

To be published in *Behavioral and Brain Sciences* (in press)  
Cambridge University Press 2003

---

*Below is the unedited, uncorrected final draft of a BBS target article that has been accepted for publication. This preprint has been prepared for potential commentators who wish to nominate themselves for formal commentary invitation. Please DO NOT write a commentary until you receive a formal invitation. If you are invited to submit a commentary, a copyedited, corrected version of this paper will be posted.*

---

## What to Say to a Sceptical Metaphysician: A Defense Manual for Cognitive and Behavioral Scientists

Professor Don Ross  
School of Economics  
University of Cape Town  
Private bag,  
Rondebosch 7701  
South Africa  
Email: [dross@commerce.uct.ac.za](mailto:dross@commerce.uct.ac.za)  
URL: <http://www.commerce.uct.ac.za/economics/staff/personalpages/dross/>

Dr. David Spurrett  
Philosophy  
University of Natal  
Durban 4041  
South Africa  
Email: [spurrett@nu.ac.za](mailto:spurrett@nu.ac.za)  
URL: <http://www.nu.ac.za/undphil/spurrett/>

**Abstract:** A wave of recent work in metaphysics seeks to undermine the anti-reductionist, functionalist consensus of the past few decades in cognitive science and philosophy of mind. That consensus apparently legitimated a focus on what systems do, without necessarily and always requiring attention to the details of how systems are constituted. The new metaphysical challenge contends that many states and processes referred to by functionalist cognitive scientists are epiphenomenal. It further contends that the problem lies in functionalism itself, and, that to save the causal significance of mind, it is necessary to re-embrace reductionism.

We argue that the prescribed return to reductionism would be disastrous for the cognitive and behavioral sciences, requiring the dismantling of most existing achievements and placing intolerable restrictions on further work. However, this argument fails to answer the metaphysical challenge on its own terms. We meet that challenge by going on to argue that the new metaphysical scepticism about functionalist cognitive science depends on reifying two distinct notions of causality (one primarily scientific, the other metaphysical) then equivocating between them. When the different notions of causality are properly

distinguished, it is clear that functionalism is in no serious philosophical trouble, and that we need not choose between reducing minds or finding them causally impotent. The metaphysical challenge to functionalism relies, in particular, on a naïve and inaccurate conception of the practice of physics, and the relationship between physics and metaphysics.

**Keywords:** Mental Causation, Functionalism, Reductionism, Metaphysics, Explanation.

## 1. Introduction

Philosophy progresses with a tide-like dynamic. Every wave, no matter how strong it seems while rolling in, is followed by a backwash, often nearly as powerful. This makes philosophical development difficult to identify except in long retrospect. For scientists who try to take philosophy seriously in their work, this is bound to be frustrating

Philosophers have been deeply involved in the development of cognitive science. The classical essays by Hilary Putnam (1963, 1967a, 1967b, 1975a) and David Lewis (1972) that articulated and promoted functionalist understandings of mind are among the foundational documents in the literature of the field. Among other things, they showed how and why the study of information processing as conducted in early AI could and should be integrated with psychology more generally. And however far in sophistication the cognitive science community has since moved from the narrowly computational models of the 1960s and 1970s, it is hard to see how it would have gotten where it is now without them. So philosophers do not exaggerate when they claim that their discipline has contributed crucial bricks to the edifice of contemporary cognitive and behavioral science.

By ‘functionalism’ we understand any position which assigns serious ontological status to types of states or processes individuated by reference to what they *do* rather than what they are *made of* – that is by reference to their effects, rather than (necessarily) their constituents. Functionalism of this sort was never without its critics, of course. From our perspective, eliminative materialists (e.g. Churchland 1981) have been the most important of these, and their arguments with mainstream functionalists have been immensely helpful in the effort to see how the neurosciences and robotics best integrate with the more rationalistic projects derived from AI. However, avowed eliminativists have always been a fringe, playing against a relatively monolithic functionalist consensus. For most of the past thirty years cognitive scientists could be assured that the main currents in the philosophy of mind, especially regarding causation and explanation, were running in a direction sympathetic to their activities. This has involved more than encouraging cheerleading, amounting to something of working scientific value. It has helped to guide choices amongst research directions by clarifying just where and how cognitive science might strive for serious integration with nearby research programs in, e.g., neuroscience and the physics of dynamical systems theory (see 3.1 and 4.2 below) without simply collapsing into them.

We regret to report, however, that the backwash has set in. Were a cognitive scientist to stroll into a typical discussion amongst the ‘purer’ philosophers of mind at a professional seminar in 2003, she would find that functionalism is under siege in such settings. Instead, ‘new wave’ reductionism<sup>1</sup> is the horse on which increasing numbers of philosophers are placing their bets.

We are, of course, being melodramatic here, and deliberately so. Good philosophers are rightly cautious about changing their minds or investing in fads, and worthwhile

philosophical activity is not best seen as a war of 'isms'. Nevertheless, as philosophers we are concerned by the rise of a new scholasticism in philosophy of mind that, in the stated pursuit of a return to 'real' metaphysics, threatens a loss of contact with empirical cognitive science. Our aims in this paper are, first, to substantiate this concern (thereby justifying the melodrama), and, second, to offer grounds for resisting the arguments that inspire it. We think that metaphysics – 'real,' professionally done, metaphysics - is an important part of all science, including cognitive science. But we also think that what is recently being promoted under this banner is based on an unhealthy disregard for the actual practice of science, and that too little philosophical discussion of it shows adequate concern for this. To the extent that some philosophers allow their discussions to drift away from relevance to and coherence with scientific activity, the short term course of cognitive science will not be much affected. However, since we would deplore a situation in which the conversation between philosophers and cognitive scientists wound down into separated silos, we think that a corrective with two aims is in order. One aim is to address philosophers themselves concerning the fundamental errors we diagnose in the new scholasticism. The other is to provide cognitive scientists with a manual for answering philosophers who try to convince them that there is something wrong with their metaphysics. After all, to the extent that cognitive scientists respond to philosophers by just shrugging and going off to another room, the conversation winds down; to the extent that they argue back, it continues.

Our discussion is organized as follows: In (1.1) we review the standard arguments for functionalism in the special sciences, and offer an account of the rise of the functionalist consensus. In (1.2) we briefly describe the recent threat to functionalism. In (2) this threat is examined in greater detail, looking first (2.1) at an influential argument against functionalism, and then (2.2) at the form of reductionism which increasing numbers of metaphysicians think is preferable. In (3) we say a little more about the argument in favour of functionalism from the perspective of the special sciences (3.1), and also argue that the reductionism suggested by the metaphysicians would be disastrous for those sciences (3.2). Section (4) contains the metaphysical meat of this paper. In it we distinguish some different ways of taking metaphysics seriously (4.1), consider the relationships between explanation and causation (4.2), distinguish two senses of causation, which we argue Kim conflates (4.3) and clarify a number of considerations relating to the nature of physics, and the relationship between physical science and the metaphysics of causation (4.4). In (5) we review our argument and offer a conclusion.

### **1.1. Functionalism, Philosophy and the Behavioural Sciences.**

Functionalism has a strong claim to being part of the methodological and ontological underpinning of any special science. By 'special' science we have in mind any science not concerned with justifying, testing or extending the generalizations of fundamental physics, and hence most science including (see section 4.4) most of physics. Functionalism offers one way in which special sciences can defend their significance against Rutherford's claim that 'there is physics, and there is stamp collecting' (Birks 1963). What potentially distinguishes a special science from stamp collecting is that it is organised around a distinctive taxonomy of phenomena and a set of processes at some level of abstraction from fundamental physical processes, which are non-redundant, amenable to scientific treatment, and to which a fully realistic attitude can be justified.

The original considerations which led to the development of functionalism were, as it happens, primarily drawn from issues in the sciences of behavior: a response simultaneously

to a simplistic behaviorist equivalence of behavior and psychological state, and to the apparent ‘chauvinism’ of expecting that the specific mechanisms which accounted for psychological states in humans should be regarded as reductive explanations of those states in general.

With respect to simplistic behaviorism, the case for functionalism runs as follows. Traditional behaviorism identified mental states with dispositions to particular behaviors, and hence expected that mental states – insofar as these were of scientific significance at all - could be read directly off surface behavior,<sup>2</sup> so reference to behavior could, and should, replace reference to mental states. One objection to this program pointed out that if mental states can, as seems likely, interact with one another, then there will be neither fixed nor simple pairings of mental states and dispositions to particular behaviors. This line of thinking suggests a place for intermediate causal roles played by (at least initially) unobservable states between ‘stimulus’ and ‘response’. Note that in the first instance these hypothesised intermediate states are characterised extrinsically, by reference to the difference they make to observable states and relations between those states.<sup>3</sup> At this stage, at least, it is possible to be agnostic about what it is that makes the difference in question, even while being confident that some difference is being made. This space for agnosticism about *what* plays the functional role in question relates closely to the second, and for present purposes more philosophically contentious, motivation for functionalism.

In this case the contrast is provided not by behaviorism but by reductionism. In the heyday of type-type reductionism it was expected that particular types of special science states would pay their ontological and causal way by being reduced to types of some science closer to fundamental physics, applied to the same systems. So, perhaps, the biological properties of a system could be reduced to its chemical properties, and from there chemistry could be reduced to physics. Classic statements of versions of this view include Oppenheim and Putnam (1958) and Nagel (1961). So, perhaps, the mental state of being in pain would turn out to be, or already was, reducible to the fact of having activated C-fibers (Smart 1959, Place 1956),<sup>4</sup> in the same way that ‘temperature’ was supposedly reducible to mean kinetic energy of molecules (Nagel 1961: 338-345). On this view the type ‘pain’ was to be considered *reducible* to the type ‘activated C-fibers’ when some bi-conditional bridge law was found enabling statements about pain to be translated into statements about C-fibers and vice versa. The other way of motivating functionalism is, then, to note that the proposed reduction is open to a charge of ‘chauvinism’ (Block 1980b: 270), because even if the biconditional linking pain and C-fiber activity in *humans* held, there presumably could be, or already were, agents physiologically different to humans than nonetheless experienced pain. So, in what came to be the standard jargon, even if what ‘realized’ pain in people was something involving C-fibers, the ‘role’ of pain could be realized in different ways in other types of agent.<sup>5</sup> This, in a nutshell, is the multiple realization argument against type-type reductionism for psychological states, and by implication an argument for a science of psychology that spans differences in realization.

The multiple realization argument can be deployed in various ways as a positive argument for the functionalist project. In the hands of, for example, Fodor (1974, 1975, Block and Fodor 1972) it is used to make clear that many, at least, of the special sciences are concerned with entities and processes which are to some degree abstracted from the details of physical realization. As noted above, practitioners of those sciences can afford to be agnostic about the physical details that realize the relevant kinds and processes, because the distinctive descriptive and explanatory contribution made by their work depends for the most part on

extrinsic, functional, relationships between role properties. A simple and classic illustrative example of the argument here is Fodor's treatment of the notion of a mousetrap (Fodor 1968), conceived in functionalist terms as a device which takes as input a live mouse, and produces as output a dead one. Clearly a wide range of devices and designs are capable of realizing the mousetrap role.

The immediately preceding discussion has referred to the concepts of roles and realizers in stating and partly defending functionalism. This distinction (which is stated and clarified in slightly different terms in Block 1980a) also allows two importantly different ways of being a functionalist. The difference turns on whether one is inclined to identify the functional states with the *role* they play, or with what the *realizer* of that role is in a given case. Put another way, a functionalist might think that 'pain' or 'money' pick out either *the property of having some other (physical) property which realizes pain or money*, or that, properly analysed, they pick out *C-fibers firing or dollar bills*.<sup>6</sup> Saying that the description of the functional state picks out the role indicates commitment to the view that even though their realizers could be very different, humans and, say, Martians could be in the *same* mental state when in pain. On the other hand, tying the function to the realizer entails that humans in pain and Martians in pain are in different states, perhaps different types of pain, just *because* the realizers of the roles in each case are different.

Note that both flavours of functionalism are entirely compatible with materialism, or physicalism. A physicalist functionalist of either type will be committed to the principle that physics is complete, or causally closed, i.e., that there are no non-physical, e.g. vital, fundamental forces (Papineau 1993, Spurrett & Papineau 1999, Spurrett 2001b). Similarly, she will be committed to the thesis that if you fix all of the physical facts, then you've fixed all the empirical facts that there are. Often, although not necessarily, this aspect of functionalist thinking is marked by saying that functionalists accept *supervenience* – the idea that there are no changes without physical changes.

Note also that there is a genuine tension between the different ways of being a functionalist. From the perspective of realizer functionalism, the role variety rides roughshod over distinctions which need to be taken seriously. So the 'equivalence' of a hundred dollar bill, a cheque for the same amount and a bag of coins with a total value of a hundred dollars doesn't amount to much when we have to try and say something about why we can only use one of them in a vending machine, or why only one of them can be bounced by a bank. On the other hand, from the perspective of role functionalism, too much attention to the realizers amounts to abandoning the apparent unity of many apparently powerful and useful generalisations. Qua 'money', it has to be granted, there is a deep sense in which any realization of one hundred dollars just *is* the same.

The importance of this distinction didn't emerge immediately during the early articulation of functionalism. In the classic papers collected in Putnam (1975b), for example, role and realizer versions are run together in a way that is, in critical retrospect, problematic. Philosophers were quick enough to unearth the tension, however. By the 1980s, central debates in the philosophy of mind revolved around arguments between role and realizer functionalists.<sup>7</sup> However, for a number of years, up to the mid-1990s, the debates were preoccupied with the question of whether semantic *meanings*, as bearers of functional roles for beliefs, desires and other 'propositional attitudes', could or couldn't be individuated for the purposes of cognitive science just by reference to intrinsic properties (causal, computational, constitutional or whatever), or were irreducibly relational. This running

controversy was known among philosophers as the ‘internalism vs. externalism’ debate, and for awhile it seemed as if the dispute between realizer and role functionalists turned mainly on it. Fortunately we need not describe its details here, because by the mid-1990s it was largely over, with the internalists – the believers in so-called ‘narrow content’ – having mostly surrendered (see Fodor 1994, Ross 1997). At that point, some thought that the philosophy of mind had made itself ready for thorough integration into cognitive science. In particular, the strong connection between externalism about semantics and the idea that narrowly computational models of thought need to be replaced or supplemented by more biological, environmentally situated and robotic ones (Brooks 1991), made the prospects for positive philosophical contributions to the scientific project look promising. Some of that promise has been realized; we cite, for example, Clark (1997) and Rowlands (1999) with approval in this connection. However, from these same years two ideas gained strength among philosophers that encouraged scepticism about, instead of participation in, mainstream cognitive science. The first of these, the conviction that qualitative consciousness is beyond the reach of functionalist method (Chalmers 1996), or, on some formulations, *any* scientific method (McGinn 1991) is a manifestation of conservative metaphysics that we thoroughly deplore, for reasons given in Ross (forthcoming) and Dennett (2001a, 2001b). This will not be our concern in the present paper, however. The second basis for metaphysical party-pooing, which *is* our present subject, encourages even deeper scepticism because it challenges not just functionalism’s adequacy in a particular domain, but its coherence in general.

For reasons we explain in Section 2 below, realizer functionalism didn’t die with semantic internalism. As far as we know, the first recognizably contemporary expression of the worry that states picked out by reference to functional roles alone can’t *cause* anything is Fodor (1987). However, at that point the worry was deeply enmeshed in the internalist / externalist controversy, so its subversive potential wasn’t clearly spotted. However, with the passing of internalism it popped clearly into wide view. In Kim (1998) it finds book-length and elegant expression,<sup>8</sup> and our experience as casual anthropological observers of fellow philosophers indicates to us that the majority of philosophers of mind are, although not unanimously persuaded by this version, inclined to take it very seriously and, if not agreeing with its conclusion, to accept the same basic picture of how things are in science, especially physics, when engaging with it (e.g. Marcus 2001, Elder 2001). In what follows we occasionally find (and cite) allies among ‘pure’ philosophers, and it is no part of our project to argue that *nobody* should pursue these problems by primarily logical methods. Given, though, that we are here confronted with a piece of metaphysics claiming consequences for science, we take it as deserving evaluation with an eye to the science *and* the metaphysics. As Marras (2000) points out, and as we will explain, what was originally supposed to be a consideration against role functionalism but *for* realizer functionalism now looms as a sceptical threat to *all* functional explanation in any science.

Our aim in this paper is to comprehensively respond to the basis for this scepticism, from the perspective of behavioural and cognitive science. Doing this, however requires some excursions deep into metaphysics. Some scientists will likely doubt that such excursions could be worthwhile trips for them to go along on. Hearing that philosophers are making themselves uneasy about the enterprise of cognitive science because of metaphysical itches, they may be inclined to respond pragmatically, saying “*we* feel fine, so *you* stop scratching!” Such responses, often heard when philosophers confront scientists with their metaphysical scruples, do not *just* express a macho attitude. It has been a widespread opinion among philosophers of science for decades that philosophy has no privileged epistemic perspective from which it legitimately can or should try to bend science to any prior ontological objective

or methodology. We endorse this stance, in a fashion to be made more precise shortly. However, we *also* agree with Kim (1998), that if metaphysics matters then it had best be done seriously. We believe furthermore that metaphysics only matters if it matters to science; and, finally, we believe, and argue below, that metaphysics matters to science. Given all this, it of course follows that if the metaphysical presuppositions of cognitive science are causing genuine itches, then everyone ought to care about scratching in the right place.

We find it necessary to say something about these grand themes for the following reason. Kim's flagship argument against the recent externalist-functionalist near-consensus *does* have a scholastic aura about it; in particular, as philosophers take up Kim's challenge and gnaw away at the problems he has raised, they focus a great deal of their attention on subtle differences amongst variations on the definition of the 'supervenience' relation. We'll ultimately conclude that this really is a scholastic's response, in the bad sense of the word (if there is a good sense). But this generates two strategic concerns at the outset. First, this conclusion may lead some philosophers to suppose that we are trying to have what Kim calls a 'free lunch', that is, simply refusing to take the demands of metaphysics seriously. Second, the fact that what we regard as a *scientifically interesting* metaphysical problem comes dressed in scholastic garb – it's even based on something *called* 'the supervenience argument' – will lead too many scientists to conclude right away that we're engaged in an in-house philosophers' quibble that isn't any proper business of theirs. These concerns present us with the following tactical burden. We must present the supervenience argument, the basic grounds for the new disquiet, in a way that does logical justice to it *and* captures the gripping intuitions behind it that we don't think you have to be scholastically inclined to appreciate.

So, here goes. Our talk about 'scientifically interesting metaphysics' gestures at the following fact. It *is* a feature of scientific epistemology, as really practised in laboratories and journals, that the various pieces of scientific inquiry must broadly cohere into a general world-view that, at least in its core, almost all signed-up members of the mainstream scientific professions can share. Furthermore, it is a legitimate job of the 'serious' metaphysician to ensure that proposals for articulating and enriching this world-view are, at least potentially, genuinely enlightening, and not merely verbal or technical. By 'genuinely enlightening' we mean that such articulations should actually be able to help scientists choose amongst theoretical and/or procedural alternatives in cases where the empirical facts remain sufficiently underdetermined to leave options open in pragmatically pressing (as opposed to just logically possible) ways. Now, what we have just said is not very precise, and so not very bold. But it is enough to help show why metaphysics *can be* (and the issues raised by Kim's supervenience argument *are*) scientifically interesting. Our bland claim makes a minimal commitment to the idea that, at some level of abstraction, the sciences need to 'hang together.' However, this commitment is in direct tension with the best motives for having special sciences, all of which turn on the facts that, along various dimensions both ontological and epistemological, different sciences do not hang together, and that we'll deny ourselves important insights and generalizations if our respect for minimal metaphysics makes us work too hard to try to get them to do so. Kim's supervenience argument is aimed precisely at this tension, and with unusually limpid clarity. Though, as we shall see, the argument generalizes all the way across the sciences, it helps its clarity, but at the same time especially challenges cognitive and behavioural scientists, that it is focused directly on this tension as it arises within the domain of their work, which sits across the fault line between generalizing and special ontologies. So we think that a working cognitive scientist who is confronted with Kim's argument will and should then notice the tension every time she goes to write up some new results, and will and should feel itchy. The problem then, we will argue, is that Kim and other philosophers,

instead of telling her where and how to scratch, counsel relief through professional suicide. We will be pleased to show that this is not called for.

Kim's argument is aimed directly at role functionalism. According to its conclusion, role functionalism is not a stable metaphysical position. Instead, it collapses into a choice between epiphenomenalism and reductionism about mental properties, objects and processes. Kim assumes that epiphenomenalism would be a dire outcome, both metaphysically and scientifically, but then spends much of his book trying to make reductionism seem palatable. As we will show, he is not convincing. The foundational assumptions of cognitive science, along with those of other special sciences, deeply depend on role functionalism. Such functionalism is crucially supposed to deliver a kind of causal understanding. Indeed, the very point of functionalism (on role *or* realizer versions) is to capture what is salient about what systems actually do, and how they interact, *without* having to get bogged down in micro-scale physical details. Functionalist understanding is, furthermore, supposed to deliver all the goods of properly causal scientific work: permitting predictions, causal explanations, sustaining counterfactuals, enabling the planning of interventions and so forth. But if reference to role properties can be shown to be causally redundant, as Kim's argument purports to show, then the appearance of causal relevance is a sham, and role functionalists, including most cognitive and behavioral scientists, most of the time, are really only telling 'just so' stories to one another.

So, apologies for some coming scholasticism duly made, let's now get this dangerous supervenience argument onto the table.

## **2. The Armchair Strikes Back**

According to Kim, the key challenge to role functionalism turns on what he calls the 'causal exclusion' problem, which arises if he is correct that putative physical and mental causes for the 'same' event can be shown to be in conflict. His problem, therefore, is to provide an answer to the question: '*Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?*' (1998: 38). This is the problem of 'finding a place' (1998: 2) for mind in a physical world, given the causal closure of physics. The fact that Kim is concerned with the problem as a *metaphysical* challenge means that it won't do simply to point out the pragmatic benefits, or indispensability, of mentalistic explanations (including causal ones) without having a good metaphysical story to tell about *how* and *why* such explanations are legitimate (cf. Marcus 2001). This would be the strategy, discussed and rejected above, of ignoring the demands of metaphysics, asking, as Kim says, for a free lunch – keeping your comfortable intuitions by refusing to notice that they commit you to anything outside of cognitive science.

### **2.1 Kim's 'supervenience argument'.**

Whether or not a cognitive scientist is in the habit of using the word 'supervenience', chances are good that some of her daily working assumptions involve at least a loose version of the concept. Starting generally: one set of (e.g., mental) properties supervenes on another (e.g. physical or neurobiological) set if, roughly, something cannot change with respect to its supervening properties without undergoing some change with respect to its subvening ('base') properties. Materialist functionalism involves commitment to supervenience in this sense, insofar as it is reasonable to suppose that what role some entity realizes cannot change without some physical changes taking place *somewhere*. This relationship of covariance plus

some kind of dependence (because physical changes need not lead to changes at the supervening level) is weaker than reduction, and does not commit you to anything like realizer functionalism (let alone internalism) unless you add that the relevant physical change has to occur *in the realizer*.

Kim's argument takes the form of a dilemma that 'apparently leads to the conclusion that mental causation is unintelligible' (1998: 39). The dilemma has two horns: on one horn mind-body supervenience is allowed to fail, and in the other it is assumed to hold. For the purposes of formulating the dilemma Kim defines the mind-body supervenience thesis as follows:

Mental properties supervene on physical properties in the sense that if something instantiates any mental property *M* at *t*, there is a physical base property *P* such that the thing has *P* at *t*, and [nomologically] necessarily anything with *P* at a time has *M* at that time (1998: 39).

This definition is not perfectly general. Philosophers have generated a large literature that debates the merits and failings of alternative definitions of supervenience. What is at issue in these arguments is the appropriate *scope* to aim for in stating generalizations about functional role-fillers. At *least*, 'pain' should apply generally and univocally to (most) people, and probably to creatures with which people share recent common (or, perhaps, any) ancestors. Perhaps almost any life form would need a trip-wire system that alarmed it by making it feel bad. If so then Martians would have pain too, however different its realizers might be in them. Now, to help discipline arguments about this sort of thing, it's a useful strategy to first fix the *essential* conditions on pain; that way you hold your semantics fixed and can test the empirical facts independently. Philosophers fix essential semantics by considering various abstract possibility classes, or 'possible worlds'. Depending on how many of these classes you want to legislate supervenience relations as having to hold across, you get different logical definitions of the relation. Fortunately for our purposes here, Kim's dilemma arises for *any* such definition, so we will treat the version just quoted as exemplary.

Here's the first horn of the alleged dilemma.

If mind-body supervenience, in general, were to fail, and we are committed to the causal closure of physics, then it seems as though we could not make sense of mental causation. Put another way, if the supervenience relation *doesn't* hold, and mental causes do have physical effects then we'd have to deny the causal closure of physics – we'd be claiming a physical consequence of a non-physical cause. As materialists, or physicalists, we can't do *this*, so it looks like the supervenience relationship has to hold. (Kim takes commitment to the causal closure of physics as being a 'minimal' requirement for physicalism). So far so good – this is a standard motivation for endorsing supervenience if you aren't willing to be a reductionist (e.g. Fodor 1987).

On, then, to the other horn.

'Suppose that some instance of mental property *M* causes another mental property *M\** to be instantiated' (Kim 1998: 41). By the mind-body supervenience thesis, *M* has a physical supervenience base *P*, and *M\** has a physical supervenience base *P\**. Kim asks us to grant that *P* causes *P\**. But, then, since *M\** is realised by *P\**, why have an apparently separate causal claim to the effect that *M* caused *M\**, especially when it seems as though once *P*, then

$M^*$  was going to happen anyway? (following Kim 1998: 38-56, see also Marras 2000). More precisely we seem to have to choose between:

$M^*$  is instantiated on this occasion: (a) because, *ex hypothesi*,  $M$  caused  $M^*$  to be instantiated; or (b) because  $P^*$ , the physical supervenience base of  $M^*$ , is instantiated on this occasion (1998: 42).

Kim notes that the apparent tension above could be relieved by accepting that ' $M$  caused  $M^*$  by causing  $P^*$ '. But, given that both  $M$  and  $M^*$  have respective physical supervenience bases, we should ultimately grant that ' $P$  caused  $P^*$ , and  $M$  supervenes on  $P$  and  $M^*$  supervenes on  $P^*$ ' so that the ' $M$ -to- $M^*$  and  $M$ -to- $P^*$  causal relations are only apparent, arising out of a genuine causal process from  $P$  to  $P^*$ ' (1998: 45).

So: If you *deny* supervenience you seem to be abandoning materialism, which would be terrible,<sup>9</sup> and if you *affirm* it you get stuck with a choice between epiphenomenalism about the mental, or reductionism. The former is an awful option for cognitive science. Therefore, the only option is reductionism. This is genuinely amazing, since the very point of endorsing supervenience was originally to allow materialism *without* reductionism!

## 2.2 Kim's reductionist proposal

Kim's reductionism is not quite the standard ('Nagelian type-type') variety that people still learn in undergraduate metaphysics and philosophy of science courses (See Marras 2002). According to that model, you reduce some type  $x$  to some type  $y$  by justifying a 'bridge-law' to the effect that all of the causal and other law-like generalizations you can state in terms of  $x$  can be re-stated in terms of  $y$ . Instead, Kim proposes a reductionism that proceeds along the lines suggested by Armstrong (1981) and Lewis (1980). The details of the proposal involve a crucial step called 'functionalization' which involves "enhancing bridge laws ... into identities" (Kim 1998: 97).<sup>10</sup> Identities, unlike bridge laws, give ontological simplification, and promise to explain why it is that the bridge laws hold true. Functionalization is to be achieved by 'priming' the to-be-reduced mental property (the proverbial  $M$ ) for reduction, which means reconstruing it in extrinsic or relational terms, i.e. specifying its causal relations to other properties. So  $M$  is now 'the property of having a property with such-and-such causal potentials, and it turns out that physical property  $P$  is exactly the property that fits the causal specification' (1998: 98). It follows that  $M$  can be identified with  $P$ , which would solve the causal exclusion problem, because one property cannot be in competition with itself over causal relevance, and Kim thinks there is no problem about the causal capacity of physical properties.

It is, of course, an open question to what extent such reductions are possible, and how extensive the scope of any given functionalizing reduction will be. The multiple realization argument (discussed above and, again, below) indicates that functionally individuated properties can have very diverse realizers, so functionalising reductions should be expected to involve some disintegration of the role properties. Kim himself seems comfortable with this, describing the upshot of his arguments as being that 'multiply realized properties are sundered into their diverse realizers in different species and structures, and in different possible worlds' (1998: 111). This is supposed to save *something* of functionalism, albeit at the expense of relinquishing the capacity to say what it is that makes some apparently similar functional properties related or the 'same' in cases where their realizations are significantly different. (We return to the question of just how much difference would count as significant

in due course.) Kim's approach, interestingly, inverts the standard image of functionalism, traditionally regarded as a major form of antireductionism, since on his view 'the functionalist conception of mental properties is *required* for mind-body reduction' and is even 'necessary and sufficient for reducibility' (1998: 101). But is this functionalism at all? Marras (2000) thinks not, and argues that Kim has 'in fact given up on functionalism' of which a central idea was that mental/functional properties retained their 'identity and projectibility across heterogeneous physical realizers'. Kim, who claims to take multiple realizability 'seriously', concedes that to those who might want to 'hang on to' functional properties as 'unified and robust ... in their own right' his proposal will be a 'disappointment' but also maintains that the conclusion in question is 'inescapable' (1998: 111).

Notice at once that if there *is* any sort of functionalism still alive in Kim's proposal, it's realizer functionalism, not role functionalism. So perhaps what Kim's argument, and his way out of it, shows is that if you want to try to be a *serious*, anti-reductionist, functionalist then you had, somehow, better be a role functionalist. As discussed in Section (1) above, many have thought that since at least 1987; but initially the implausibility of semantic internalism was the main reason. Now it turns out that there's a more general reason: if you try to be a realizer functionalist, you'll turn 'inescapably' into a reductionist, and you won't be able to do cognitive science (or biology, or economics, or ...)! Or so we now aim to show. Remember, though, that showing we'd be *in trouble* if we followed Kim, no matter how *big* the trouble, doesn't show that we're *not* in trouble. Acknowledging that is the price of taking metaphysics seriously.

### 3. Special Sciences without Functionalism

In section (1) of this paper we outlined the reasons for the establishment of a broad functionalist consensus in the behavioral sciences, and the special sciences more widely. Functionalism seemed, was *devised* to be, ideal for such sciences, insofar as it offered a justification for focussing on role properties and extrinsic relations, coupled with a well-motivated degree of agnosticism about the exact physical details of the systems studied. In section (2), though, we described Kim's supervenience argument, contending that functional causal claims, understood as being claims about properties which supervene on more basic physical properties, are epiphenomenal, and can only have their causal status saved by reducing them to physical properties.

It is not essential that anyone view this as a problem. One simple way to avoid the challenge Kim poses is to be an instrumentalist about functional claims. That means contending that metaphysical questions about the causal status of scientific claims just aren't important, and that what really matters is whether science is, in some sense or other, 'useful'. It isn't, after all, *compulsory* to worry about metaphysics. If you are indeed willing to say that, ultimately, the validity of some piece of science is determined on pragmatic grounds, then this is your stop, and you can disembark right now. In so proceeding you are allowing that you don't mind if the behavioral sciences are considered to be a kind of stamp collecting – a process of arranging the artifacts of our own epistemic limitations in interesting or useful-seeming ways. (As we argue shortly, in so doing, whether you like it or not, and more to the point whether *he* likes it or not, you're agreeing with Kim, because the only place he leaves open for the special sciences is an instrumentally justified one.)

If you're still here then perhaps you want to be more than a stamp-collector. Perhaps you want a defensible functionalist conception of 'pain' that generalizes across species, or of

‘competition’ that generalizes across organisms and ecologies, or even of ‘mousetrap’ which does justice to the varied assortment of gadgets you have around for the purposes of killing mice. In this section we aim to do two things: first (in section 3.1), extend section (1) above by developing stronger and more sophisticated arguments against reductionism in the special sciences; and second (in section 3.2), make clear that Kim’s proposal *does* amount to turning special scientists into stamp collectors.

### 3.1 Explanation and Causation

It is a manifest fact about science that the various special sciences are partly constituted by parochial types of causal relations. Indeed, this is one of the principal things making them *special*. These relations are, furthermore, reciprocal functions of the accepted explanatory schemata in the relevant sciences. This fact is, at least in the first place, sociological rather than metaphysical. One way of being what Kim derides as a free lunch seeker is to take this as a brute fact in need of no explanation, supposing instead that the ‘specialness’ of each special science, taken individually, is somehow self-justifying. *Part* of what is involved in heeding Kim’s enjoinder to take metaphysics seriously is acknowledging the need to say something about the circumstances under which special science accounts are genuinely explanatory, where it is presumed that a genuine explanation is not merely something psychologically satisfying to someone, but must cite explanans that are both true and informationally non-redundant. In this section, we will show that, in light of leading accounts of explanation from the philosophy of science literature, Kim’s version of reductionism would disqualify many or most *prima facie* powerful special science explanations.

Where special sciences are concerned, we can inquire about the explanatory value of a specific account at either or both of two levels. An account might be genuinely explanatory just relative to the particular ontological and causal structure of the science in which it is embedded, but remain mysterious from the perspective of the wider standpoint at which science as a whole is expected to ‘hang together’. Kim, of course, contends that explanations citing mental causes have just this status unless we embrace his reductionistic version of realizer functionalism. The inquirer who takes metaphysics seriously seeks accounts of phenomena that are explanatory *both* relative to the ontological presuppositions of her special science, and to whatever wider metaphysical principles unite the sciences as a whole. The project of seeking explanatory generality of this sort is historically, actually and normatively, *part* of the business of science. That is to say that the naturalistically oriented metaphysics that we engage in is continuous with, rather than separate from, what ‘scientists’ do. Our main criticism of Kim’s proposal, to which we will devote Section 4, is that the particular wider metaphysical perspective he takes for granted has no persuasive justification. At the moment, however, we are concerned with tracing the consequences of Kim’s proposal for the special sciences, and for the cognitive and behavioral sciences in particular. But since we contend that one such consequence would be the disqualification of a whole class of important (putatively, at least) explanations, we cannot avoid some introduction at this point of general considerations from the philosophy of science. For the moment, these considerations are intended to facilitate our discussion of special-science explanations. In Section (4), we amplify them in a general treatment of the demands of serious metaphysics.

Kitcher (1976, 1981) has argued that *ontological unification*, either within a special science or across two or more special sciences, consists in the justification of common *argument patterns* that hold within or across, respectively, the science(s) in question. This claim is then substantiated through detailed analysis of the concept of an argument pattern, which is a set of

ontological and structural primitives featuring recurrently in the explanations given in the unified domain. Thus, for example, evolutionary biology is unified by its recurrent use of explanations that cite measurable effects of environmental or other selection on the distribution of varying and heritable properties within populations. A biologist does not, *qua* biologist, query the cogency of this sort of explanation in general, since accepting the soundness of its generic logic and its general ontological appropriateness is part of what makes her a biologist. We need not here endorse all the details of Kitcher's analysis in agreeing that this idea identifies one plausible element of the vector of (soft) constraints on explanatory unification. Over the course of his recently truncated career, the late Wesley Salmon explored another element of the vector, one lying more clearly and directly in the metaphysical tradition that seeks a basis for ontological monism in one fundamental kind of 'stuff'. That is, Salmon endeavored to show something enlightening about the ontologies of all sciences by reference to general micro-structural relations that bind all real objects and processes. In the philosophy of science literature, Kitcher's and Salmon's approaches are taken as offering rival bases for identifying good scientific explanations in the shared context of scientific realism.<sup>11</sup> We agree with Salmon's (1990) view that while neither his approach nor Kitcher's may furnish a complete and ultimate analysis of explanation, they form a complementary pair of answers to a general question about what science wants and needs from philosophy of science.

In the context of our response to Kim here, we will be following a road that Kitcher and Salmon have mapped quite explicitly in dialogue with one another. Kitcher (1989) characterizes his work as analyzing 'top-down' explanation, wherein we explain phenomena by fixing their roles in wider ensembles of regularities, and he contrasts this with 'bottom-up' explanation, the sort analyzed by Salmon, which consists in identifying the causal-mechanical processes that generate a phenomenon being explained. Salmon (1990) endorses this idea of a 'duality' of explanatory approaches, which he takes to apply across the board. Thus, to cite one of Salmon's examples, we provide a top-down explanation of industrial melanism in peppered moths by means of the familiar story embedded in population genetics and evolutionary ecology, and we would furnish a bottom-up account to supplement it if we added facts about the synthesis of proteins that lead up to the production of differently colored wings.

As will be clear from our discussion in the previous sections, Kim can be happy enough with this sort of duality in *explanation*. His difficulties turn on the fact that, according to his analysis, top-down accounts of the Kitcherian sort cannot be causal. Neither Kitcher nor Salmon would necessarily disagree with this, since the duality they endorse is epistemological rather than metaphysical. However, many typical explanations in the behavioral and cognitive sciences seem to be simultaneously top-down *and* causal.

Consider the following example, based on Hutchins (1995) which echoes many others found in the current cognitive-science literature on intentional action. Some navigation systems on large ships require two specialist 'pelorus operators', one on either side of the ship, each reporting, with the aid of a special instrument (the pelorus), the angular position, or bearing, of visible landmarks. Pelorus operators do not select the landmarks, instead they are specified by other members of the navigation team. Imagine a pelorus operator, recently ordered to 'stand by to mark' the bearing of a particular landmark, and so expecting immanently to be asked to report the continually changing bearing upon being ordered to 'mark'. The actual response to the 'mark' instruction will be constituted by a series of neural, nervous and muscular events that the pelorus operator can't directly access for description to himself, or

subsequently report as distinct from one another (even if he knows on theoretical grounds that they must have been).

His actions – including adjusting the orientation of the pelorus, maintaining a state of readiness to report the current bearing by frequently consulting the apparatus and what is visible through it, and rehearsing and reporting the reading, will largely consist of pre-prepared subroutines that can be executed as relatively autonomous wholes. These subroutines will be the product of training, guided by personal habits and primed by ritualized social cues. Some subroutines will be specialized at gathering information from the world (reading the instrument, decoding instructions about landmarks), some at controlling the information gatherers (lining the apparatus up on a landmark given an external instruction), some at producing responses according to strict conventions (reporting the bearing when instructed to ‘mark’, inter alia by producing the required phonemes in the required order). Others still will co-regulate the activity of those already mentioned – preventing the reporting system from being executed until the ‘mark’ instruction has been decoded, etc. The routines will therefore partly be coded as dispositions in particular synaptic firing pathways, amenable to being triggered by some small subset of those synapses.

Further, the pelorus operator’s *entire brain* must, on balance, be so configured that the output of the instrument reading subroutines, when released by the decoded ‘mark’ command, controls reporting behavior, preventing him from becoming enraged when remembering the ‘Mark’ is also the name of a romantic rival, or abandoning his station to tie a shoelace, and so on. He must instead be neurally primed to check and report the bearing at the moment of hearing the mark command, and do nothing else. So, there’s the setup. And now, action! The command ‘mark’ is uttered and decoded, the visual position relative to the calibrations on the instrument consulted, the markings transduced and processed, the result slotted into the conventional template, the phonemes rehearsed and uttered ‘[Landmark X] 237’.

This explanation is relentlessly causal, but it is very far from strictly bottom-up. The ‘subroutines’ to which we casually referred are black boxes, top-down characterizations of networks of connections that include both triggers and inhibitory links. At every stage, we picked out these black boxes as pure role-fillers, by reference to a rich conceptual network that we already know the operator must have learned. Perhaps, though, we were just being lazy, or deferring to our own ignorance: if we could have provided the whole bottom-up story, individual electrochemical event by individual event, wouldn’t we then have provided the *exclusive* causal story? Let us postpone that question for now. Notice that even if a full specification of ordered synaptic potentials *is* the exclusive causal story, then, as functionalists of all sorts have long emphasized, reciting the specification would be a poor explanation of what happened, because there’s nothing systematically special about *these* particular synaptic sequences that ties them to bearing reports from one occasion to the next. Furthermore, at one point we had to cite the dispositional state of the pelorus operator’s *entire brain!* But this state will likely never occur again, exactly, no matter how long the operator’s career or how many bearings he reports. And knowing the state in one case would do very little to illuminate different cases: what would those neurons, let alone the operator, have been doing were an alternative landmark to have been specified? Or were the ‘mark’ command to have been given a moment later? Thus the strictly synaptic account would miss almost all of the counterfactuals relevant to behavioral explanation. The fundamental basis for this is the servosystematic nature of the control architecture at work here. If some synaptic paths wander away from the central task, then feedback generated from other regions concerned with

attentional focus will quickly recruit backup or alternative resources. Restricting explanation to the *actual* microcausal chain misses this structural fact.

The account of the pelorus operator's action given above is an instance of what Jackson and Pettit (1988, 1990) call *program explanation*. Social pressures operating on the pelorus operator ensure that one of many possible overall configurations of his brain that keep him focused on his task will (likely) be in place as the moment for action looms. This in turn 'programs for' one suitable chain of synaptic events or another, by virtue of the feedback mechanisms through which brains embedded in environments control behavior in general. Here is what Kim says about program explanation. First, he invokes one half of Salmon's duality in asserting that "to explain an event is to provide some information about its causal history." Then "what can be done is to define, say, the 'causal network' of an event which is closed under both causal dependence and its converse, and then explain the idea of explanation in terms of providing information about the causal network in which an event is embedded. Pointing to an epiphenomenon of a true cause of an event [thus] does give some causal information about the event" (Kim 1998, 76). In offering this analysis, Kim does not disagree with Jackson and Pettit themselves. According to them, the 'programming for' relation provides 'causally relevant information' but is not itself a causal relation. That is, to them, knowing about the pelorus operator's role plus the changing position of the landmarks tells us that *some* causal process sufficient for a bearing of '237' being reported will unfold, but not *which* one.

Suppose a scientist explains an animal's hunting by saying that it's hungry – in advance of knowing enough to have 'sundered' hunger by reduction into hunger<sub>lion</sub>, hunger<sub>mantis</sub>, hunger<sub>snake</sub> and so on. She would then be giving us a program explanation of the hunting. Based on his remarks just quoted, Kim would concede that this explanation 'gives some causal information'. Furthermore, he seems to have no grounds for denying that it gives the *right* causal information so far as prediction and generalization are concerned; for as the example of the pelorus operator is supposed to show, the program explanation supports the relevant counterfactuals. So why are we supposed to still be worried about the causal exclusion problem? Here's why:

I believe it is only this sort of extremely relaxed, loose notion of explanation that can accommodate Jackson and Pettit's program explanations. Explanation is a pretty loose and elastic notion – essentially as loose and elastic as the underlying notions of understanding and making something intelligible – and no one should legislate what counts and doesn't count as explanation, excepting only this, namely that when we speak of 'causal explanation' we should insist ... that what is invoked as a cause really be a cause of whatever it is that is being explained (Kim 1998: 76).

Implicit in this response is a metaphysical restriction on what sorts of states can and cannot figure in 'real' causal explanations. Kim interprets minimal physicalism (that is, commitment to the causal closure of the physical) as requiring that all properties that cause things must *be* (perhaps by reductive identification) physical properties. This, of course, invites us to ask what makes a property 'physical.' Kim does not provide an analysis, but merely a recursive restriction that ties the physical to the 'micro.' That is: "First, any entity aggregated out of physical entities is physical; second, any property that is formed as micro-based properties in terms of entities and properties in the physical domain is physical; third, any property defined

as a second-order property over physical properties is physical” (Kim 1998: 114-115). Then the idea is that as long as the domain of ‘real’ causal explanations is restricted to explanations that cite only micro-based properties, we are guaranteed never to violate the principle that physics is causally closed. Now we want to know what ‘micro-based’ means. Here is Kim’s definition of a micro-based property:

*P* is a *micro-based property* just in case *P* is the property of being completely decomposable into nonoverlapping proper parts  $a_1, a_2, \dots, a_n$ , such that  $P_1(a_1), P_2(a_2), \dots, P_n(a_n)$ , and  $R(a_1, \dots, a_n)$ . (Kim 1998: 84)

Micro-based properties are thus macroproperties that are not shared by the micro-constituents of the macro-systems that bear or instantiate them. So  $\text{hunger}_{\text{ion}}$  could be a macroproperty, though hunger in general presumably couldn’t (see Section 3.3).

Thus on Kim’s view, whatever macroproperties ‘really cause’ molar behavior must be decomposable into individual, nervous system–based, properties. This does *not* amount to the absurd thesis that all *causal powers* at the macro-level *are* actually micro-properties; Kim knows that cars can get people down the street while parts of cars can’t. Rather, what he’s committed to is the thesis that a system’s causally effective macroproperties derive their effectiveness entirely from interactions among causally effective microproperties that are both regular and intrinsic to the same system.<sup>12</sup> He answers worries about radical multiple realizability of mental properties, with which both parts of this commitment are inconsistent, by suggesting that the possibility of practically interesting psychology shows that, as a matter of fact, multiple realization is not out of hand:

The idea that psychology is physically realized is the idea that it is the physical properties of the realizers of psychological states that generate psychological regularities and underlie psychological explanations. Given an extreme diversity, and heterogeneity of realization, it would no longer be interesting or worthwhile to look for neural realizers of mental states for every human being at every moment of his / her existence. If psychology as a science were possible under these circumstances, that would be due to a massive and miraculous set of coincidences (Kim 1998: 94-95).

Many cognitive scientists who see program explanations as playing ubiquitous and irreducible roles in their domain (along with those of other special sciences) do *not* agree that it must be “the physical properties of the realizers of psychological states that generate psychological regularities and underlie psychological explanations.” Most will likely concede that similar neurophysiology (and other physiology) from one individual to the next makes it possible for people to share comparable natural capacities, saliences and learning histories, which is a necessary etiological condition for cultural learning. However, the operations of the natural devices that do this learning are not *equivalent* to our molar selves. Mental states are individuated by a process of triangulating under equilibrating pressure from similarity of cognitive and perceptual apparatus, similarity of social pressures on our histories of self-construction, and shared ecologies (especially social ecologies).<sup>13</sup> The basis for an interpretation of some set of synaptic potentials in the pelorus operator’s brain as being ‘the state of believing that the bearing to the landmark was 237 degrees at the time he was ordered to mark’ is, *in part*, reference to his history as someone conditioned to perform social roles, and, in particular, a role in a practice that has such-and-such conventions.

Explanations of this triangulating kind are pervasive enough across the behavioral sciences that their genus constitutes a recognizable Kitcherian argument pattern. We identify hunger-states by triangulating amongst physiological, ethological and evolutionary-ecological factors; and then we furnish explanations of particular events in animal lives by supposing that hunger programs for displays of search and consumption. We identify productive activities in economics by triangulating amongst considerations of energetic output, behaviorally derived utility functions and culturally evolved rules of exchange; and then we try to explain particular decisions of firms by supposing that production-possibility frontiers and profit-maximization functions (given some cost of capital) program for the appearance at particular prices of goods on the market. There has been no shortage of attempts to rigorously ground this loose argument pattern of triangulation in a generic but rigorous common logic – dynamic game theory, in which any of a variety of selection mechanisms sifting amongst rival strategies for allocating scarce resources lead to predictable shifts in the distribution of behavioral tendencies, is the current favorite candidate (Gintis 2000, Ross 2001). In these respects, the behavioral and cognitive sciences look no obviously worse off, no intrinsically less unified as a suite, than the various wings of physics and chemistry taken as a group. But Kim's contention that special sciences are only genuinely explanatory if they can survive a reductionistic re-interpretation does *not* depend on his finding that their typical explanatory attempts fail Kitcher's criteria. Clearly, for Kim, the unifying strategy championed by Salmon trumps Kitcher's: the epistemological duality is not mirrored at the ontological level. Scientists cannot reasonably be expected to share this intuition, however, and throw away what look like powerful explanations from one leading philosophical perspective, unless serious, professional-class metaphysical arguments show that Salmon was more obviously holding trumping aces than even Salmon himself thought.

We thus best press at the strength of the basis for Kim's 'hyper-Salmonian' intuitions about explanation by asking first how they are supposed to make sense of actual explanations in the behavioral and cognitive sciences, then, if that strains the prospects for accommodation, inquiring into the persuasiveness of their roots in general metaphysical analysis by itself.

The second of these tasks is taken up in Section (4). In pursuit of the first question, let us first note that the triangulational approach to the individuation of mental states in psychology is compatible with two possible situations where the macro-micro relation is concerned. On the one hand, mentalistic psychology and neurophysiology might employ typologies that cross-classify across their putative micro-bases. In Kim's words:

To say that a given taxonomic system cross-classifies another must mean something like this: there are items that are classified in the same way, and cannot be distinguished, by the second taxonomy (that is, indiscernible in respect of properties recognized in this taxonomy) but that are classified differently according to the first taxonomy (that is, discernible in respect of properties recognized in that taxonomy), and perhaps vice versa. That is, a taxonomy cross-classifies another just in case the former makes distinctions that cannot be made by the latter (and perhaps also conversely). (Kim 1998: 68-69).

According to Kim, this amounts to a denial of supervenience as a one-way relation, permitting what Meyering (2000) calls 'multiple supervenience' (see below). Kim says that "this is a serious form of dualism, perhaps an approach worthy of serious consideration".

Kim's two uses of 'serious' here must prevent us from regarding this as name-calling. On the other hand, we really don't think that 'dualism' is quite the apt word here, since in this context it is clearly supposed to indicate views which deny the causal closure of physics. We will indicate reasons for doubting that acknowledgement of multiple supervenience implies such dualism, after first indicating just what multiple supervenience amounts to and why special sciences constantly traffic in it.

### 3.2 Multiple Supervenience and Special-Science Explanation

Meyering (2000, 191) introduces the concept of multiple supervenience by means of an analogy with dispositional explanation, and referring to the imagined example of Mary, electrocuted while atop an aluminum ladder:<sup>14</sup>

... dispositions, just like macro-properties, fail to produce causal effects independently of their categorical base. And yet their explanatory power clearly differs from, or exceeds, that of their bases. This becomes intelligible when we recognize that one and the same categorical base 'realizes' more than one disposition. Even so, only one of those is usually relevant for a given event. Thus Mary's death is related to the electrical conductivity of her aluminum ladder. But the categorical base thereof (the cloud of free electrons permeating the metal) also 'realizes' such diverse dispositions as the thermal conductivity or the opacity of the metal.

The key point here is that the categorical base on its own, given that it realizes more than one disposition, plays a less effective role in an explanation than does one particular disposition it realizes. Referring to the realizer is insufficiently precise compared to citation of the disposition, or role. So the "actual realizer state is not merely inessential because a different state might have realized the same causal role. Rather it is inessential because the *very same realizer state* may yield a wide range of very different causal trajectories" (2000: 193). (This nicely exemplifies why even Salmon came to recognize the need for duality in explanation: Kitcherian top-down explanations are often more informative than bottom-up ones, and *objective* informativeness is surely a *metaphysically serious* aspect of explanation on any reasonable account.) One way of describing the state of affairs Meyering considers is to say that there are supervenience relations (i.e. relations of covariance plus dependence) going in two directions at once here. On the one hand the disposition supervenes on a particular set of micro-properties, but the disposition could be realized by different micro-arrangements. On the other, the relevant micro-properties realize multiple dispositions, and if a given disposition is picked out in *relational* terms, it turns out to supervene on the system of macro-relations. (The earliest explicit appearance of this idea in the literature is Dennett (1981), who argues that explanations in cognitive science often rely on 'macroreductions'.)

If one acknowledges the possibility of multiple supervenience, then one disagrees with Kim's supposition that all supervenience relations point unidirectionally to physics. This might suggest a basis for a quick answer to Kim's supervenience argument, since if you reject its implicit premise that supervenience relations must all be 'downward,' then you won't get impaled on the first horn of Kim's dilemma (see 2.1 above), because *this* kind of breakdown of supervenience has no consequences at all for the causal closure of the physical. To clarify this last claim, multiple supervenience does not imply the spooky idea that you could change the global psychological state of the world while making no physical changes *at all*. But it does imply that, even given ideal science, you couldn't necessarily predict which *particular*

physical changes would have to accompany a given psychological change; i.e., that these relations aren't, in general, systematic. Avoiding Kim simply by abandoning supervenience, though, wins a cheap victory by burying more substantial issues at stake. Kim would presumably deny that an explanation citing upward-supervenient dispositions can be a *causal* explanation; and Jackson and Pettit, in shying away from regarding program explanations as causal, presumably agree about this. This brings us to what we think is the deepest bedrock beneath the new metaphysical unease with the special sciences, with which we grapple in Section (4).

Meanwhile, however, let us press on by asking what the special sciences actually *do* that leads them to pick out entities, processes and kinds which don't end up in neat supervenience relations with physics. Meyering offers the following suggestion:

What gets studied in the special sciences is in fact huge systems of concatenated micro-systems which are systematically organized in such a way that their typical causal antecedents prompt typical patterns of causal processing to eventuate in typical effects, which in their turn serve as typical inputs for yet other causal sequences of events to take place. Regimented in this way the system produces emergent effects that have no salience at the level of physics, and yet constitute the preconditions for the recurrence of the sequence in question, or for the emergence of related processes which are significant at that same level of special science description (2000: 193-4).

For an example of 'emergent effects which have no salience at the level of physics', consider the huge collection of physical particulars which happen to constitute a given stock market crash. Such an event is clearly of considerable importance to the group of special sciences we call economics. The claim being made here is that *without* the perspective provided by the special science explanations in question there would be no way of picking out *that* collection of particulars as being an event at all. It just wouldn't be on anyone's list of 'things to be explained', any more than the particular things counting as 'money' would cry out to be classified together on grounds recognizable to physics. So, it would appear, if you want to have descriptions, let alone *explanations*, of phenomena where functional, and especially multiple, supervenience obtains, then you need to grant the irreducibility of the kinds which feature in such explanations. To be blunter still, we are faced with a choice between embracing reductionism, or being able to construct the explanations we do in fact construct.

Faced with this choice, some thinkers have supposed that there just *can't* be anything wrong with our apparently causal explanations, and hence that Kim just has to be wrong. One version of this response argues that if Kim is correct about the mental causal exclusion problem (2.1 above), then all of the special sciences are in the same trouble. Taking it as more or less self-evident that *that* can't be the case, they reason that Kim's problem isn't a real problem at all. Burge (1993), Baker (1993) and Van Gulick (1993) all offer versions of this 'generalization argument'. Kim's response is twofold: he argues that if the problem *did* generalize, to reason that there isn't a problem because we find the conclusion outrageous amounts to demanding a metaphysical free lunch, and he argues that the problem does not, in fact, generalize very far.

If it seems like the causal exclusion problem *should* generalize, it is because the supervenience argument looks like it should apply to *any* non-physical property, including

chemical, geological, biological and other special science properties. In the limit, this suggests that *all* causation should ‘seep down’ to the level of micro-physics. Kim argues that this supposition trades on vague intuitions about a hierarchy of ‘levels’ of properties, which need to be handled more rigorously. Specifically he argues that we should distinguish between the realization relation and the macro-micro relation, and, having done so, recognize that the ‘*realization relation does not track the macro-micro relation*’ for the reason that both ‘*second-order properties*<sup>15</sup> and their first-order realizers are properties of the same entities and systems’ (1998: 82). To supplement this argument Kim develops a notion of a ‘micro-based’ property (discussed above in 3.1 and below in 4.2) so as to save the physical status of ‘micro-based macroproperties’ such as hardness, transparency, conductivity, and the like, as well as the objects in which we standardly locate them such as tables, windows and nerves. Kim’s reflections here are, we think, partly salutary: physicalism ‘need not be, and should not be, identified with micro-physicalism’ (1998: 117). (Clapp (2001) develops independent arguments against misleading ‘level’ talk, in the context of a defence of non-reductive physicalism.)

Marras (2000) argues, however, that Kim’s attempt to limit the extent to which the causal exclusion problem generalizes is of limited success. At best Kim’s arguments show that the causal exclusion problem is not an *inter*-level problem, indicating that the only causation is micro-physical. What his arguments do not show is that it is not an *intra*-level problem for every individual special science. The possibility left open by Kim is that every special science is ontologically confused, in virtue of classifying the world into types which cannot be reduced to physics. In the light of what has been said above, it should be clear that the causal exclusion problem generalizes, at least, to every case of multiple realization of a functional or relational property. (In a complementary contribution Clapp (2001) shows that Kim’s argument has the ‘unsavory consequence’ that it makes *all* multiply realized properties, including most paradigmatic physical ones, illegitimate partly *because* most properties are associated with causal/functional roles.) So the question how many of the special sciences are threatened by Kim’s arguments is the question how many of them trade in multiply realized functional kinds. We think that *all* of them do, but this is not the place to defend this claim by means of an enumerative induction. A few examples, then, will have to do a lot of work. Consider water.

On Kim’s view ‘being a water molecule’ is a straightforward physical property, which he regards as the ‘micro-based’ property of ‘having two hydrogen and one oxygen atom in such-and-such a bonding relationship’ (1998: 84). This assertion is either false, or runs in the face of the practice of chemistry. A sample of liquid water does not consist only of H<sub>2</sub>O monomer molecules, but also, at any moment, of various polymeric molecules such as (H<sub>2</sub>O)<sub>2</sub>, and (H<sub>2</sub>O)<sub>3</sub>, in a condition of statistical equilibrium involving rapid reciprocating transformations (van Brakel 2000; Millero 2001, Ponce MS<sup>16</sup>). If we allow polymeric forms of H<sub>2</sub>O to count as water, then water is multiply realized, and Kim is simply wrong about what kind of property ‘being water’ is. Further, and more importantly, what chemists recognize as procedures for determining sameness or heterogeneity of substance, or establishing whether something is a pure element or a compound, are a variety of tests of which the most crucial involve attempts to separate a sample into its different constituents, and to determine whether it is hydropic under phase shifts (Needham forthcoming, Ponce MS). These procedures track relational, or dispositional properties – what it is that a sample *does* rather than what exactly it is made of. Following an account of these procedures Ponce (MS) concludes that ‘chemical kinds are not, within chemical thermodynamics, individuated by reference to their microstructure or micro-composition, but rather by reference to their macroscopic physical

properties, including their behavioral or dispositional properties.’ Water is, perhaps, an especially telling example, just because if multiple realization operates at chemical scales, then it seems more likely to manifest at larger scales, where the smaller scale variability could be inherited.

This is definitely what we see in cell biology, where strict (Kim-style) reduction to molecular biology seems impossible because key biological phenomena such as ‘signal sequences’ are multiply realized and context dependent, and because functional roles specified in biological terms are indispensable. As Kincaid (1997) argues, many different sequences of amino acids function as signals (multiple realization) but whether any given sequence does so is partly dependent on context (since the same sequences in other contexts *don’t* play the signalling role – i.e. multiple supervenience), and, furthermore, ‘signal sequences’ cannot be defined without reference to *biological* functions. (See also Hull 1972.)

No matter how far Kim’s argument generalizes, though, we will not follow those who try to call for a free lunch. We are simply after the interim conclusion that Kim’s problem, if it is a problem at all, affects almost all of the special sciences. It could well seem as though what is being argued for here is a kind of anthropocentrism, or pragmatism, where if something seems to *us* (or to chemists, biologists, etc.) like a good or powerful explanation, then, whether or not it is amenable to being reduced, it should be regarded as legitimate.<sup>17</sup> Realism about special-science types does not require any such abandonment of metaphysical seriousness, however.

Macroscopic states need be neither anthropocentric nor pragmatically justified if there is some way of making sense of their being real, in the sense of ‘real’ which involves it not being up to us whether an ontology respecting Occam’s razor would have to recognize them. Dennett (1991b), confronted with demands to take a position on whether ascriptions of beliefs should be thought of in realist terms, or as merely instrumentally justified devices, answered by offering a ‘mild realism’ in which the reality of basic physical states was unproblematic, and in which macroscopic *patterns*, understood in information-theoretic terms as structures that encode non-redundant, objective information by means of compression, could be considered real enough to settle the debate. One of us (Ross 2000) has argued elsewhere that Dennett’s position should be modified into a more thoroughgoing pattern-realism, suggesting that a pattern should be considered objectively real if and only if:

(i) it is projectible under at least one physically possible perspective

and

(ii) it encodes information about at least one structure of events or entities *S* where that encoding is more efficient, in information-theoretic terms, than the bit-map encoding of *S*, and where for at least one of the physically possible perspectives under which the pattern is projectible, there exists an aspect of *S* which cannot be tracked unless the encoding is recovered from the perspective in question (Ross 2000).

So considered, it is a contingent and empirical matter whether any particular real pattern is reducible to another, and, crucially, the question of the reality of any pattern is not to be decided on anthropocentric grounds. This is so because patterns are required to be projectible under a *physically possible* perspective, rather than a perspective which is an artifact of human perceptual or cognitive capacities, so if there is a physically possible perspective from

which some phenomenon recognized by our current working ontology could be more efficiently represented under an alternative ontology, then our current ontology is false, regardless of whether we are or are not, or shall ever be, aware of the existence of the alternative possible perspective in question.

What realist special scientists do on this view, then, is seek to find real patterns in particular domains of reality, domains defined by sets of particular structures and/or processes at some level of abstraction from fundamental physics. These patterns are what Meyerering needs to cash out his talk of ‘huge systems of concatenated micro-systems which are systematically organized in such a way that their typical causal antecedents prompt typical patterns of causal processing to eventuate in typical effects’.

A defender of Kim’s line can object that what we have just said about explanation, and the irreducibility, indeed even the objective *reality*, of irreducible functional properties, doesn’t automatically make any headway against the *causal* exclusion problem. It is, after all, in the name of solving that problem that we are supposed to ‘give up’ on these irreducible properties. That is, it is just these properties that we are supposed to learn ‘to live without’ so as to preserve a coherent and univocal concept of causation. Looked at this way, our banging the table and complaining about how difficult it would be to live without the properties isn’t a good an answer to Kim at all.

It is true that the possibility of non-reductive realism about special-science types doesn’t make *direct* headway, because it doesn’t yet say anything yet about how to show that special science generalizations invoking irreducible properties could be *really* causal. But it’s more than a *mere* request for a free lunch, since it is crucial to showing what’s at stake for the special sciences in evaluating the importance of Kim’s argument. We maintain that Kim’s position is based on serious misunderstandings about how things are in the special sciences, and in order to make our more direct argument against him we need to outline and defend what we take to be a more defensible picture. Because of his inaccurate picture of special sciences, Kim doesn’t seem to think the costs of his proposal are intolerable. We aim to show that they are utterly intolerable, requiring that we regard almost all explanatory activity in the special sciences as confused.

### **3.3 Stamp collecting**

In section (2.2) above we briefly outlined Kim’s reductionist proposal, which he urges as the proper response to his supervenience argument for the instability of non-reductive physicalism. His proposal involves ‘sundering’ the types referred to in special science explanations in accordance with the particular reductive bases for them we discover empirically. We argue now that this effectively urges us to abandon functionalism entirely, which goes against Kim’s claim to the effect that his brand of reductionism is consistent with taking multiple realizability ‘seriously’ (1998: 111).

Here, to recap, is why Kim’s proposal is supposed to include elements of functionalism. The process of reduction he describes gets started with a role property (pain, say) and proceeds via the discovery of the particular physical realizers of that property to a series of reductive identifications, ‘sundering’ the role into as many realizers as turn out to be empirically warranted. One immediate difficulty here is that without access to the role properties scientists wouldn’t know where to start looking for realizers, or what the realizers were supposed to be realizers *of*. That is, as we argued above, if they *started* from physical

particulars and were prohibited from making reference to role-properties, it's not clear that there would be any way at all for them to tell a collection of particulars that was the realizer of a functional property from one that wasn't, or to tell what manner of functional property it realized. This would be mission impossible: trying to look at some huge mass of physical detail, and hoping to be able to say at some point 'Ah ha! It's a stock market crash, and it started at *this* moment, and the proper boundaries of the physical event constituting the crash are *here*.' Kim's proposal, that is to say, requires that his metaphysically justifiable types of science are parasitic on the very types he argues are epiphenomenal.

A defender of Kim's position may point out here that Kim does allow that by "grouping properties that share features of interest to us" it is possible that "important conceptual and epistemic needs" could be served (1998: 110). Perhaps, then, what we are calling parasitism is what he would call serving an important conceptual need. This is an unsatisfying answer, though, since it leaves those hunting real causal relationships using others whose work is epiphenomenal as trackers. It also makes clear, as we suggested above, that on Kim's view the only justification for functionally motivated special science work is indeed instrumental: by doing that kind of science you help the reductionists figure out where to start digging, so as to dismantle the foundations of that very work. If we set this point aside and continue, matters only get worse for the special sciences of Kim's future world.

As we've seen, it is when empirical work turns up diversity in the realizers of some functional property that we're supposed to dismember the role-property into its parts. Let's assume that Kim's hunch, or hope, that realizers are likely to turn out to be species specific is right – then perhaps we'd sunder pain, irrespective of how well it paid its way as a single notion in behavioral science, into *pain<sub>h</sub>*, *pain<sub>m</sub>*, *pain<sub>o</sub>* (for, say, human, Martian, octopus). If we did this, we'd be proceeding as though we'd discovered (so far) that pain was actually *three* things. How, though, would we decide whether this was the case, or whether we'd really found out that there was no such thing as pain in general? Or, perhaps, that only *one* of the three was pain (in which case *which* one?), and that the other two were something else? (See also Marras 2000, 2002.)

Looked at another way, had our scientists somehow managed, despite the parasitism worry noted above, to *start* with a set of realizers (not having to work out from raw physical data what is a realizer of a function and what isn't) it's not at all obvious that they would group the realizers in the same way as they would given access to the role properties too. It could well be that, say, the realizer of Martian flatulence was structurally more like the realizer of human pain than the two pain realizers were like one another. In this case scientists working with only a collection of empirical descriptions of realizers might be expected to group the realizers quite differently, if they were to group them at all. There would be nothing to stop them supposing, like neoplatonist medical thinkers, that walnuts might be therapeutic for some brain conditions because they *look* rather like brains. Again, it seems, reference to role properties by Kim's rules has to be an instrumental necessity arising from the fact that the reductive relations are unknown at the outset of any enquiry. Worse, the enquiry proceeds by making the role properties obsolete. If we call the three imagined realizers of the pain role different versions of pain, it seems we are doing so out of a kind of nostalgia for when we thought (if we ever did) that pain was in some sense *one* thing, rather than out of clear-headed recognition of what, by Kim's lights, we subsequently discovered.

The most important reason why the costs of going down Kim's road are prohibitively great is thus that it requires, in the end, giving up on the prospect of a unitary psychology, and in fact

on any unified science referring to functionally individuated kinds. (Economics and biology are obvious instances.) As noted, Kim is willing to allow that by “grouping properties that share features of interest to us” it is possible that “important conceptual and epistemic needs” could be served. But he is also adamant that, in the end, functional properties with diverse realizers are properties “we will have to learn to live without” (1998: 106). In other words the only justification for unitary sciences having as their objects functionally individuated kinds is *instrumental*, because multiply realized properties turn out not to be metaphysically acceptable.

To hammer this point home, let us examine a real example. Consider hunger and satiety. Hunger is multiply realized (perhaps, therefore, a property Kim thinks we may have to ‘live without’), by several mechanisms with distinct effects on different parts of the brain. We can be stimulated to eat by, inter alia, the mechanical sensation of an empty stomach, glucose level monitoring by the liver, the sight of others eating, the smell or taste of novel food, and stress, not to mention combinations of these and other factors. One of the various realizers of satiety, or of ‘contra-realizers’ of hunger, it seems, is hormonal, particularly but not exclusively via peptides occurring in the gut. These play a role both in modulating other gut secretions, hence participating in the control of digestion, and in sending an ‘enough’ signal to the brain. Whether these hormones do in fact realize one or the other of these functions, or neither, or both, at any given time, depends on relational factors, so we here have a case of *multiple supervenience*. They might be active, yet we stop eating for other reasons (an artificially filled stomach triggered our mechanical sensors), or continue eating despite their action, perhaps because people around us are eating, or because we’re anxious, or because the novel dessert is more attractive than the unfinished main course.

A special science which studies what Meyering describes as ‘huge systems of concatenated microsystems’ or what we have suggested should be thought of as real *patterns*, has a shot at tracking typical patterns produced in consequence of the systematic organization of those systems. Such scientists get to make explanatory, and *predictively powerful* statements like:

When dietary variety is produced by providing a meal or diet composed of several foods, animals generally become hyperphagic relative to single-food meals or diets (Raynor and Epstein 2001).

At the risk of laboring one point we have been emphasizing, it is worth drawing attention to the term ‘variety’ in the above quotation. Pattern-hunting special scientists, such as behavioral scientists interested in motivated behavior, in this case eating behavior, are able to justify broad-scope predictive generalizations referring to ‘variety’. In humans, the variety in question is strikingly multi-modal, which is to say that the effect is stronger if the foods differ in more than one way, including taste, color, shape, smell, texture, and presentation. ‘Variety’, though, just *has* to be multiply realized (there are different ways of being different) and multiply supervening (structural features of the food may ground various dispositions, only some of which contribute to ‘variety’ in a given context), and so by Kim’s lights it is one of those properties we’re going to have to learn to live without.

That, though, just cannot be acceptable. We hereby bet the farm that any *possible* life form which metabolizes and is faced with resource scarcity will have *something*, and in all likelihood several things, playing the role of hunger, and that some of the generalizations of, inter alia, our psychology, ecology, and micro-economics, will apply to it.

Kim doesn't have a direct argument against multiple supervenience. It's off his radar insofar as it is more powerfully anti-reductionist than anything he seems willing to consider. We haven't yet shown to be him wrong. What we have done is show just how bad it would be for the special sciences were Kim's position to be generally endorsed. Now we need to look at how to disarm his argument that anything much needs to be changed in the special sciences at all.

#### **4. Taking Metaphysics Seriously**

We've stressed repeatedly that answering Kim requires taking metaphysics seriously. This does *not* mean respecting any particular *a priori* hunch about the objects of any special science. Rather, it means acknowledging such demands on the structure of scientific inquiry as transcend the disciplinary boundaries of individual special sciences, with the aim of productively applying these demands to guide interpretations of the relationships amongst hypotheses generated across separate sciences. Part of our diagnosis of what's wrong with Kim's approach is that he is mistaken about the relationship between one metaphysical problem and the work of physics, so we begin this section in (4.1) by distinguishing a number of metaphysical questions relevant to the issues at hand. In (4.2) we return to multiple supervenience, and the related questions of what to count as physical, and how to draw the macro-micro distinction in a way consistent with the account of realism offered in (3.2). In (4.3) we distinguish two ways in which 'cause' has been understood in the history of philosophy, and argue that Kim equivocates between them. Finally in (4.4) we argue that Kim erroneously supposes that physics provides us with the answer to a metaphysical question, and furthermore that he is seriously mistaken about how things are with physics.

##### **4.1 What metaphysics demands (or: How to pay for lunch)**

Clearly the metaphysical question bothering Kim is the following: *What explains the fact that the supervenience relations which do in fact hold, hold at all?* Kim thinks the answer to that question would be a solution to the causal exclusion problem (2.1 above), and his own reductionist proposal (2.2) is supposed to show how the supervenience relations hold because some stronger, reductive, relationship holds between physical facts and functional (special science) facts: the causal capacity of special science properties can be inherited from the unproblematic causal capacities of the physical properties with which we find they are identical. We agree with Kim that mere *invocation* of supervenience cannot answer the metaphysician's question about the place of mind in a physical world. If the special sciences that deal in supervenient types are not to be isolated from the rest of our scientific ontology, we must indeed be able to explain why the particular supervenience relations (both general and specific) which in fact hold, hold at all.

As just indicated the causal exclusion problem, considered very generally, is a problem about the *unity* of our scientific worldview, as briefly introduced in Section (1.2) above. In the context of the naturalistic, broadly empiricist, conception of knowledge and reality presupposed here, the task of the metaphysician, if she has any task at all, is to systematically investigate the ways in which relatively separated and special tracks of scientific inquiry 'hang together' to imply a whole greater than the sum of their respective parts. This is important not just because people like having unified world-views. Principled, if always necessarily tentative, answers to metaphysical questions are required to help scientists make sensible bets on which special-science kinds they should be trying to explain and which ones they would be better advised to try to explain *away*. Study of unification as a distinctive

enterprise can be predicted to co-vary in importance with the extent to which individual sciences develop specially. In the heyday of positivism, the demand for unification was typically given the strongest possible reading by philosophers who supposed that special-science generalizations should be logically derivable from more fundamental generalizations, and/or that all special-science types should be logically constructible from fundamental types and relations. Insisting on such ‘strong unification’ amounts to asserting reductionism as a metaphysical hypothesis, which is just what, in disagreeing with Kim, we are here rejecting. As indicated above, the history and practice of actual sciences, especially the behavioral, cognitive and life sciences, honors no such reductionist constraint. However, the history and practice of science *does* demonstrate consistent concern for unification in a weaker sense. To the extent that the conclusions of a given special science are isolated from those of all other special sciences, in the sense that their generalizations and/or ontological typologies are strictly ‘brute facts’ from all available exogenous perspectives, we find ourselves with a mystery or set of mysteries (Friedman 1974); and science is never content with mysteries.

We can try to describe the generic ambition for unification a bit more precisely by distinguishing three specific kinds of project that might collectively constitute it:

1. Identifying a unifying ontological structure that justifies the argument patterns accepted across all of the sciences.
2. Saying something genuinely enlightening about the ontologies of all sciences by reference to general structural relations of some kind.
3. Identifying the ‘glue’ that holds all objective relations in place.

Notice that none of the three metaphysical problems we have identified is necessarily about *causation* although all can be read as having something to do with it. For the time being (although see 4.3 below) to remain agnostic about whether the ‘glue’ might be something worth calling causation. In recent philosophy of science the first problem has been most strikingly associated with the work of Philip Kitcher, the latter two with that of Wesley Salmon. We discuss relevant details of their respective positions shortly.

Talk of ‘binding ontologies together,’ or of the metaphysician’s ‘universal glue’ is unabashedly metaphorical. Positivism was, among other things, an attempt to explicate unification without resort to superficial metaphor, but like most similarly motivated projects in the history of philosophy, it failed because it committed itself to claims that were too strong and specific to fit the full complexity of actual science. We won’t, then, be able to avoid metaphor here - ‘glue,’ indeed! - in trying to say what metaphysical explanation aims at. What we can do, and will, is as far as possible allow the dominant analyses in recent philosophy of science (Salmon’s and Kitcher’s) to constrain what it is that we do with our metaphor.

## 4.2 Supervenience and Physical Causation

We saw in (3.2) how the prospect of a breakdown of ubiquitous one-way supervenience struck Kim as tantamount to dualism. That shouldn’t be surprising, since we also saw in (2.1) that the first horn of the dilemma forming the supervenience argument has it that denying supervenience just *is* denying the causal closure of physics. Kim demands something stronger than general commitment to supervenience in the form of a principle requiring that there be ‘no changes without physical change’, though. He wants (see 1.1) ‘narrow’ supervenience,

where the supervening properties of some entity must supervene on its internal, or intrinsic properties. Failures of *this* kind of supervenience don't by themselves imply anything about whether physics is causally closed. One way such failures can arise, consistently with the closure principle, is from cross-classifying taxonomies, which in turn can arise from triangulational individuation of mental states, as discussed in (3.1) above.

Though triangulational individuation is *compatible* with cross-classification, it doesn't imply it. Social-ecological properties relevant to mental state individuation in the case of the pelorus operator (section 3.1) *could* be micro-based in Kim's sense. Perhaps cognitive scientists could work adequately with a system of mental-state classification sensitive to two or three micro-based taxonomies of properties among which it banned conflicts with its own coherence rules.<sup>18</sup> However, cognitive scientists are just not, as a matter of fact, trying to regiment their macro-properties in the way relevant to this scheme. (As Wallace (2003) argues, physicists faced with the logically identical issue in relating quantum-level properties to macro-properties don't try this either.) So this apparent possibility for reconciling Kim with cognitive science is worth pursuing only if Kim's independent metaphysical motivations for needing some such reconciliation are truly pressing. As we now argue, they aren't.

Kim provides no direct analysis of the concept of a physical property. Instead, as we have seen, he assumes that the domain of physical properties is antecedently clear, and then analyzes putative *non*-physical properties as micro-based macroproperties. Cognitive and behavioral scientists might imagine that this way of proceeding reflects consensus among metaphysicians, appealing to some well-established analysis of the physical. This is not the case. Recall, first, that the distinction between 'top-down' and 'bottom-up' prevalent in philosophy of science is drawn – by Kitcher and Salmon – by reference to the logic of *explanation*, not by appeal to a brute concept of the 'physical'. We need to ask, then, what makes something a macro-state, relative to some other set of states that are micro-states. It would be circular in this context to say that *M* is a macro-state relative to micro-states  $m_1, \dots, m_k$  just in case *M* is specified in terms of properties that supervene on the properties in terms of which  $m_1, \dots, m_k$  are specified; and Kim, given his project of showing that supervenience does not explain the relationship between the mental and the physical, would have to agree. Since we want to test Kim's picture against the prevailing metaphysic in general philosophy of science, we need to relate the macro-micro distinction directly to differences in kinds of explanations. This can be done following Kim's own lead, as given in his remarks on the relationship between explanation and causation quoted in (3.1) above. Let us say that *M* is a macro-state relative to  $m_1, \dots, m_k$  just in case the (mere) information that *M* obtains fails to carry information picking out a particular member of  $m_1, \dots, m_k$  as causally relevant. Let us add that *M* is a *scientifically reputable* macro-state just in case restrictions on the set of micro-states, one or more of which must have obtained, can be stated in a scientific vocabulary more general than that of the special science that generalizes over states of type *M*. The point of this way of restricting the scientifically reputable macro-states is to reflect the weak unification requirement that special sciences cannot be metaphysically comfortable in complete isolation. Thus: an individual's performing an action that constitutes a move in a social game (e.g., the pelorus operator's uttering '[Landmark x] 237') carries the information that some set of dispositions selecting the relevant action, encoded by the potentials along some synaptic pathways in that individual's brain, was available to be triggered, and was in fact triggered, by some state of affairs encoded as an instance defined according to the rules of the social interaction by some other set of synaptic pathways in that same individual's brain. By virtue of what might knowledge of the social action carry information about such generic sorts of brain processes? By virtue of the actual and particular content of some empirical theory of

mental architecture and its relations to neural structures, on the one hand, and behavior, on the other. Similarly, to pick up the final example from (3.3), to say that ‘hungry things are more likely to eat’ compresses information about a range of multiply realizable states and mechanisms, and is arguably not further compressible.

Notice that this way of analyzing the macro-micro relation is strictly relative to a particular special-science context; we have said nothing yet about what might make some state or property ‘intrinsically’ or ‘absolutely’ micro. This is because commitment to the non-isolation of special sciences does not imply commitment to the idea that all special sciences admit of hierarchical analysis in terms of one basic science. Kim, however, must suppose that there are ‘intrinsically’ micro states, since only this could justify his implicit restriction on scientifically reputable macro-states being, as it is, stronger than the one just given. The point is not that he must suppose that psychological states reduce directly to such states; rather, the claim is that if there are no such states to end a potential regress and do the ‘real’ causal work, then Kim’s supervenience argument would lead to an antinomy rather than to a disjunction with a preferred horn as he supposes. Our key question, then, is: *does (serious) metaphysics lend support to this intuition?* Non-philosophers might be disappointed, although not surprised, to hear that the answer is complicated. It requires examining some details of the tensions and complementarities in the two generic perspectives, the Kitcherian and the Salmonian, on the scientific realist’s epistemology.

Kim and his supporters can, *prima facie*, draw strong support from Salmon (1984, 1999) whose most general goal is to articulate and defend a realist interpretation of the point and nature of science. In particular, according to Salmon, science aims to describe the causal structure of the world. In the end, we’ll raise grounds for doubting that ‘causal’ is an unambiguously perspicacious word here. So let us say for now that the essence of this sort of realism seems to us to be crucial to any sort of realism worth having, and describe that essence thus: Science aims to tell us how the world is structured, that is, how its various processes and classes of entities constitute a single working machine.<sup>19</sup> In trying to describe how a machine works, a natural approach is to try to lay out its various internal processes and indicate how they influence each other. Salmon aims to justify a picture of science that, as a whole, is engaged in this project. It is a virtue of such ambitious realism that it must go beyond mere affirmation of an independently existing world and wrestle seriously with Hume’s epistemological challenge, to wit: How could anyone know, by any amount of observation, which links between processes are causal and which are not? Salmon’s answer here is that we can observe something that is precisely *diagnostic* of causation. That is, we can see that certain processes transmit information about their antecedent stages while others do not. Only the former are genuine processes. Following Reichenbach (1957), we can put this in terms of the transmission of marks. In the absence of specific structure-preserving (and, ultimately, structure-constituting) activity, entropy will eliminate marks on objects that carry information about their histories. A structure is, by definition, something that resists entropy, even briefly. Therefore, wherever marks are preserved we have structure. The goal of science is to discover the structures in nature. We can discover such structures because, as fairly sophisticated information-transducing and processing systems, we can detect, record and systematically measure mark-transmitting processes.

This is a terrifically powerful and, we think, deeply inspiring idea. It captures the core component of scientific realism – that science describes mind-independent natural structure (and activity), in an ontologically systematic way – while respecting the essence of empiricism. This latter constraint is that science has no place for inherently hypothetical

events or processes that are in principle beyond our capacity to physically detect, e.g., events on the other sides of space-like or time-like singularities, such as the interiors of black holes or the far side of the big bang, or events outside of our collective light-cone. One of us (Ross 2000) has exploited this idea to suggest a general metaphysics of existence; so we could hardly think it more important as a *metaphysical* insight. What, though, does it have to do with causation?

### 4.3 Two notions of causation

Salmon takes the idea described in the preceding section to be, first and foremost, an analysis of causation. Is it? As our remarks immediately above make clear, it is certainly an analysis of something<sup>20</sup> quite fundamental. But its primitive notion is information-transmission (in the physical and mathematical, not pragmatic, sense of ‘information’), not causation. It therefore amounts to a semantic proposal to treat causation as an information-theoretic concept. Should we accept the proposal? Since Salmon recognizes Hume’s challenge to the effect that causation cannot be picked out by some observational procedure independent of the analysis itself, this evaluation must proceed pragmatically. What effect would accepting the semantic proposal have on our broader conception of science, and of particular sciences? In particular, will it justify Kim’s intuitions about intrinsically micro-causal relations?

Kitcher (1989) provides a detailed critique of Salmon’s analysis, which we will summarize. First, we must reproduce Kitcher’s gloss of Salmon’s analysis:

(CP)  $P$  is a causal process iff there are spacetime points  $c, e$  such that  $P$  links  $c$  and  $e$  and it is possible that there should be a modification of  $P$  (modifying a characteristic that would otherwise have remained uniform) produced at  $c$  by means of a single local interaction and that the modified characteristic should occur at all subsequent points from  $c$  to  $e$  without any subsequent interaction (1989: 462).

This rests the idea of a causal process on the prior idea of a *causal interaction*, demanding an analysis of causal interactions in non-causal terms. Here is Salmon’s analysis of causal interaction:

(CI) Let  $P_1$  and  $P_2$  be two processes that intersect with one another at the spacetime point  $S$ , which belongs to the histories of both. Let  $Q$  be a characteristic that process  $P_1$  would exhibit throughout an interval (which includes subintervals on both sides of  $S$  in the history of  $P_1$ ) if the intersection with  $P_2$  did not occur; let  $R$  be a characteristic that process  $P_2$  would exhibit throughout an interval (which includes subintervals on both sides of  $S$  in the history of  $P_2$ ) if the intersection with  $P_1$  did not occur. Then the intersection of  $P_1$  and  $P_2$  at  $S$  constitutes a causal interaction if:

(1)  $P_1$  exhibits the characteristic  $Q$  before  $S$ , but it exhibits a modified characteristic  $Q'$  throughout an interval immediately following  $S$ ; and

(2)  $P_2$  exhibits the characteristic  $R$  before  $S$ , but it exhibits a modified characteristic  $R'$  throughout an interval immediately following  $S$ . (Salmon 1984: 171).

We then have a case of causal interaction between  $P_1$  and  $P_2$  at  $t$  iff there exist  $S, Q, R$  at  $t$  satisfying CI. Two features of this analysis are crucial to Kitcher's criticism. First, it depends essentially on counterfactuals: we need to be able to pick out characteristics that *would have* carried on inertially in the absence of the interaction. Second, it makes the concept of a macro-level cause depend on the idea of a micro-level cause. (This is just what Kim assumes is unproblematic.)

These two features interact to generate the main criticism.<sup>21</sup> First, note that macro-processes typically involve vast ensembles of interactions. To use Kitcher's example, if a batted baseball breaks a window, then we have, along with the interaction between the bat and the ball, interactions between the ball and gusts of wind, the ball and changes in the Moon's gravitational field, these changes and the window, etc.. We thus need to be able to pick out the *relevant* counterfactuals to identify the macro-cause, viz.,

(A) If the bat had not intersected  $P_1$  [the process that is the history of the ball's spacetime coordinates] then the momentum of  $P_1$  would have been different;

(B) If the momentum of  $P_1$  after its intersection with the bat had been different then the momentum of  $P_1$  just prior to its intersection with  $P_2$  (the window) would have been different;

(C) If the momentum of  $P_1$  just prior to its intersection with  $P_2$  had been different, then the momentum of  $P_1$  just after the intersection would have been different (specifically, the window would not have broken!) (Kitcher 1989: 471).

But how do we know to pick out *these* counterfactuals? By reference, it would seem, to what we already know about the general causal structure of the world! Notice also that if we have these counterfactuals picked out, then we might be tempted to analyze the causal process just in terms of *them*; a detour through informational considerations would seem redundant.

A defender of Salmon could reply here that his analysis takes as its proper object only an *ideal* causal process, which would be a micro-process such that, given the restricted predicates available in its physical description,  $S$  is exhaustively and exclusively *defined* by a finite set of characteristics  $Q, \dots, Q'$  and  $R, \dots, R'$ . Now, however, it seems that we must know the causal structure of the world in order to pick out the class of ideal interactions  $S$ .

We do not know if these technical problems can ultimately be solved. For our purposes here, this matters less than the general complaint Kitcher draws on the basis of them. That is: Salmon's analysis "requires that we provide an account of the way in which the causal structure of the macroscopic world results from the stringing together of elementary processes. Even if we already had such an account, the emerging picture of our causal knowledge is one in which the justification of *recherché* theoretical claims about idealized processes seems to be fundamental and our ordinary causal knowledge derivative" (Kitcher 1989: 469). Now, we do not think that this constitutes a serious objection to Salmon's substantive accomplishment, where that is interpreted as articulating the kinds of real structures in the world that science aims, in the limit, to discover. However, we *do* think that Kitcher's point has force against the idea that an analysis such as Salmon's, even if it can be made technically bullet-proof with respect to its intended sphere of application in

fundamental metaphysics, can be pressed into service as an analysis of the elaborately macroscopic, feedback-driven processes cognitive and behavioral scientists seek to characterize when they talk about mental causation, and the similarly complex causal patterns characteristic of science in general.

We will now present an alternative interpretation of Salmon's achievement, intended to shed light on what we see as the equivocal nature of the concept of causation. We take our cue here from Redhead's (1990, drawing on Kuhn 1971 and Russell 1917) discussion of causation and physics. Redhead notes that classical physics, in which forces played a crucial role, has given way to forms of physical theory in which forces have been eliminated. Redhead, we think justly, accuses those metaphysicians who wish to retain forces of anachronistically clinging to a distinction between natural and forced motion. In general relativity, says Redhead:

There is no such thing as a non-natural motion. To most physicists the old-fashioned idea of cause arises from the idea of our interfering in the natural course of events, pushing and pulling objects to make them move and so on. In modern physics there are just regularities of one sort or another (Redhead 1990: 147).

This attitude represents a principle that seems to us to be well justified by induction on the history of science. The central concepts of traditional metaphysics, including the Aristotelean distinction between natural and forced motion, are *folk* concepts. 'Eliminativism', in the usual sense of that word in the philosophy of mind, is the thesis that folk concepts tend to be progressively eliminated from scientific practice. We don't endorse eliminativism in *that* sense. The concept of mind, for example, may enable us to pick out and generalize over real patterns in nature. Furthermore, the folk concept of agent causation may be *biologically necessary* in the sense that no functioning agent could get by without it. Science will therefore have things to tell us about both minds and agent causation. However, there is no compelling reason to think that folk intuitions about which patterns, if any, must be *general* should survive as the scope of scientific knowledge widens. The concepts and axioms of Euclidean geometry pick out and organize some real patterns – the class of (approximate) physical Isosceles triangles, for example. But Euclidean geometry is not general, in the sense of describing most space adequately. We may gloss Redhead as suggesting that the concept of causation has its uses in describing the doings of agents, and perhaps in a range of other special inquiries, but that these uses do not generalize to physics.

Redhead is clearly asserting that metaphysicians *should not* use the concept of causation in talking about general physical relations. This is a stronger conclusion than we endorse. Suppose that a scheme developed from Salmon's proves empirically adequate and logically perspicacious for bringing us closer to an analysis of the universal glue that metaphysicians seek. Suppose furthermore that Salmon's use of the term 'causal structure' to describe what he is analyzing sticks, and not simply out of semantic inertia, but because – after all – the idea of analyzing causation as at bottom an informational relation isn't *silly* or *pointless*. Then it would be right to say that the concept of causation had generalized. However, it would have done so along only one or a few of the dimensions that compose its historical semantic vector. Other such dimensions, those peculiar to the concept's origins in describing the interventions of agents, would have been discarded. Alternatively, we might end up (on a similar outcome in the philosophy of science) with 'causation<sub>1</sub>' and 'causation<sub>2</sub>'. Our argument will not require a preference among these or other semantically plausible scenarios. The claim we need is merely a bit of conceptual history: that causation has its origins as a folk concept

associated with agency, and that the concept as it figures in realist fundamental metaphysics, such as Salmon's, is intended to have no such associations, since it must shed them if it is to do the work Salmon wants from it. We will then see that Kim's challenge to functionalism depends on these very associations.

Before going further, it will help briefly to substantiate this conceptual history. Prior to the modern period we find no concept equivalent or isomorphic to the kind of causal notion analyzed by Salmon.<sup>22</sup> Aristotle's efficient causation, considered apart from his wider metaphysic, maps best onto Redhead's 'pushing and pulling' of objects by agents. The full Aristotelean story, with its multiply composed causes and deep teleology, is an elaboration of the folk notion modeled on the execution of a plan for intervention by an agent. The rise of science disturbed this picture. Famously, the rationalists were led to continual controversy amongst themselves over how to relate agent-causation to mechanical accounts; thus we have Descartes's immaterial will that nevertheless exerts mechanical effects, Malabranche's occasionalism, Leibniz's pre-established harmony, and so on. We suggest that Hume's attempt to analyze causation away is best interpreted in light of this history. Following Hobbes, but with much greater sophistication, Hume sought to explain all mental activity as mechanical (Ross 1991). Furthermore, on his account all mental activity had its ultimate impetus outside the mind, in the sources of impressions. Hume's was thus the most thoroughgoing denial of Aristotelean agency observed in Western philosophy to that point. But the elaborated folk notion of causation he inherited from his tradition was rooted in the idea of the intrinsically active agent. Attempting to drain *this* element out of the concept of causation, Hume found almost nothing positive left in it; and so, in his hands, it becomes merely a superstitious over-interpretation of regularity.

Hume thus set the philosophy of science along a trajectory that, with respect to its treatment of causation, finds maturity in the analyses of Reichenbach and Salmon. On this whiggish reading of the history, Kant represents a regressive step, trying to regiment the folk notion within the necessary operations of the understanding, and positivism a recovery of the Humean path from *within* the framework of Kantian metaphysics (Friedman 1999). Had this been the only major development in theories of causation after Hume, then it would be appropriate to describe the modern history of causation as a steady re-analysis away from the original folk notion and towards an idea – whatever its exact content – that could find its conceptual gravity wholly within the framework of fundamental physics. In evaluating efforts like Salmon's *as analyses of causation*, we would then be asking, in effect, whether the concept ultimately finds a role within that framework, or is fated for elimination.

However, this post-Humean development is not *all* that has happened to causation in recent philosophy. With the rejection of positivist and behaviorist accounts of mind in the 1960s and 1970s, functionalists reasserted the metaphysical significance of the mental, in a way that was not a re-working of Kant's attempted compromise with empiricism. Functionalism –when it does not drift toward epiphenomenalism – seeks to give the mental a real and distinctive causal role. Most importantly for present purposes, it understands that role in a way that resurrects the folk idea of causation, since the minds defended by functionalists are analyzed precisely as the ontological basis for agency. Thus, while one tradition within the philosophy of science continued the project of trying to drain the agency out of causation, a parallel department worked assiduously at putting it back in! The contemporary metaphysical muddles represented by Kim, against which we are taking issue, are consequences of this double development.

As discussed in section (1), functionalists have disagreed significantly amongst themselves over *how* mental causation could best be rehabilitated. The attributionist school of thought, following Dennett, has articulated a metaphysics of mind according to which the componential analysands of mind have been progressively distanced from the microcausal gears of behavior – brains and nervous systems. This is the perspective whose consequences for behavioral and cognitive science is discussed in section (3). Philosophers in this camp can excuse themselves from any particular commitments with respect to the fundamental metaphysics of causation-in-general, *except* insofar as some particular account of such causation turns out to be essential for respecting weak unification constraints on all sciences. Describing as they are an unabashedly macroscopic set of phenomena, and with no ambitions in the direction of reductionism,<sup>23</sup> they can respond to demands for explanation of mental causation in the same way as any special scientist asked to discuss the kinds of processes picked out by the scope constraints of her discipline. Ask a geophysicist about ‘geological causation’ and you will be told about tectonic plates and flows of undersea lava and so forth. Similarly, a functionalist cognitive scientist might address issues related to mental causation by talking about feedback mechanisms, servosystematic control architectures, modules built by natural selection, neural networks simulating Von Neumann computers, and so forth. To a philosopher who regards the topic of mental causation as *essentially* a part of fundamental metaphysics, such answers will look like cases of changing the subject. However, they are no more illegitimate than the geophysicist’s similar answer to the similar question. The point here is just that a scientist’s ‘taking metaphysics seriously’ does not imply slavery to semantic legislation by metaphysicians. Saying that special sciences are sensitive to metaphysical issues isn’t to say that special sciences *are* exercises in (highly specific) metaphysical inquiry. If, in the end of the day, metaphysicians convinced us that the concept of ‘causation’ descended from Hume is more confusing than helpful, then either cognitive scientists will find other ways to talk about evolved macroscopic patterns in behavioral control built by interactions of genetic and cultural evolution, or, alternatively, we’ll collectively ‘decide’ to let the Aristotelean semantic heritage triumph, regard cognitive scientists as having provided a naturalistic analysis of agent causation, and conclude that ‘causation’ is a concept *restricted* to application in cognitive science and other disciplines that study agents. From the present state of play, it seems to us, either future semantic trajectory is possible; but neither threatens functionalism.

To return to the general point and summarize it, the history of philosophy incorporates a tension between two quite different notions of causation, both of which survive because both have been intended to serve legitimate but differing projects. On the one hand, special sciences are partly constituted by parochial types of ‘interaction-transmission’ relations, where by such relations we refer to Salmon’s ‘glue’ without pre-judging the details of the relationship between this and any causal concept as used in any particular special science. As Kitcher has emphasized, parochial, special science-relative varieties of the interaction-transmission relations are reciprocal functions of accepted explanatory schemata in the relevant sciences. Aristotelean agent causation, or folk psychological causation, is one such special interaction-transmission relation. Contemporary functionalism has significantly revised the Aristotelean or folk notion, particularly in denying the coherence of the idea of a unified ‘Cartesian’ will with direct causal capacities of its own, but there is a clear lineage relationship nevertheless. Those Kim charges with being ‘free-lunchers’ stop here, content to point out how useful this notion is. Kitcher’s analysis of the importance of unification is one aspect of paying for lunch, in that it aims simultaneously at explaining and transcending (without abandoning), the cluster of parochial interaction-transmission relations. That is, Kitcher takes the existence of this cluster *as a fact for metaphysics to explain*. This is why

Salmon, who was engaged directly in constructive metaphysics, can view his project and Kitcher's as complements.

Salmon's project, however, simultaneously continues to appeal to the *other* philosophical tradition with respect to the concept of causation and its use. As we explained above, the origin of this tradition lies in Hume's conviction that the folk concept of causation, as rooted in the kinds of program explanations peculiar to invocations of agent causation, fails to generalize. That concept therefore fails to be a suitable candidate for universal glue. Salmon is in pursuit of such glue, and this pursuit is the core of the serious metaphysician's job. What potentially confuses matters is that Salmon thinks his candidate for glue preserves enough traditional associations to make 'causation' the appropriate name for it. We argued above that this is a semantic decision on his part. It is an understandable and not unreasonable decision, but it is not forced science and it doesn't commit a follower of Salmon to thinking that an analysis of some concept of causation deployed *outside* the project of seeking universal glue must be illegitimate. As we explained, that the substance of Salmon's effort can and should survive a revision of his semantic decision. Whether or not what Salmon gives us is best described as an analysis of 'causation', his work demonstrates the need for understanding the structure of the world in terms of objective informational properties if one is to reconcile realism and empiricism. This involves no retreat from understanding Salmon's work as a deeply illuminating contribution to fundamental metaphysics; and the semantic revision need in no way obscure the fact that the contribution evolves out of earlier inquiries, such as Hume's and Reichenbach's, into the nature of causation. Ultimately, if Salmon's work has shown us the way toward a successful account of *universal* glue, then all parochial special-science causal relations must be susceptible to analysis in terms of it. This is what the definition of existence in terms of information-transmission defended in Ross (2000), and cited back in Section (3), is supposed to achieve. Here, we do not depend on the unqualified success of that analysis. Our ultimate goal – not yet accomplished – is to show that Kim's whole project relies on a metaphysical intuition about causation that is itself less secure than the role-functionalist explanatory processes it seeks to undermine. We therefore need only demonstrate alternatives to aspects of this intuition, not definitively to replace it with a better one.

However, the fact that we have had to reinterpret Salmon's project in a special way might still seem worrying. All the weight of our answer to Kim, it may appear, rests on this reinterpretation, so anyone finding it unpersuasive, or fearing that, *whatever* we choose to call Salmon's glue-candidate, it won't be compatible with the cognitive scientist's parochial concept of causation, will be unsatisfied. Here we must work on locating the burden of argument. There is, first of all, no question that Salmon analyzes what he calls causation in terms of information-transmission. So if Salmon-style analysis is to support Kim's reductionist picture, then we should be able to find an 'information-transmission exclusion problem' analogous to the causal exclusion problem. Defending such an analogy would require a justified intuition to the effect that a map of all the real causal-transmission paths in the universe has to have a simple, low-dimensional geometry, so that if information borne to a receiver by, say, the collisions of particles, were fully transduced and analyzed, then all information borne to that receiver down all other available paths would be redundant. Can any of the work in fundamental metaphysics we have discussed here ground such an intuition?

Perhaps. Suppose someone thinks that the sort of information-transmission relation necessary for performing Salmon-style analysis of special-science causal relations just *is* the causal relation delivered by physics. In that case, the fact that science does observe a rule to the

effect that special sciences are not allowed to contradict the generalizations of physics, conjoined with the view that the Salmon-style analysis in question is approximately correct, would lead straight to the intuition just presented, and hence to an information-transition exclusion problem. Kim indeed seems to believe that the serious metaphysician's master-concept of causation, to which all special-science causation concepts are then answerable, comes from physics. This, at least, would explain his convictions that micro-*physical* causes exclude mental ones, that supervenience must be a one-way relation, and that allowing some program explanations to count as causal amounts to a form of dualism. A similar assumption underlies Jackson's and Pettit's convolutions to the effect that program explanations may be 'causally relevant' but cannot be causal. Our reinterpretation of Salmon's project allows it to go forward unencumbered by this assumption that physics supplies one concept of causation for every legitimate purpose. This is a good thing for that project, because the assumption is *false* of physics as we find it.

#### 4.4 Physics and the physical

In common with much metaphysical philosophy of mind, Kim's arguments trade on a particular image of how things are with physics. This image includes commitment to the view that there is no controversy about whether the apparently causal claims of physicists are indeed causal,<sup>24</sup> and that the distinction between physics and the, 'special', sciences is simple and exclusive. Much of the bite of the causal exclusion problem arises from the contrast between physics thus understood, and the special sciences. Both assumptions, though, are false: physics in general is not enquiring into ultimate causes, and much, perhaps all, of physics consists of a large collection of special sciences.

We do not deny that there is a metaphysically important sense in which physics is fundamental. Physics is alone among the disciplines in being required to aim at generalizations that hold across all materially possible worlds (by which we mean: all worlds that could actually exist within the boundaries of the singularities, space-like and time-like, that limn in-principle observable spacetime). This leads to the fundamental asymmetry in the structure of the sciences we mentioned above, according to which no other discipline may violate currently accepted generalizations of physics, but physics itself respects no corresponding limitation. This entails a (weak) version of the principle of the causal closure of the physical: no special science may traffic in information-transmission relations, and, therefore (again, presuming the adequacy of a Salmon-style analysis), in parochial causal relations, that are spooky according to physics. But Kim, as we have shown, needs something stronger than this. The intuition that brings Salmonsque metaphysics into the service of his causal exclusion problem requires that physics *supplies* the form of general causal relation that must then generalize. But it doesn't.

Working physicists sometimes talk about causes. But, as Cartwright (1983, 1989) has argued, this is because most working physicists, most of the time, are not in search of nomic generalizations holding across the whole scope of materially possible reality. They are, therefore, working within special sciences *within* physics. If we ask, as our engagement with Kim has now forced us to, 'What is a physical cause, *in general*?', we must answer the question by reference to the part of physics that seeks generalizations good across the whole scope of the science, that is, fundamental physical theory (general relativity and quantum mechanics and electrodynamics). What turns up, by way of examples, is nothing.

In section (4.3) we referred to Redhead's remarks on the elimination of forces from physics. Redhead goes further than that, and argues that much physics has very little to do with causes. Instead of causal laws, he maintains, physicists are interested in finding 'laws of functional dependence' such as Boyle's law, where pressure and volume co-exist in certain specifiable ways without it making sense to say that one causes the other. Galileo's law describes the behavior of falling objects, but doesn't identify any general 'cause' of the displacement of objects. Acceleration, for example, just *defines* the kinematic relationship expressed by the law. A standard move is to say that the law measures a 'force,' and that *that* causes objects to fall. But, as Redhead (1990: 146) notes, since the notion of 'force' derives directly from the Aristotelean analysis, this move adds no content to Galileo's law beyond transference of an anthropocentric metaphor. Redhead intends skepticism about the idea of causation as a scientific concept altogether, but we need not go this far for the sake of the argument here. Nor need we claim that if some special branches of physics invoke parochial causal notions of their own, these must be equivalent to the Aristotelean folk notion – we claim that there is a plurality of special interaction-transmission relations, not just a folk notion and a scientific one. What is important here is the factual point Redhead makes about the practice of physics, which is that it doesn't feature the use of any *general* sort of thing – forces, fields, charges – that is a characteristic kind of cause, picked out by physicists in contrast to other possible general kinds of causes. If this is plausible where classical physics is concerned, it is surely that much more persuasive when we attend to contemporary physics.

In fact, the message obtained from careful attention to physics is about as bad for Kim's hunch as can be imagined. Loewer (2001: 323) joins us in noting that Kim requires a "generation and production conception of causation," but then writes:

The fundamental laws (for example, Schrödinger's law) relate the totality of the physical state at one instant to the totality at later instants. The laws do not single out parts of states at different times as being causally related. If  $S'$  is the microphysical state in a region  $R$  at time  $t'$  and  $t'$  is a time prior to  $t$ , then nothing less than the state  $S$  of the region  $R^*$  that fills the backward light cone of  $R$  can be said to produce  $S'$ . We cannot say that one event (one part of the physical state) produces another part since the laws do not connect parts in this way.

It gets still worse. Batterman (2000) argues that most *theoretical* (as opposed to purely manipulative) activity in physics consists in searching for physicists call 'universalities'. By this they do not mean, as a philosopher likely would, metaphysical principles necessarily holding everywhere, but *physical* facts that allow them to extract "just those features of systems, viewed macroscopically, which are stable under perturbations of their microscopic details" (129). In particular, they search for suitably abstract topological characterizations of systems in which basins of attraction emerge that corral microphysically heterogeneous processes around the universalities – for example, renormalization group fixed points among Hamiltonians in Hamiltonian-space descriptions of fluids, gases, magnets, pendulums and other diverse systems that display 'critical' behavior with respect to phase-states. Thus, far from invoking generation-and-production causal relations at the micro level to explain functional dependencies among physical states, physicists look for principled physical reasons for *ignoring* most of the aspects with which such relations might be identified.

A follower of Kim might object that this is all just epistemology. If physicists can extract useful generalizations by ignoring lower-level causal detail, just like psychologists do, then

this is all to the good; but it doesn't, and couldn't, show that necessary causal work isn't actually being done 'down there' where micro-events are generating and producing other micro-events. However, this interpretation is at best gratuitous, and at worst a contributor to confused physics. Physicists do *not* begin by identifying a micro-level of generation-and-production relations and *then* find bases for abstracting away from some of these relations. They instead engage in measurement, manipulation and re-paramaterization of whole systems until universalities emerge. Wallace (2003) argues that failure to *shake off* the intuition that 'down there,' under the level of system-level patterns that show stability within restricted measurement time-scales, lies a realm where all measurement values are definite independently of scale, leads to apparent paradoxes and pseudo-problems. For example, there is a widespread belief that the established quantum formalism is incompatible with definiteness of measurement at the macro-level – the famous problem of Schrödinger's cat – and so needs to be supplemented with empirically unmotivated parameters for finding connection principles that temporally link multiple worlds, or link multiple observers into continuous minds, so as to allow for superpositions of micro-states without corresponding macro-superpositions (a cat being simultaneously dead and alive). Such mangling of the formalism to make it square with our old metaphysical hunches has a severe cost in terms of physical theory: it "almost inevitably spoils the relativistic covariance of the theory." We best dissolve these insoluble dilemmas, Wallace suggests, by dropping the hunches and thinking of 'reality' as measurement-scale-relative patterns in structural properties of quantum states 'all the way down.' (Reading Wallace's argument alone, one might worry that his point is itself just a qualitative philosophical hunch, but this is not so. Recent work by Nottale (1993, 2000), for example, gives formal details motivated from within physics.)

The above survey of physical ideas is not intended to represent a settled picture of or a committed prediction about the metaphysical implications of contemporary physics. The point, rather, is this. Physics supplies no 'master-concept' of causation that is motivated independently of some particular explanatory program. Physics does not encourage, and may well even actively *discourage* – as Wallace effectively claims – the Salmonsque interpretation of causation-as-metaphysical-glue on which Kim relies, unless that glue is reinterpreted structurally. By 'structurally' we intend reference to networks of informational relations as suggested earlier, or to topological structures of fractal spacetime following Nottale (1993, 2000), or to something else that features a key property incompatible with Kim's hunches about physics, namely that basic measurements are indexed by global rather than local analytic procedures, in the strictly mathematical sense of these terms. (That is, measurement values are not indexed to neighborhoods of points.<sup>25</sup>) What all such interpretations have in common is that, far from undergirding the one-way local supervenience that Kim transforms into reductionism as a general metaphysical principle, they suggest it to be a scientific anachronism. There is indeed a deep tradition of 'causation as universal glue' to which Kim implicitly appeals; but that tradition has found its most sophisticated contemporary expression in analysis of causation as a special type of information-transmission or other global-structural relation.<sup>26</sup> One cannot derive a basis for insisting that physicalism implies one-way supervenience from analysis of this concept. So Kim's particular reductionist image, from which follows all the trouble for special sciences identified in Section (3), ends up resting on a vague and unmotivated hunch about an unanalyzed class of absolutely general causal relations.

The upshot of our discussion to this point is negative: we have shown that there is no rational basis for thinking that special sciences that traffic in multiply realized and multiply supervenient functional kinds should expect to have to endure conceptual revolutions under

pressure for general unification of science, or because they presently fail to track real or non-redundant causal relations, thereby missing the genuine explanations of the phenomena in their domains. We thus hope to have shown cognitive and behavioral scientists how, at least in general terms, to see off the complaints of skeptical metaphysicians. However, we noted at the outset that a more straightforward way of doing that, by a simple appeal to epistemic pragmatism, has always been available to special scientists, and is no doubt the sociologically dominant response. The justification for all the work we have been asking behavioral scientists to do in working through our argument depends on our claim that metaphysics *should* be taken seriously, that is, can actually contribute something *positive* to cognitive and behavioral inquiry. The promises made at the beginning of the discussion, therefore, have not been discharged until, now that we have swung a wrecking ball at the structures of conservative metaphysics of mind, we say something about how to build a scientifically useful structure above the rubble. To this we therefore turn in our concluding section.

## 5. Conclusion

The general consequences of the preceding discussion for behavioral and cognitive scientists can be consolidated as consisting of one negative point and one positive one. In order to state the latter from a clear basis, we begin here by summarizing the negative upshot. Recall that we identified a general and important metaphysical task as that of identifying whatever it is that holds all objective relations in place, metaphorically calling this a kind of glue. Recall also that we have distinguished two different senses of ‘cause’ discernible in the history of philosophical reflection on causation, and relevant in different ways to metaphysics and reduction. Kim’s conviction that the special sciences are answerable to physics amounts to the conviction that the particular causal claims produced by physics amount to statements about the metaphysical glue. He thinks that physical causal claims are already metaphysically unimpeachable, and that that is the reason why the causal claims of special sciences have to answer to them. But physics is itself largely composed of special sciences, physicists aren’t best seen as in the business of discovering causes, and the primacy of physics does not consist in the fact that physicists are, simply in virtue of being physicists, automatically doing fundamental metaphysics. Kim is thus multiply mistaken.

The intuition among philosophers that ‘down there’, in physics, lies an unproblematic and univocal concept of causation that can directly inform metaphysics runs very deep. As we have seen, even Jackson and Pettit, whose work has been motivated by a concern to defend and articulate the basis for the strong autonomy of special sciences, succumb to this picture when they assume – without argument or even much discussion – that program explanations, no matter how important they may be to scientific explanation, cannot be causal. We have argued, however, that the metaphysical ground is nowhere close to being sufficiently settled or uncontroversial to drive a set of conclusions as logically strong, or as troubling for scientific practice, as the new reductionists imagine. Indeed, we suggested in the last preceding section that current trends in physics and the philosophy of physics make their commitment to a *localist* conception of ‘causal glue’ look like an increasingly poor bet.

We thus claim to have shown cognitive and behavioral scientists how to see off metaphysicians who are skeptical about the explanatory adequacy of their hypotheses and conclusions on grounds that these rely on ultimately reducible or eliminable causal mechanisms. They can say: “We’re scientists, not metaphysicians. We aren’t trying to explain *in general and at one analytic stroke* how our macro-phenomena relate to micro-phenomena. This doesn’t mean that we dismiss metaphysics as irrelevant; we’d worry if you were right

that we're positing scientifically isolated mystery processes. But we don't need to integrate ourselves with other sciences by identifying mental causes with non-mental causes – there isn't any single scientific concept of causation to govern this. We'll stay integrated by piecemeal connections as we go, and if a metaphysician offers us a more general unifying principle that actually sheds potential light on our subject – mind and behavior – then we're all ears.” In light of our discussion's length and complexity, however, we can't exactly claim that this gift has come for free. Many may be inclined to think that all we have done is provided a tediously unnecessary justification for doing something they could always do, but without work: ignore philosophers. A scientist who believes that *all* metaphysics is gratuitous to her activity will not thank us for buying her a lunch she had no interest in eating.

We have operated from the assumption, however, that metaphysics can and should be taken seriously *as a part of, and for the sake of, science*. We thus owe some demonstration of payoffs in these terms. There are, we think, two. First, cognitive and behavioral scientists *do*, like most special scientists outside of physics, invoke and rely on distinctive causal concepts; but these are frequently implicit, and this implicitness can and does complicate debates over investigative methods and interpretations of conclusions. We think we are now in a position to say something enlightening about the causal concepts at work in cognitive and behavioral science. Second, special sciences, despite being separable by definition, *do* lean on one another in a variety of practically significant ways. We will be able to say something useful about the details of that too.

Virtually all models in the broad domain of cognitive and behavioral science rest on the idea that nervous systems, in interaction with environments, are engines that 'produce' behavior and perhaps – a recurrently controversial point – representations. Behaviorists, Gibsonians, some connectionists and many neuropsychologists have motivated their research programs by, and thus apparently made their importance hostage to, the conviction that representations are the wrong sorts of things for carrying ultimate causal efficacy. Those who make more robust use of representational structures typically counter their skeptical *scientific* critics by noting that computer programs manipulating representations are undeniably causally significant, and that anti-representationalist projects are simply refusing to avail themselves of helpful resources. Still, it is usually conceded, the causal sources of behavior can't be representational 'all the way down'; somewhere, somehow, there must be a level of activity analogous to that of electrical circuits in computers that fully explains the 'mental' patterns. To suppose otherwise is to allow the possibility of dualism or magical emergentism or some similarly irresponsible license for seceding from the legitimating sphere of real science. Anyone who doubts that real scientists worry about these issues, in the course of criticizing, defending and building new work upon serious models, need merely review a sample of back issues of this journal.

The set of assumptions that drives these debates is distinguished from Kim's only in being (typically) a bit less explicit. They arise because the legacy of two concepts of causation in tension is a *general* inheritance, alive in psychology as well as philosophy. Functionalist analyses of representations as making irreducible differences to behavior are piecemeal vindications of the scientific significance of old-fashioned agent causation. This sort of causation *is* disturbingly unlike the kind of modern causation by mechanical bumping or (later) magnetic or gravitational pulling that all of science is still often imagined to be 'ultimately' about. It seems to us that in much of cognitive science explanation by allowing mental control of physical behavior *is* viewed as a kind of pragmatic compromise: ever so useful for getting work done, but *someday* ...

Suppose, though, that the appropriate way of dissolving the tension is to allow a refined and sophisticated kind of agent-causation as a parochial special-science type, while giving up on the modern, generic, kind of causation. Such positive news for cognitive scientists should be as surprising – if comforting rather than threatening – to some cognitive scientists as to Kim and his followers among philosophers. Yet so far as other sciences – emphatically including physics – *and* careful metaphysical speculation are concerned, it is just as plausible to suppose that the fundamental ontological structures governing all of science are global and structural as to suppose that they are local and mechanical. Contemporary cognitive and behavioral science is dominated by accounts of feedback-driven servosystems and hypotheses about how natural and cultural selection can build and maintain them. It is very natural to suppose that such complex dynamics must be ‘built out of’ simpler processes in an additive way. This, however, is a metaphysical assumption, derived from meta-reflection on the history of science, which is not now standing up well under concerted pressure.

Here is an alternative metaphysical image: the dynamic patterns studied by the cognitive and behavioral sciences are instantiations, at particular scales of metric identification and measurement, of *more global* dynamics characteristic of the physical universe in general. Recent work in mathematical physics, information theory and analytical metaphysics shows how to make this claim relatively precise and non-fuzzy from the perspective of mathematics,<sup>27</sup> but it *is* more than a little boggling with respect to prevailing intuitions about explanation. However, the current adventures of the conservative metaphysicians can help to remind us that explanation by reference to ‘ultimate’ collisions of particles is no less *logically* puzzling; we simply grew accustomed to it during the long march from the days of Galileo and Kepler. Let us be clear: we are not here *asserting* that, as matters have turned out metaphysically, the world is made of informational topologies (or some other kind of globally structuring manifold) ‘all the way down’ and demanding that everyone sign up. Scientists are justified in their prevailing pragmatic intuition that metaphysical inquiry doesn’t generate clean, resolute satisfactions of this sort, and that this is related to the grounds on which they should keep some distance from it in their pursuit of lasting explanatory accomplishment. However, metaphysical frameworks guide science as *constraints* whether we like it or not. Furthermore, for reasons we will now briefly discuss by explicit reference to the causal concepts of cognitive science, some such constraints are, at any given time, necessary for scientific progress. It is equally crucial that these constraints be allowed to evolve, to the extent of complete replacement over time. Such constraint management is sufficiently delicate, and sufficiently important, to justify philosophical activity. We will illustrate the point by reference to some actual recent debates and activities in cognitive science.

We have already mentioned one way in which implicit localist metaphysics influences activity in cognitive science: it leads those who develop representationalist models to constrain them by the idea that they must be amenable to vindication through implementation in some set of ‘lower level’ local mechanisms. Often, all this amounts to is that a few speculative paragraphs on possibilities for such implementation get tacked onto the backs of papers describing representationalist models; and this is hardly a problem, if it is a problem at all, over which to get worked up. However, it expresses the fact that most scientists *do* feel a responsibility not to leave their models isolated from the wider, unified explanatory project. Vague speculations about implementation are, at their limit, lip-service acknowledgements of this. The principle becomes important to science when it is taken truly seriously. The leading expression of genuine commitment to the principle that has recurrently characterized work in behavioral and cognitive science is *restriction* of modeling approaches to domains of

explanation that are taken to be *already* unproblematic from the perspective of localist implementation.

Glimcher (2003) has recently given us a history of neuroscience from this explicit perspective. The Sherrington program for explaining all ‘determinate’ behavior by reference to passive reflexes, each of which responds in isolation, according to fixed condition-action rules, to a finite menu of possible stimulations, is as perfect an instance of commitment to Kim-style localism as can be found anywhere in science. Since Sherrington doubted, empirically, that all behavior is determinate in this way, he was a dualist; note, then, that his dualism was not a *directly* metaphysical thesis, but a *scientific* response given an implicit metaphysical constraint on hypothesizing. More interesting, as Glimcher shows, is that grounds for doubt about the capacity of pure reflexology to explain even the plausibly ‘determinate’ behavioral patterns were made evident (by Graham Brown) during Sherrington’s lifetime, and more systematic critiques in the mid-twentieth century, both theoretical and experimental, by von Holtz, Mittelstaedt, Weiss and Bernstein accumulated decisive refutation. Nevertheless, Glimcher argues, the most emulated and productive neurophysiological investigations *right now* - connectionist models of learning using backpropagation, and Shadlen *et al*’s (1996) celebrated work on visual perception of motion in monkeys are Glimcher’s examples – continue to honor Sherrington’s localist paradigm. The contemporary work of course invokes a range of new mechanisms Sherrington could not have imagined; but the commitment to input-driven, localized, non-hierarchically governed processes remains in place.

Glimcher’s point is not that these studies don’t merit celebration or emulation. His point, rather, is that neuroscientists, despite knowing that localism in their domain isn’t generally true, and despite their not being willing to allow dualism, concentrate their best energies on such phenomena as *can* best be modeled in localist terms. We now suggest that this should be interpreted not as a retreat from unification with other special sciences but as an indication of extremely serious commitment to it *given a prevailing, mostly implicit, localist metaphysic*.<sup>28</sup>

With Kitcher, Friedman, Kincaid and other philosophers we have cited approvingly in the course of this paper, we agree that special scientists are right to care about avoiding completely isolated explanations. Here an overt normative principle is in order: If what science mainly delivered was a chaos of scattered descriptions of unrelated phenomena, we would be justified in feeling crushingly disappointed in it. This would be science as, at best, a pure under-laborer to engineering, not a collective project for increasing our understanding of the universe. Like Glimcher, we intend no criticism of cognitive scientists who respect this norm by doing powerful work that preserves unity by leaving wider metaphysical assumptions unchallenged. However, explanation is no virtue if we don’t care whether explanations are, in addition to being comprehensible, *true*. This implies that we mustn’t leave metaphysical presuppositions unchallenged in practice unless philosophical reflection convinces us that the presuppositions in question are actually justified. The justification of a metaphysical presupposition should rely mainly on its fruitfulness in science, so commitment to localism was a healthy restriction for a long time. But in some sciences – in physics, in economics, in many parts of biology and cognitive science – that time has passed. Recent experience and reflection suggests that explanation of local phenomena as instances of global dynamic structures is a viable alternative route to unification.

This is just what Kim and the conservative metaphysicians deny. Kim’s commitment to fundamental metaphysics is expressed as the insistence that one does not solve the mind /

body problem by offering particular accounts of intentional processes in non-intentional terms. Rather, one must try to explain how and why in general “mental properties and physical properties are related, and hopefully also explain why they are so related” (Kim 1998: 5). We agree. We also agree that functionalism cannot be vindicated as providing such an account by mere appeal to supervenience. But functionalism *could* license agent-causation as a legitimate, special-science-parochial, sort of causation after all, if more general accounts of informational or other topological dynamics can show it to be non-mysterious – and the prospects here look promising (see Juarrero 1999). Surely we have, up to an important standard of generality, explained how mental properties and physically non-problematic properties are related if we produce a broad account of feedback-driven servosystems and the ways in which evolution has built nervous systems that support them. If dynamic systems theory *is* a way of doing metaphysics – and that is what we are suggesting – then servosystematic control without localist reduction is not isolated as a basis for explanation (and the same goes, in spades, for evolution). Here, perhaps, is the source of the technical tools through which the special problem of mental causation and the general questions about universal glue find a common logic of address. But the working problem of mental causation, as we see it, is the very old problem of how agency is possible. ‘Causation’, in this context, means something special: the processes, whatever they are, by means of which thoughts and decisions, beliefs and desires, make a real difference in the world. We see no reason to believe that there is any more ‘general’ a way of addressing this problem than by the approach of contemporary behavioral science, with its plethora of servosystematic control processes grounded in neuroscience and ethology.

Kim’s problematic is what you get if you reify the folk and the post-Humean concepts of causation. On the one hand, you find yourself wanting to show how interventions by agents can make a difference to what actually happens in the world. But then, on the other hand, you insist that these interventions must be micro-processes, or decomposable into micro-processes, that agents must not just turn out to be programs. Well, we think it’s overwhelmingly likely that agents are programs, and they aren’t anything else. Some mental states are reliable bearers of information about other mental states, even though no particular state in the supervenience base of one is a reliable bearer of information about any other particular state in the supervenience base if the other. If something is a running series of such states, then it’s a program, something that acts and exists *by* compressing information. Indeed, living systems are only possible at all thanks to the fact that some of their states, including mental states in those with brains, extract and emit useful (accurate) information in compressed form. That they *can* do this is empirically evident. It is also not, pace Kim, mysterious: thanks to the dynamics of servosystematic feedback structures, multiple supervenience is possible (and actual).

This contradicts nothing that physicists either presuppose or tell us. Physicists, like all scientists, study patterns of compressed information at whatever scales they can be found, not, by elision, some non-existent level of ‘ultimate’ micro-banging and colliding. This is good news, we take it, for most cognitive scientists. But the full news is even better. When we ‘do’ metaphysics in the naturalist’s way – by standing back and looking hard at collections of special sciences in abstraction – then moving our attention from the cognitive sciences to the physical ones doesn’t involve a discontinuous leap from spooky or redundant causal relations to good old-fashioned mechanical ones. We see instead convergent dynamical accounts that can be swapped across the boundaries of the many special sciences in a profoundly interdependent intellectual market.

### **Acknowledgements:**

Previous versions of this paper have been presented at conferences in Dubrovnik, Stellenbosch and Ghent. We thank the audiences on those occasions. We are also grateful to John Collier, Daniel Dennett, Harold Kincaid, James Ladyman, Ausonio Marras, Veronica Ponce, and Alex Rosenberg for critical feedback and comments, to the editors and referees of this journal for comments and criticisms of an earlier version of the paper, and to Nelleke Bak for careful reading of the text.

### **Notes**

- <sup>1</sup> The phrase is due to Bickle (1998). However, we will not here be engaging with Bickle's interesting thesis, which has enough direct empirical content to be a piece of cognitive science in its own right. The philosopher who has done most to inspire the backwash is Jaegwon Kim, and his most influential argument, as given in Kim (1998) will be the target of our discussion.
- <sup>2</sup> Or, at least, so philosophers often say. As a claim about actual behaviorist psychologists, this claim is largely nonsense, flatly untrue of, for example, E.C. Tolman or Karl Lashley. But the importance given to knocking over this straw man in the history of the rise of functionalism is indisputable, and is what is of relevance to us here. We would encourage more footnotes like this one in the philosophical literature, however.
- <sup>3</sup> Functionalism, thus understood, can be a kind of behaviorism - just one allowing for some intermediate behaviors between stimulus and response. Our own favored variety of functionalism is in fact of this behaviorist sort; but this will not play any direct role in our argument in this paper.
- <sup>4</sup> We use this example because it is standard in the literature we are describing. We should point out, though, that the philosophers who introduced it knew from the outset that it was at best neurologically implausible, and were using it as a place-holder for some imagined future reduction of a psychological to a neurological state.
- <sup>5</sup> The same argument structure has also been (e.g. Horgan 1997) used to argue for functionalism within a species (on the basis of significant neural and other differences between conspecifics), or, over time, within a single individual.
- <sup>6</sup> There are various particular ways of being a realizer functionalist in the broad sense indicated here. One, particularly strong, way is via the 'functional analysis' strategy associated with Armstrong and Lewis, as discussed in section (2.2). Another way, canonically defended in Pylyshyn (1984), requires that types be individuated either by reference to intrinsic properties of members of the type, or by reference to intrinsic properties of independently specifiable sets of tokens of the type.
- <sup>7</sup> On our broad conception of realizer functionalism; see note (6) above.
- <sup>8</sup> Kim (1998) is not the first expression of the problem, merely an elegant, sophisticated and up to date version of it. Yablo (1992) is a clear and widely cited statement of the

issues as of the early 1990s, and the papers in Kim (1993) show many of the lines of argument and thinking that lead up to Kim (1998).

9 We assume throughout that we're talking amongst people who would regard the admission of supernatural causes into science as the end of the world.

10 Talk of 'enhancing' is somewhat sloppy, as Marras (2002) points out, but the details need not detain us here.

11 The philosophical literature on explanation is enormous, and so some philosophers might object to our announcing that we can boil it down to consideration of just two approaches. A few meta-comments on that literature are therefore in order. It divides naturally into two piles. The first pile, concerned directly with the way in which the search for explanation descriptively and normatively guides scientific activity, really *does* mainly revolve around the dialectic established by Kitcher's and Salmon's long argument with each other (see Kitcher and Salmon 1989). The second pile, highlights of which include van Fraassen (1980), Garfinkel (1981), and Achinstein (1983), concerns the logic of explanatory statements. Both piles descend from the classic work on explanation in philosophy of science by Hempel (1965), which, in the way of positivism, saw these two concerns as indistinguishable. To a post-positivist of whatever stripe, however, they are distinct, and to a considerable extent orthogonal. That is, just about any combination of views from the first and second debates can be made compatible. (See, for example, Kincaid 1997, Chapter 5, whose work depends on subtle recombinations of them.) For our purposes in this paper, only the first set of issues about explanation are directly relevant. Non-philosophers are cautioned, however, against taking our summary as a mini-survey of the whole literature on the subject.

12 As Batterman (2000: 118, n. 4) notes, "Kim's argument won't go through unless the causal properties of the macroproperties just are the resultant (or 'sum') of the microstructural properties."

13 For a sample of the literature urging this perspective, see McClamrock (1995); Wilson (1995); Clark (1997); and, especially influential with respect to what we say here, Dennett (1991a). Pettit (1993) provides the most systematic, though very cautious, investigation of these ideas.

14 According to Menzies (1988) this line of argument was suggested by Lewis.

15 i.e. A property possessed by an object (such as dormativity) in virtue of its having some more basic (e.g. chemical) properties.

16 We are especially indebted to Ponce's treatment here.

17 Clapp (2001) successfully argues that some leading defenses of the autonomy of special sciences, such as Fodor's (1974), are guilty of this lapse of metaphysical seriousness. We should therefore note explicitly that none of our arguments in this paper depend on the idea that the kinds of any special science must be preserved as kinds just *because* people find it useful to think with the concepts they represent. Indeed, on one interpretation this is what 'taking metaphysics seriously' in our sense here *means*.

18 Stich (1983) devoted a book to arguing that this sort of picture, intended as a way of  
reconciling a plausible cognitive scientific typology of states with folk psychology,  
couldn't work. Kim must believe that Stich is wrong about this.

19 Cartwright (1983, 1999) has famously argued that the world is *not* a single, working  
machine, but is instead 'dappled', by which she means ontologically disunified. Dupré  
(1993) has urged a similar thesis. For reasons given in Spurrett (1999, 2001a), we  
reject this conclusion. The fact that science is never finished, and therefore never  
completely unified, may mean that its current description of the world at any given  
time will always be of a world that is 'dappled'; but to derive *as a metaphysical  
conclusion* the claim that the world *is* dappled is to simply abandon the regulative  
ideal that informs Salmon's project, and, for that matter, Kim's. Answering Kim *this  
way would* simply amount to shrugging off the significance of realist metaphysics,  
another way of trying to have lunch for free.

20 The 'something', we would say, is indeed fundamental structure; Ross (2000) takes it  
to be the network of Schrödinger-style negentropic relations. That network is *our*  
favorite candidate for universal glue.

21 Kitcher develops, at length, additional criticisms based on counterexamples to  
Salmon's technical criteria for distinguishing genuine causal processes from pseudo-  
processes. We will not incorporate these into our summary here, since they contribute  
little to the issues relevant to our discussion, and since even if Salmon's apparatus is  
repaired so as to block the counterexamples, Kitcher's main critique is unaffected.

22 We rely here – with liberal interpretive work of our own – on Sorabji (1988).

23 It is common outside philosophy for Dennett to be called a 'reductionist' because he  
analyzes intentionality and consciousness without recourse to any entities or processes  
incompatible with the causal closure of physics. However, Dennett in fact denies, like  
us, that there is any *general* relation between physics and special sciences stronger  
than global supervenience; and in the context of most debates in philosophy of  
science, this makes him as anti-reductionist as the recent tradition allows. Thus, for  
example, when Kincaid (1997, 86-90) defends anti-reductionism, he feels he needs to  
spend a few pages showing that he needn't go as *far* in that direction as the 'radical'  
Dennett. Ross (2000) explains in detail the sense in which Dennett's anti-reductionism  
is radical.

24 Philosophers typically grant that our current physical theories are open to revision, so  
the point here is slightly more complicated. Still, for philosophers of mind an ideal  
physicist is generally assumed to be making unproblematically causal claims, whereas  
an ideal economist, say, would need to do additional philosophical work over and  
above her economics to justify thinking of her claims as causal.

25 We owe this insight to Andrei Rodin.

26 The logic of this, and comparison of the causation concept's role in different branches  
of science, is made formally explicit in a recent paper by Thalos (2002).

27 We allude to the earlier references to the work of Nottale (1993, 2000).

28 This point has also been vividly argued by Dennett (1991a).

## References

- Achinstein, P. (1983) *The Nature of Explanation*, Oxford University Press.
- Armstrong, D. (1981) *The Nature of Mind and Other Essays*, Cornell University Press.
- Baker, L. R. (1993) Metaphysics and Mental Causation. In: *Mental Causation*, ed. J. Heil and A. Mele, Clarendon Press.
- Batterman, R. (2000) Multiple Realizability and Universality. *British Journal for Philosophy of Science* 51: 115-145.
- Bickle, J. (1998) *Psychoneural Reduction: The New Wave*, MIT Press / Bradford.
- Birks, J. B. (1963) *Rutherford at Manchester*, W. A. Benjamin.
- Block, N. (1980a) Introduction: What is Functionalism? In: *Readings in the Philosophy of Psychology* (volume 1), ed. N. Block, Methuen.
- Block, N. (1980b) Troubles with Functionalism? In: *Readings in the Philosophy of Psychology* (volume 1), ed. N. Block, Methuen.
- Block, N. and Fodor, J. (1972) What Psychological States are Not. *Philosophical Review* 8(2): 159-81.
- Brooks, R. A. (1991) Intelligence without Representation. *Artificial Intelligence* 47: 141-160.
- Burge, T. (1993) Mind-Body Causation and Explanatory Practice. In: *Mental Causation*, eds. J. Heil and A. Mele, Clarendon Press.
- Cartwright, N. (1983) *How the Laws of Physics Lie*, Oxford University Press.
- Cartwright, N. (1989) *Nature's Capacities and their Measurement*, Clarendon Press.
- Cartwright, N. (1999) *The Dappled World*, Cambridge University Press.
- Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press.
- Churchland, P. (1981) Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78: 67-90.
- Clapp, L. (2001) Disjunctive Properties: Multiple Realizations. *Journal of Philosophy*, 98: 111 – 136.
- Clark, A. (1997) *Being There*, MIT Press / Bradford.
- Dennett, D. (1981) Three Kinds of Intentional Psychology. In: *Reduction, Time and Reality*, ed. R. Healey, Cambridge University Press. Reprinted in Dennett (1987).

- Dennett, D. (1987) *The Intentional Stance*, MIT Press / Bradford.
- Dennett, D. (1991a) *Consciousness Explained*, Little Brown.
- Dennett, D. (1991b) Real Patterns. *Journal of Philosophy*, 88: 27-51.
- Dennett, D. (2001a) Are We Explaining Consciousness Yet? *Cognition* 79: 221-237.
- Dennett, D. (2001b) The Zombic Hunch: Extinction of an Intuition? In: *Philosophy at the New Millennium.*, ed. A. O'Hear, Cambridge University Press.
- Dupré, J. (1993) *The Disorder of Things*. Harvard University Press.
- Elder, C. (2001) Mental Causation versus Physical Causation: No Contest. *Philosophy and Phenomenological Research* 62 (1): 111-127.
- Fodor, J. (1968) *Psychological Explanation*, Random House.
- Fodor, J. (1974) Special Sciences, or the Disunity of Science as a Working Hypothesis. *Synthese* 28: 77-115.
- Fodor, J. (1975) *The Language of Thought*, Harvard University Press.
- Fodor, J. (1987) *Psychosemantics*, Cambridge, Mass.: MIT Press.
- Fodor, J. (1994) *The Elm and the Expert*, MIT Press / Bradford.
- Friedman, M. (1974) Explanation and Scientific Understanding. *Journal of Philosophy* 71: 5-19.
- Friedman, M. (1999) *Reconsidering Logical Positivism*, Cambridge University Press.
- Garfinkel, A. (1981) *Forms of Explanation*, Yale University Press.
- Gintis, H. (2000) *Game Theory Evolving*, Princeton University Press.
- Glimcher, P. (2003) *Decisions, Uncertainty, and the Brain*, MIT Press / Bradford.
- Hempel, C. (1965) *The Logic of Scientific Explanation*, Free Press.
- Horgan, T. (1997) Kim on Mental Causation and Causal Exclusion. *Philosophical Perspectives*, 11: 165-184.
- Hull, D. (1972) Reduction in Genetics – Biology or Philosophy? *Philosophy of Science*, 39: 491-499.
- Hutchins, E. (1995) *Cognition in the Wild*, MIT Press / Bradford.
- Jackson, F., & Pettit, P. (1988) Functionalism and Broad Content. *Mind* 97: 381-400.

- Jackson, F., and Pettit, P. (1990) Program Explanation: A General Perspective. *Analysis* 50: 107-117.
- Juarrero, A. (1999) *Dynamics in Action*, MIT Press / Bradford.
- Kim, J. (1993) *Supervenience and Mind*. Cambridge University Press.
- Kim, J. (1998) *Mind in a Physical World*, MIT Press / Bradford.
- Kincaid, H. (1997) *Individualism and the Unity of Science*, Rowman and Littlefield.
- Kitcher, P. (1976) Explanation, Conjunction and Unification. *Journal of Philosophy* 73: 207-212.
- Kitcher, P. (1981) Explanatory Unification. *Philosophy of Science* 48: 507-531.
- Kitcher, P. (1989) Explanatory Unification and the Causal Structure of the World. In: *Scientific Explanation*, eds. P. Kitcher and W. Salmon, University of Minnesota Press.
- Kitcher, P., & Salmon, W., eds. (1989) *Scientific Explanation*, University of Minnesota Press.
- Lewis, D. (1972) Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50: 249-258.
- Lewis, D. (1980) Mad Pain and Martian Pain. In: *Readings in the Philosophy of Psychology* (volume 1), ed. N. Block, Methuen.
- Loewer, B. (2001) Review of Kim: *Mind in a Physical World*. *Journal of Philosophy* 98: 315-324
- Marcus, E. (2001) Mental Causation: Unnaturalized but not Unnatural. *Philosophy and Phenomenological Research* 63 (1): 57-83.
- Marras, A. (2000) Critical Notice of Kim: *Mind in a Physical World*. *Canadian Journal of Philosophy* 30:137-160.
- Marras, A. (2002) Kim on Reduction. *Erkenntnis* 57(2): 231-257.
- McClamrock, R. (1995) *Existential Cognition*. University of Chicago Press.
- McGinn, C. (1991) *The Problem of Consciousness*. Oxford University Press.
- Menzies, P. (1988) Against causal reductionism. *Mind* 98: 551-574.
- Meyering, T. (2000) Physicalism and Downward Causation in Psychology and the Special Sciences. *Inquiry*, 43:181-202.
- Millero, F. J. (2001) *The Physical Chemistry of Natural Waters*, Wiley-Interscience.
- Nagel, E. (1961) *The Structure of Science*. Harcourt, Brace and World.

- Needham, P. (forthcoming) The Discovery that Water is H<sub>2</sub>O. *International Studies in the Philosophy of Science*.
- Nottale, L. (1993) *Fractal Space-Time and Microphysics: Towards a Theory of Scale-Relativity*, World Scientific.
- Nottale, L. (2000) Scale Relativity, Fractal Space-Time and Morphogenesis of Structures. In: *Sciences of the Interface: Proceedings of International Symposium in Honor of O. RöSSLer*, eds. H. Diebner, T. Druckrey and P. Weibel, Genista.
- Oppenheim, P. and Putnam, H. (1958) Unity of Science as a Working Hypothesis. In: *Minnesota Studies in the Philosophy of Science*, vol. 2. eds. H. Feigl, G. Maxwell and M. Scriven, University of Minnesota Press.
- Papineau, D. (1993) *Philosophical Naturalism*. Blackwell.
- Pettit, P. (1993) *The Common Mind*. Oxford University Press.
- Place, U. T. (1956) Is consciousness a brain process? *British Journal of Psychology* 47: 44-50.
- Ponce, V. (MS) Chemical Kinds, Microscopic Essences, and Chemical Laws.
- Putnam, H. (1963) Brains and Behavior. In: *Analytical Philosophy*, Second Series, ed. R. Butler, Basil Blackwell & Mott.
- Putnam, H. (1967a) Psychological Predicates. In: *Art, Mind and Religion*, eds. Captain, W. H. and Merrill, D. D. University of Pittsburgh Press.
- Putnam, H. (1967b) The Mental Life of some Machines. In: *Intentionality, Minds and Perception*, ed. Hector-Neri Castañeda, Wayne State University Press.
- Putnam, H. (1975a) Philosophy and Our Mental Life. In: *Mind, Language and Reality: Philosophical Papers*, vol.2. Cambridge University Press.
- Putnam, H. (1975b) *Mind, Language and Reality: Philosophical Papers*, vol.2, Cambridge University Press.
- Pylyshyn, Z.W. (1984) *Computation and Cognition: Towards a Foundation for Cognitive Science*, MIT Press.
- Raynor, H. A. and Epstein, L. H. (2001) Dietary Variety, Energy Regulation, and Obesity. *Psychological Bulletin* 127 (3): 325-341.
- Redhead, M. (1990) Explanation. In: *Explanation and Its Limits*, ed. D. Knowles, Cambridge University Press.
- Reichenbach, H. (1957) *The Philosophy of Space and Time*. Dover.

- Ross, D. (1991) Hume, Resemblance and the Foundations of Psychology. *History of Philosophy Quarterly* 8: 343-456.
- Ross, D. (1997) Critical Notice of Ron McClamrock: *Existential Cognition*. *Canadian Journal of Philosophy* 27: 271-284.
- Ross, D. (2000) Rainforest Realism: A Dennettian Theory of Existence. In: *Dennett's Philosophy: A Comprehensive Assessment*, eds. D. Ross, A. Brook and D. Thomposon , MIT Press.
- Ross, D. (2001) Dennettian Behavioural Explanations and the Roles of the Social Sciences. In: *Daniel Dennett*, eds. A. Brook and D. Ross, Cambridge University Press.
- Ross, D. (forthcoming) Chalmers's Naturalistic Dualism: A Case Study in the Irrelevance of the Mind-Body Problem to the Scientific Study of Consciousness. In: *The Mind as Scientific Object*, eds. C. Erneling and D. Johnson, Oxford University Press.
- Rowlands, M. (1999) *The Body in Mind*, Cambridge University Press.
- Russell, B. (1917) On the Notion of Cause. In: *Mysticism and Logic*, Allen and Unwin.
- Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton University Press.
- Salmon, W. (1990) Scientific Explanation: Causation and Unification. *Critica Revista Hispanoamericana de Filosofia* 22: 3-21.
- Salmon, W. (1999) *Causality and Explanation*. Oxford University Press.
- Shadlen, M., Britten, K., Newsome, W., & Movshen, J. (1996) A computational analysis of the relationship between neuronal and behavioural responses to visual motion. *Journal of Neuroscience* 16: 1486-1510.
- Smart, J. J. C. (1959) Sensations and Brain Processes. *Philosophical Review* 68, 141-156.
- Sorabji, R. (1988) *Matter, Space and Motion*. Cornell University Press.
- Spurrett, D. & Papineau, D. (1999) A note on the completeness of 'physics'. *Analysis* 59 (1): 25-29.
- Spurrett, D. (1999) *The Completeness of Physics*. Doctoral dissertation at the University of Natal, Durban.
- Spurrett, D. (2001a) Cartwright on Laws and Composition. *International Studies in the Philosophy of Science* 15 (3): 253-268.
- Spurrett, D. (2001b) What Physical Properties Are. *Pacific Philosophical Quarterly* 82(2): 201-225.
- Stich, S. (1983) *From Folk Psychology to Cognitive Science*. MIT Press / Bradford.

- Thalos, M. (2002) The Reduction of Causal Processes. *Synthese* 131: 99-128.
- van Brakel, J. (2000) The Nature of Chemical Substances. In: *Of Minds and Molecules: New Philosophical Perspectives on Chemistry*, eds. Nalini Bhushan and Stuart Rosenfeld, Oxford University Press.
- Van Fraassen, B. (1980) *The Scientific Image*, Oxford University Press.
- Van Gulick, R. (1993) Who's in Charge Here? And Who's Doing All the Work? In: *Mental Causation*, ed. J. Heil and A. Mele, Clarendon Press.
- Wallace, D. (2003) Everett and Structure. *Studies in History and Philosophy of Science B: Studies in History and Philosophy of Modern Physics* 34: 87-105.
- Wilson, R. (1995) *Cartesian Psychology and Physical Minds*, Cambridge University Press.
- Yablo, S. (1992) Mental Causation. *Philosophical Review* 101 (2): 245-280.