] Clark (1999) op. cit. pp. 75-76
[20] B. Holldöbler and E.O. Wilson, Journey to the Ants (Cambridge, MA: Harvard University Press 1994) p. 107.
[21] Clark (1999)op. cit. p. 107
[22] Clark (1999)op. cit. p. 107
[23] Kim (1992) op. cit. p 120. Italics Kim's.
[24] I specify that dynamically emergent properties are present down to the level of the lowest *complete* components of the system because dynamically emergent properties can be layered, and while present in a sense, it is not obvious that dynamically emergent properties at the highest level are appropriately invoked to explain states at lower emergent levels (or nonemergent levels). For example, it seems likely that insects employ a similar mechanism for walking as robots like Attila: rather than having a dedicated central processor for identifying obstacles and moving the various parts of the body to navigate over and around obstacles, insects have multiple specific processors that respond to certain stimuli in certain ways, yielding skilled walking. The behaviours of whole insects, of which walking is a dynamically emergent behaviour, when in colonies, yield other dynamically emergent features of the colony, such as the oscillations in activity patterns in Iæptothoracine ants. Insofar as the activity of a particular ant within a colony can be understood as a result of the dynamically emergent oscillation of the activity of individuals, the fact that one of its legs is not moving, or that it is metabolising at a certain rate, or that the atoms in its body are moving, &c, can also be so understood, but it is probably not appropriate to think of the oscillatory activity patterns across the colony as *explaining* why a particular leg of a certain ant is not moving at a given time. It is, however, responsible for the cycling of activity patterns in whole individual ants, appropriately invoked to explain why a particular whole ant in the colony is or is not active at a certain time.
[25] Clark (1999) op. cit. p. 104
[26] Clark (1999) op. cit. p. 113
[27] Clark's (1999) note 8 to pages 104-115 explains: "A nonlinear system is one in which the two quantities or values do not alter in a smooth mutual lockstep. Instead, the value of one quantity may (e.g.) increase for some time without affecting the other at all, and then suddenly, when some hidden threshold is reached, cause the other to make a sudden leap or change." Clark goes into considerably more detail on the nature of nonlinear relations. The note appears on page 236.
[28] Clark (1999) op. cit. p. 113
[29] See Clark (1999) op. cit. p. 114
[30] The resolution of complex dynamically emergent properties to physical properties may have to proceed through more than one layer of dynamic emergence, but this is unproblematic.
[31] John. McCrone, Going Inside: A Tour round a Moment of Consciousness (London: Faber and Faber Ltd., 1999)
[32] Whether or not a subject is conscious is not relevant to whether or not cells in the visual cortex will fire in response to stimulation.
[33] C. Gross, C. Rocha-Miranda, and D. Bender, "Visual properties of Neurons on the Inferotemporal Cortex of the Macaque," *Journal of Neurophysiology* 35 (1972): Pages.
[34] IT, the inferior temporal cortex, is not properly a part of the visual cortex. It does receive information from (and present information to) the visual cortex, among other areas.
[35] McCrone (1999), op. cit. p. 84
[36] McCrone (1999), op. cit. p. 84
[37] See McCrone (1999) op. cit. 84-86
[38] McCrone (1999), op. cit. pp. 88-89
[39] McCrone (1999) op. cit. p. 89
[40] McCrone (1999), op. cit. p. 92

[44] S. Zeki, "The Visual Image in Mind and Brain", in Editors of Scientific American, The Scientific American Book of the Brain with introduction by A. Damasio (New York: The Lyons Press, 1999)
[45] McCrone (1999) op. cit. p. 77
[46] McCrone (1990) op. cit. p 94
[47] McCrone cites: A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner, "Neuronal population coding of movement direction", *Science* 233 (1986); C. Lee, W. H. Rohrer and D. L. Sparks, "Population coding of saccadic eye movements by neurons in the superior colliculus", *Nature* 332 (1988); and A. Georgopoulos et. al., "Mental rotation of the neuronal population vector", *Science* 265 (1994).

The atom, ant, and ammonite graphics used herein were commissioned especially for this work and executed by David at Draw4u.com. All other graphics are public domain, original, or have been used with the permission and acknowledgement of their creators.