International Studies in the Philosophy of Science
Vol. 19, No. 3, October 2005, pp. 319–333

Routledge
Taylor & Francis Group

Against Functional Reductionism in Cognitive Science

Muhammad Ali Khalidi

Functional reductionism concerning mental properties has recently been advocated by Jaegwon Kim in order to solve the problem of the 'causal exclusion' of the mental. Adopting a reductionist strategy first proposed by David Lewis, he regards psychological properties as being 'higher-order' properties functionally defined over 'lower-order' properties, which are causally efficacious. Though functional reductionism is compatible with the multiple realizability of psychological properties, it is blocked if psychological properties are subdivided or crosscut by neurophysiological properties. I argue that there is recent evidence from cognitive neuroscience that shows that this is the case for the psychological property of fear. Though this may suggest that some psychological properties should be revised in order to conform to those of neurophysiology, the history of science demonstrates that this is not always the outcome, particularly with properties that play an important role in our folk theories and are central to human concerns.

1. Introduction

After a hiatus of a few decades, various forms of reductionism have witnessed a revival in current philosophical work on the mind. In some of his most recent writings, Jaegwon Kim has championed functional reductionism as a solution to the problem of the 'causal exclusion' of the mental. In response to concerns about whether psychological properties can be truly causally efficacious, given that the underlying neurophysiological properties can be seen to be doing all the work on their behalf, Kim has argued that psychological properties can be functionally reduced to these neurophysiological properties, which are in fact causally efficacious. Since the former can be reduced to the latter, this implies that psychological properties have no causal powers of their own. In this paper, I will take a closer look at functional reduction, with a particular view to seeing whether it is a viable strategy in psychology and cognitive science. I will argue

Muhammad Ali Khalidi is at the American University of Beirut, Lebanon
Correspondence to: Muhammad Ali Khalidi, Department of Philosophy, American University of Beirut, Beirut 1107 2020, Lebanon. Email: khalidi@aub.edu.lb

ISSN 0269–8595 (print)/ISSN 1469–9281 (online)
DOI: 10.1080/02698590500462448

that even though functional reduction can be squared with the multiple realizability of the psychological in the neurophysiological, there is some evidence of a more radical 'mismatch' between the former and the latter, which would preclude the possibility of functional reduction of the psychological to the neurophysiological. I will cite evidence from a recent research program in cognitive science to support this claim. Since psychological and neurophysiological properties are mismatched, this blocks any attempt at functional reduction. The phenomenon of mismatched properties is not always clearly distinguished from multiple realizability, though it is importantly different, as I will try to show.

In Section 2, I will briefly rehearse Kim's causal exclusion problem and say why he thinks that functional reduction resolves it; then I will take a closer look at the nature of functional reduction as explicated by David Lewis. In Section 3, I will present a case study from cognitive science that illustrates the mismatch between psychological and neurophysiological properties, thus blocking functional reduction. Then, in Section 4, I will show in more detail why this case study counts against functional reductionism, and will argue that it would not be prudent to revise our psychological predicates in all such cases so that they line up with our neurophysiological ones. Finally, although I will not attempt to offer a solution to the causal exclusion problem, I will speculate in Section 5 that it might be an occasion for rethinking some of our intuitive ideas about causation.

2. Causal Exclusion and Functional Reduction

Though Jaegwon Kim is perhaps its most vigorous exponent, the 'causal exclusion argument' has been rehearsed many times in the philosophical literature in the past decade. For our purposes, it is enough to outline the main steps of the argument, based on one of Kim's recent expositions (see Kim 2003, 155–59). In ordinary cases of mental causation, a mental event instantiating some mental property M causes another mental event instantiating some other mental property M*, for example my *fear* of snakes causes my *desire* to exit the herpetology section at the zoo. But, if we subscribe to the *supervenience* of the mental on the physical, then M* has some physical property P* as its supervenience base. Indeed, it would seem as if M caused M* by causing P*.[1] But it is also true, because of the *causal closure* of the physical domain, that P* also has a physical cause P, which occurs at the same time as M. If we subscribe to the *irreducibility* of the mental to the physical, then we must hold that M is not identical to P. Thereby, we have competition between M and P, and since this is not a case of causal overdetermination, one of the two must exclude the other. If we adhere to the causal closure of the physical domain, then P must be chosen over M as a candidate for causing M*, thereby excluding M.

Kim (2003, 165) tells us that: 'The real aim of the argument … is not to show that mentality is epiphenomenal, or that mental causal relations are eliminated by physical causal relations; it is rather to show "Either reduction or causal impotence"'. This is because the irreducibility of the mental to the physical was crucially used as a premise in the argument, as the above exposition shows (other crucial premises have also been

flagged by being italicized). Accordingly, Kim's recent response to the argument is to block it simply by rejecting the irreducibility premise, thereby embracing reductionism. In fact, he holds that psychological properties are 'higher-order' functional properties defined over 'lower-order' properties. The higher-order properties are generally possessed by just those entities that possess the lower-order properties, and are functionally reducible to them. Since higher-order properties are reducible to lower-order properties, they must be identical with them. Indeed, since they are identical, the higher-order properties should not be considered to be distinct properties at all; they can be thought of as higher-order concepts or expressions that do not introduce new properties into the world. Strictly, one should speak here of higher-order *predicates* that quantify over lower-order properties (Kim 1998, 104–106).[2] Therefore, the problem of causal exclusion can be resolved in the case of psychology or cognitive science. Indeed, the problem of causal exclusion is not even allowed to crop up because higher-order properties are not real properties in their own right; they are predicates logically constructed out of predicates that denote lower-order properties, and can be analyzed in terms of them via a process of functional reduction. Naturally, therefore, they do not have their own causal powers and do not causally exclude lower-order properties.

Kim's functional *reduction* is quite different from the more familiar versions of functionalism. He asserts clearly that functionalism of the traditional variety is a non-reductive thesis, whereas he adopts a reductionist line (Kim 1998, 7–8). Typically, we think of functionalism in the philosophy of mind as giving an account of a mental state such as fear in terms of causal links to inputs, outputs, and other internal states. The functional account is not ordinarily taken to reduce psychological states to physical or physiological states. Among the causes and effects of a certain type of psychological state, a functional specification generally lists other types of psychological state. This kind of functional account is clearly of little help with the causal exclusion problem, since it continues to let in psychological states as causally efficacious entities.

By contrast, a functional *reduction* of the psychological to the physiological (so as to avoid the problem of causal exclusion) attempts ultimately to remove mental causes from the picture. Kim's account of functional reduction is based on that of David Lewis, who introduces this reductionist program by way of a suggestive analogy. Lewis imagines that we are assembled in the drawing room of a country house in the presence of a detective proposing a theory about who murdered Mr. Body. The detective utters the following sentences:

> X, Y and Z conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, X was Body's partner … Last week, Y and Z conferred in a bar in Reading … Tuesday night at 11:17, Y went to the attic and set a time bomb … Seventeen minutes later, X met Z in the billiard room and gave him the lead pipe … Just when the bomb went off in the attic, X fired three shots into the study through the French windows … (Lewis 1972/1999, 249–50; original ellipses)

Although they are written in the form of open sentences with three unbound variables (X, Y, Z), the detective uttering these sentences is ordinarily presupposing that X, Y, and Z are actually existing persons. One might rephrase his opening utterance as: 'There exist three persons X, Y, and Z, such that X, Y, and Z conspired to murder Mr. Body

…'. Thus, the conjunction of sentences that he utters comprises the detective's theory about the crime; more precisely, it can be rewritten as that theory's Ramsey sentence:

$$\exists X\, \exists Y\, \exists Z\, (\ldots X \ldots Y \ldots Z \ldots)$$

When the detective goes on to specify that X is none other than Plum, Y is Peacock, and Z is Mustard, then the reduction is complete. In the first step of the functional reduction, the Ramsey sentence supplies the *roles* of each of the key players (X, Y, and Z). In the second step, these variables are identified with the names of actual people, thereby providing us with the *occupants* of those roles. Lewis holds that the Ramsey sentence in the first step supplies an 'implicit functional definition' of the bound variables, X, Y, and Z (Lewis 1972/1999, 251–52). The second step occurs when the bound variables are subsequently identified with persons that occupy the roles designated in the theory's Ramsey sentence.

By adopting such a method, the mentalistic predicates used in psychology (say, M_1, M_2, …), should be capable, at least in principle, of being replaced entirely by functional constructs of neurophysiological predicates (say, N_1, N_2, …) that denote neurophysiological properties.[3] Causal exclusion worries are purportedly resolved since the mentalistic terms merely represent certain roles, the occupants of which are the causally efficacious properties that bring about particular effects. Being in a state of fear, for example, is the second-order 'property' of having a certain first-order neural property. Thus, when a psychological state of fear, say, is predicated of a particular person, we are merely picking out a certain role for which that person's neurophysiological properties are causally efficacious.[4] Ned Block uses the analogy of the property of dormitivity to illustrate the point: 'Dormitivity in one sense of the term is the property of having some first order property that causes sleep. The first order property is the realizer of the second order property of dormitivity' (Block 2003, 140). Thus, dormitivity becomes a kind of place-holder for whatever underlying physical constitution in different compounds is causally responsible for sending people to sleep.

Lewis observes that when we come to identify the mental states, M_1, M_2, … , with the neural states, N_1, N_2, … , this could hold '[i]n general, or for a given species, or in the case of a given person.' (Lewis 1972/1999, 257 n. 12). He adds: 'It might turn out that the causal roles definitive of mental states are occupied by different neural (or other) states in different organisms.' (Lewis 1972/1999, 257 n. 12). This is meant to allow for the possibility of multiple realization, for example that the role played by the psychological state of fear, say, could be occupied by different neural processes in different species or even in different individuals. This would be equivalent to saying that the detective's theory about the murder plot could be realized in different cases by different actors. In each of these cases, the roles assigned to X, Y, and Z would be the same, though their identities and the identity of their victim would be different. This would allow us to say that, for example, X represents the role of the principal plotter, Z the role of the assassin, and Y the role of the person who arranged for the cover-up, and that these roles were played by different individuals in two different plots. Similarly, the psychological predicates, M_1, M_2, … would pick out the same roles, but these roles

would be played by different neural properties, N_1, N_2, …, or P_1, P_2, …, and so on, in different species or individuals.

It is crucial to distinguish such cases of multiple realizability from those in which we discover that the roles implicitly defined by the detective's story do not neatly match those of the purported occupants. Suppose we find that the role represented by X was not played by Plum, for Plum only carried out some of the actions attributed to X in the story, while Peacock carried out the rest. We would naturally conclude that Plum was not the occupant of role X and that there could be no identification between role and occupant, as envisaged. We would reach a similar conclusion on finding that Peacock carried out some of the roles attributed to X as well as some attributed to Y. In the first case, the role is *subdivided* among two different occupants, while in the second case the roles are *crosscut* by the occupants. In both cases, there is a mismatch between roles and occupants. In such instances, we would be inclined to conclude that there is no such person who played the roles defined by the theory, that the theory was false, and that it should be modified accordingly.

Though the difference between cases of multiple realizability and cases of mismatch between roles and occupants may seem obvious when illustrated by Lewis's detective story, the corresponding contrast when it comes to a functional reduction of the mental may not be so obvious. In the following section, I will focus on a case study from cognitive science, arguing that it can be used as evidence of a mismatch between psychological and neural predicates. By taking a close look at some recent work in cognitive science, I will try to make the case that there is no prospect of an across-the-board functional reduction of the psychological to the neurophysiological.

3. Evidence from Cognitive Neuroscience

Let us consider the psychological state of fear in more detail, which would seem to be a prime candidate for functional reduction to a neurophysiological state. In this section, I will focus on an ongoing research program in cognitive neuroscience on fear in infant rhesus monkeys and in human children, undertaken by Ned Kalin and others (notably, Kristin Buss on human children). Kalin and his collaborators have carried out extensive experimental work on rhesus monkeys, particularly the conditions under which fear behavior is first acquired at ages of 9–12 weeks. They find that monkeys behave quite differently in three different experimental conditions, all of which involve threatening stimuli (Kalin and Shelton 1989, 1720; see also Kalin 2003, 42):

- Alone condition (A): monkeys are separated from their mothers or other conspecifics and placed in a cage alone.
- No-eye-contact condition (NEC): an unfamiliar human intruder comes into the monkey's presence, presents a profile, but does not make direct eye contact.
- Stare condition (ST): an unfamiliar human intruder comes into the monkey's presence and stares directly at the monkey, while maintaining a neutral expression

In A, the rhesus monkeys emit frequent 'coo' calls; in NEC, the monkeys tend to freeze; while in ST, the monkeys engage in 'aggressive gestures and vocalization', barking,

staring back, producing threat faces, and baring their teeth (Kalin and Shelton 1989, 1720). They sometimes mix the threatening displays with submissive behavior, such as 'fear grimaces, which look something like wary grins, or grinding of the teeth' (Kalin 2002, 74). They also coo in this third condition, though the experimenters speculate that the function of cooing in this condition is different from that in the alone condition. Whereas cooing in the alone condition may reflect an affiliative need, cooing in the stare condition may also be an urgent plea for help (Kalin 2002, 78). They state that behaviors exhibited in the stare condition 'are associated with fear in other species; thus, it is likely that these actions by infant rhesus monkeys are defensive and represent attempts by the infant to protect itself in a threatening situation' (Kalin and Shelton 1989, 1720). Clearly, these monkeys exhibit rather different behaviors in these different threatening situations; in other words, 'changes in the context of threatening situations resulted in dramatic changes in fear-related behaviors' (Buss et al. 2004, 585).[5]

At first glance, this may seem to be precisely a case in which we will be able to produce a functional reduction. Omitting some niceties, fear might be functionally represented as: (a) the state that mediates between an input consisting of separation from conspecifics and an output consisting of cooing, (b) the state that mediates between an input consisting of the presence of an intruder and an output consisting of freezing, and (c) the state that mediates between an input consisting of the stare of an intruder and an output consisting (*inter alia*) of barking, baring teeth, and cooing. Schematically, we can represent the situation as follows (with arrows representing causation):

$$\text{A} \rightarrow \text{FEAR} \rightarrow \text{COOING}$$
$$\text{NEC} \rightarrow \text{FEAR} \rightarrow \text{FREEZING}$$
$$\text{ST} \rightarrow \text{FEAR} \rightarrow \text{BARKING (ETC.)} \,\&\, \text{COOING}\cdot$$

Furthermore, this formulation would seem to provide us with the makings of a functional *reduction*: any neurophysiological state that plays the functional role indicated in these causal processes can be considered to be a state of fear. To revert to Lewis's schema above, we can think of the psychological theory pertaining to these rhesus monkeys as a long conjunctive sentence containing three conjuncts, each of which involves a place-holder standing in for the role of the psychological state of fear (X):

$$\ldots\ (I_1 \rightarrow X \rightarrow O_1)\ \&\ \ldots\ \&\ (I_2 \rightarrow X \rightarrow O_2)\ \&\ \ldots\ \&\ (I_3 \rightarrow X \rightarrow O_3)\ \&\ \ldots,$$

where the Is stand for certain environmental inputs (here, experimental conditions) and the Os for corresponding outputs (behavior of various kinds). When we preface this theory with the existential quantifier that quantifies over X, we obtain the theory's Ramsey sentence. In this theory, 'X' marks a particular psychological role (fear), which can eventually be identified with a certain neurophysiological state.

The problem with this purported reduction is that the second step (identification with a neurophysiological state) does not seem to be borne out by the empirical facts assembled by Kalin and his collaborators. Even though the investigators regard all three

of the experimental conditions as threatening and consider that the responses given by the monkeys show signs of fear or involve 'fear behaviours', they do not find any neurophysiological commonalities exhibited in the three conditions. Indeed, these investigators conclude that the monkeys' behaviors in the three conditions are not caused by the same neural processes:

> Interestingly, these different behaviors appear to be controlled by different neurotransmitter systems. Thus, manipulations of the opiate system affected coo frequency without affecting barking induced by ST or freezing induced by NEC. If the effects of altering the opiate system were mediated simply by changes in the infant monkey's level of arousal, then barking and freezing would decrease with morphine and increase with naloxone. This was not the case. Conversely, diazepam reduced barking and freezing without significantly affecting cooing. (Kalin and Shelton 1989, 1720)

They also state that 'Opiate and benzodiazepine systems may also mediate the development of human psychopathology characterized by excessive or inappropriate fear responses' (Kalin and Shelton 1989, 1720). The internal physiological processes that take place in each of these situations appear to be quite different. That is to say, X is not underwritten by the same neurophysiological state in each case; it cannot therefore be reduced to such a state.

There does not seem to be a single common physiological element that is always and only associated with cases of fear behavior, which is overlaid with additional factors or processes in each of the three different situations. In the literature on the neurophysiology of fear in (human and non-human) primates, one prime candidate that has been suggested as a neural correlate for fear is cortisol, a corticosteroid hormone secreted into the bloodstream by the adrenal cortex. Some studies have indicated that, 'Emotional stressors, such as novelty and uncertainty, involving fearful responses result in cortisol increases' (Buss et al. 2004, 584). But it is also the case that 'several studies have failed to find an association between fear behaviors and stress cortisol levels' (Buss et al. 2004, 584). Moreover, 'Elevations in cortisol are not unique to fear-related behaviours; they have also been observed for bold and exuberant behaviors' (Buss et al. 2004, 585). Thus, elevated cortisol levels are not invariably associated with fear; nor is fear always associated with elevated cortisol levels.

Instead of a one-to-one relationship, there is a mismatch between fear and certain neurophysiological states and processes, which could be a case either in which fear is *subdivided* by neurophysiological predicates (i.e. one-to-many-relation) or in which it is *crosscut* by such predicates (many-to-many relation). Now, if it transpires that there is a one-to-many relationship between fear and certain neurophysiological states, it may be tempting to consider this a case of multiple realizability. One might say that fear is multiply realized in rhesus monkeys, the same psychological role being played by different neurophysiological processes in different types of situation. However, a quick reflection on Lewis's detective story will show that it is importantly different from cases of multiple realizability. As Lewis points out, multiple realizability of psychological properties in neural (or other physical) properties comes about when the roles played by mental predicates M_1, M_2, … in the functional specification of their roles are

occupied by different P_1, P_2, … in different organisms or species. However, this is a case in which the role played by M_1 is not wholly occupied by P_1, but sometimes by P_2 and P_3. (It may also be a case in which P_1 sometimes occupies the role played by some other mental state, M_2, which would be a case of crosscutting rather than subdividing.) Rather than finding that a certain psychological predicate can be identified with a single neurophysiological property (in a particular species or organism), we have found that there are different neurophysiological properties in different situations occupying the role that is identified in the psychological theory. At the very least, it appears that M_1 itself does not correspond to a single P, and it may be that some of the Ps to which it corresponds may also be involved in manifestations of some of the other Ms.

Multiple realizability is sometimes divided into type and token varieties. The basic notion is that mental states can be multiply realized relative to physiological states in one of two ways. There can be different realizations of a mental state in, say, different species or structures (whether actual or possible). Alternatively, there can be different realizations of a mental state in different individual organisms. In both cases, it is assumed that something like a local reduction is possible, if only one indexed to a specific individual or set of individuals. Despite the difference between phenomena that are multiply realizable in types (e.g. all rhesus monkeys) and those that are multiply realizable in tokens (e.g. in a particular rhesus monkey with a particular history of neural development and neural organization), they continue to consist in phenomena that can be *generally* characterized at the physiological level. Though the generalizations are more or less restricted, there is something that such states have in common from the vantage of neurophysiology. However, there is another type of case in which the physiological subdivides the psychological. At first sight, it may seem feasible to consider such cases to be ones of multiple realizability of a yet more restricted sort, namely multiple realizability relativized to a type of situation. But that would be to misunderstand the nature of the case at hand. It is true that the evidence cited above does not suggest that fear is associated with a different neurophysiological correlate in each token situation, but rather that it is so associated in each *type of situation* (corresponding to the three experimental conditions: A, NEC, and ST). However, once we conclude that manifestations of fear in an individual or a species in different types of situation are underwritten by different neurophysiological processes, we are no longer dealing with a single functional state that is multiply realized. Rather, we are saying that nothing plays the unitary functional role outlined by our psychological theory of fear, and therefore that there is no such thing as fear from the vantage of neurophysiology. As shown clearly by Lewis's analogy with the detective story, if no single individual played the single part ascribed to X, we would have to conclude that there was no such role to be played, contrary to the detective's theory.

4. Reductionism or Psychological Autonomy?

Functionalist reductionists can avail themselves of an obvious line of response in the face of a one-to-many relationship between psychological and physiological predicates. They could simply propose that we were wrong about fear being a single property or

state, and insist that there are, say, three different properties or states, $fear_1$, $fear_2$, and $fear_3$, corresponding to the different neural processes that underwrite each. Thus, they could argue that what folk psychology has identified as a single state is in fact three different psychological states, reducible to three different neurophysiological substrates. In this section, I will try to argue that this is not a palatable alternative.

Let us begin by examining more closely the proposal that fear has been found to be three states or properties, not one. Consider how Kalin and his collaborators identify the state as one of fear in the first place. By modifying the functional characterization of fear given above, we can postulate that three different psychological states are responsible for mediating the inputs and outputs involved:

$$\begin{aligned} &A \rightarrow FEAR_1 \rightarrow COOING \\ &NEC \rightarrow FEAR_2 \rightarrow FREEZING \\ &ST \rightarrow FEAR_3 \rightarrow BARKING\ (ETC.)\ \&\ COOING, \end{aligned}$$

where '$FEAR_1$', '$FEAR_2$', and '$FEAR_3$' are accidental homonyms. But the scientists undertaking this research resist such an analysis on the grounds that the experimental conditions in all three cases (ALONE, NO-EYE-CONTACT, and STARE) are 'situations perceived as threatening' by the monkeys (Kalin, Shelton and Takahashi 1991, 1176). Indeed, the category of 'fearful or threatening situations' was employed in order to design the experimental setup that was supposed to reveal something about the internal states of the monkeys. Therefore, there is clearly an assumption on the part of these scientists that fear is a real psychological state or property and that it is important to uncover its neural substrates. However, it might be objected that that is not necessarily an indication of the reality of fear, since this is precisely a case in which science may have discovered that what we once thought to be a unitary state or property is not in fact so. One might cite historical precedents to show that even though scientists might begin with a working assumption that a certain category is real, they may go on to discover that it must be discarded in light of the evidence. In this case, *fear* could be a category that scientists employ in their initial inquiries but end up jettisoning as their inquiries progress and as they learn more about the underlying neurophysiology of the situation.

Although there are undoubtedly instances from the history of science in which we revise our old taxonomies and replace existing categories with newer ones, that is not always the case. Existing taxonomies are often particularly resilient when they are part of our lay or folk theories. These categories are not always revised when a different set of categories is uncovered by scientific investigation, particularly when they play a role in certain anthropocentric activities or inquiries. Examples abound from the history of science: *lizard* is not a category recognized by biological systematics, nor is *bird* (unless it is widened to include dinosaurs), nor are such categories as *onion*, *garlic*, *tree*, *weed*, *parasite*, or *livestock*. Similar examples can be cited from other domains: *glass*, *vitamin*, *poison*, and *energy* (if this is taken to exclude mass), to mention a few.[6] In all such cases, folk categories or categories pertaining to older sciences have been retained for many purposes—including scientific purposes, especially in those sciences directly relevant

to human concerns and interests—despite the fact that they do not coincide neatly with the categories picked out by newer sciences, or sciences of micro-phenomena. I would argue that that is precisely the kind of situation we face in this case, since the psychological category of *fear* is one that plays an important role in our folk theories, as well as in scientific theories particularly pertinent to human concerns and interests. Replacing the category *fear* with three new categories would entail foregoing certain explanations that link threatening situations to defensive behaviors. In the experimental situations described in the previous section, there would no longer be any similarity between the three 'threatening' conditions, A, NEC, and ST, since each would have to be characterized in terms of the proximal stimuli on the monkeys (retinal impressions, auditory cues, and so on) or in terms of distal stimuli unintentionally characterized, which would not necessarily bring out any significant similarities among them. Similarly, there would no longer be certain natural affinities between the 'fear-induced behaviours' in the three conditions. Thus, we could no longer frame explanations that would link the three types of stimuli and the three types of subsequent behaviors by way of the unitary psychological category of *fear*. The reductionist position would require us to dispense with *fear* and replace it with three new categories, denoting entirely different internal states. This is importantly different from introducing three *additional* categories, $fear_1$, $fear_2$, $fear_3$, which represent a more fine-grained taxonomy than our earlier one, since that would be to implicitly retain the overarching category *fear*. Rather, abandoning the existing category would be tantamount to a denial that there is some commonality among the three types of psychological state (which is why I referred to them as 'accidental homonyms' in the previous paragraph).

This reductionist position is not only undermined by many precedents in the history of science but also resisted by the scientists involved. Some of the researchers I have cited do express some doubts as to whether the internal states involved are the same in different experimental setups (though in this instance they are commenting on similar experimental results concerning 24-month-old human children rather than infant rhesus monkeys):

> Problems arise when these rather distinct behaviors are discussed as if they were part of the same construct … Failing to distinguish components of the fear family of behaviors can thus lead to failure to discern physiology–behavior links. (Buss et al. 2004, 591)

Nevertheless, it is quite clear that even though they may advocate making finer-grained distinctions among different types of fear, they are not calling for abandoning the category of *fear* altogether: 'As the current study demonstrates, each type of fear reaction (e.g., inhibition, the fear behavior composite, and freezing) may have different associations with physiology' (Buss et al. 2004, 591). In other words, despite the fact that they clearly recognize the lack of a one-to-one correspondence between psychological and physiological categories, they do not seem to be making a plea for replacing the category of *fear* but at most distinguishing between different types of fear that arise in different situations. Moreover, their insistence stems partly from the fact that they see certain important psychological similarities between individuals who behave in a non-standard fashion in different experimental conditions. In conducting research on

human children, these investigators are attempting to identify individuals who have a 'fearful temperament'. They find that such individuals exhibit a contextually inappropriate (or 'dysregulated') response consisting of 'a high level of freezing duration across all the stranger contexts' (Buss et al. 2004, 591). They hypothesize that this inappropriate response exhibited by human children is analogous to that exhibited by a small percentage of monkeys, who exhibit freezing behavior in the stare condition (as opposed to the majority, who only engage in this behavior in the no-eye-contact condition). Thus, one thing these scientists are interested in investigating are the commonalities among those humans who have a 'fearful temperament', an investigation that can only be pursued if one employs the category *fear*. Interestingly, they do cite evidence to suggest that there are some neurophysiological features common to human children who have such a temperament, but that is not the same as finding that there is a neurophysiological correlate of the state of fear itself: 'Children with an extremely fearful or shy temperament have greater relative right frontal EEG activity at baseline … and during stressful tasks …' (Buss et al. 2003, 11).

At this point, it may be objected that distinguishing among three different types of psychological state may not be such a bad thing in the particular case under discussion. Rather than impoverish our psychological theorizing and explanations, it may in fact enrich them. For example, further reflection on the three experimental conditions described in the previous section might lead one to propose that the alone condition (A) ensues in a psychological state of *distress*, while the no-eye-contact condition (NEC) results in *anxiety*, and the stare condition (ST) leads to *panic* (or perhaps this is the only condition that leads to *fear* proper). These categories should be thought of as new terms of psychological art, though they gain some plausibility from the fact that our corresponding folk psychological terms seem intuitively to apply to the experimental conditions as described in the previous section. Perhaps a revision of our initial assumption that all three conditions are ones involving fear is precisely what is needed here. However, before endorsing this proposal, the objector should keep in mind that this alleged enrichment of our explanatory resources still deprives us of the ability to say that the three situations have a single property in common, which, as I have argued, is suggested by our folk theory as well as by the psychologists and cognitive scientists who are undertaking these investigations. As we saw above, the researchers studying the neurobiology of fear link threatening situations to defensive behaviors in terms of the internal state of fear, and they attempt to come up with generalizations about individuals with a 'fearful temperament'. Formulating such explanations and generalizations requires them to use a common psychological term to characterize internal states of the monkeys in the situations under investigation. In all three experimental conditions that they examine, the explanations would be unavailable to them without appealing to the unitary category of *fear*. Note that this leaves open the possibility of distinguishing among three different types of fear, which may indeed be the conclusion to which this evidence points. Thus, in order to achieve this enrichment of our theoretical vocabulary, we need not dispense with the category of *fear* altogether.

This may not amount to a general argument against dispensing with our existing psychological vocabulary and replacing it with new vocabulary that correlates better

with the underlying neurophysiological facts, but it does suggest that replacement of existing categories is not always the outcome when the categories of one science conflict with another, more recent science, or when the categories of science conflict with those of the folk. In many such cases, particularly involving anthropocentric inquiries, there is an important loss of information and explanatory power involved when existing explanatory categories are discarded. Besides, some of the evidence cited in the previous section indicates not only that physiological categories are more fine-grained and that there is a one-to-many relationship between psychological and physiological states, but that there may be a many-to-many relationship between them (i.e. crosscutting among them). As indicated above, there may be physiological commonalities among some states of fear and other types of mental state, such as those leading to 'bold and exuberant behaviours'. If these or similar findings are upheld by future research, this would preclude the possibility of replacing our current psychological categories with ones that are more fine-grained, but would rather involve replacing these categories with an entirely different taxonomy that crosscuts the existing one.

I have argued that revision of existing categories in the face of subdivision or crosscutting relative to the purported reduction base is not always the course of action that is pursued in the history of science, and that this is particularly so in the case of categories that play an important role in human activities and inquiries.[7] An altogether different course would be to abandon reductionism and give up on finding neural or physiological occupants for the roles specified in the psychological theory, concluding instead that the functional theory has validity in its own, separate domain. This would amount to retaining functionalism without reductionism. Psychological states could then be analyzed in terms of inputs, outputs, and other psychological states, without necessarily needing to identify the roles specified in the theory with certain neurophysiological properties in a one-to-one fashion. This assertion of the 'autonomy' of the psychological is by no means a blanket defense of all existing categories and theories in folk or scientific psychology. It is a claim to the effect that the absence of a one-to-one correspondence between psychological categories and neural ones need not entail revising or eliminating the psychological categories—although in some cases, it may. Reductionism requires that *all* such psychological predicates be reducible, whereas the autonomy of psychology is vindicated even if there is only one psychological state or property that is worth retaining that is not reducible to the physiological.

5. Conclusion

Functional reductionism would seem to be an attractive compromise between reductionism and functionalism. Though it calls for the reduction of psychological properties to those of neuroscience, it is compatible with the multiple realizability of such properties, since the kind of reductions it countenances are local in nature. In addition, it appears to provide a solution to the causal exclusion problem propounded by Kim: if psychological properties are reducible to neurophysiological ones, then their causal powers are identical, and they have no independent causal powers of their own. But despite the apparent attractions of functional reductionism, I have presented evidence

from cognitive neuroscience showing that distinctions at the neurophysiological level subdivide or crosscut those made at the psychological level, thus rendering reduction impossible. Though this may appear at first to be simply a case of multiple realizability, closer inspection reveals that it is importantly different, since it precludes the possibility of even a local reduction of the psychological to the neurophysiological. In response, reductionists might advocate replacing existing psychological categories with ones that conform to those of neurophysiology, but I have argued that such a course of action is not an attractive option in this particular case, given the explanatory value of such psychological categories in our folk theories as well as in current theories in psychology and cognitive science; nor is it necessarily recommended by reflection on the history of science.

One implication of this argument is that Kim's solution to the problem of causal exclusion is unworkable. Is there another way of resolving the problem? Though much philosophical ink has been spilled over this topic, there appears to be no consensus even concerning whether it is a genuine problem or a pseudoproblem. Moreover, there is no agreement among philosophers as to whether the problem generalizes to other special sciences or not, though some recent responses to Kim argue that it does (see e.g. Bontly 2002; Block 2003). If that is indeed the case, then exclusionary worries should be endemic concerning causation in the 'special sciences'. But this should give us pause and lead us to reflect more thoroughly on our conception of causation. Kim avers that the causal exclusion problem arises from 'what strikes [him] as a perfectly intuitive and ordinary understanding of the causal relation' (Kim 1998, 67). However, it may be that some of the intuitive features of our hallowed concept of causation, influenced by two and a half millennia of philosophical theorizing and derived from familiar macro-phenomena and common-or-garden variety objects at the same level, need to be re-examined. Terry Horgan advocates 'causal compatibilism', the position that 'there is genuine causation and genuine causal explanation at multiple descriptive/ontological levels, and that despite the causal closure of physics, physics-level causal and causal-explanatory claims are not really incompatible with mentalistic causal and causal-explanatory claims' (Horgan 2001, 98). To this end, Horgan suggests that the 'concepts of causation and causal explanation are contextually parameterized notions, with an implicit contextual parameter keyed to a specific descriptive/ontological level …' (Horgan 2001, 102). This seems like a step in the right direction, in contextualizing causation and not regarding causal accounts at different levels as being in competition. Since explanation is widely recognized to be interest-relative, why not causation? Many philosophers are likely to balk at such a suggestion, regarding causation as a metaphysical rather than an epistemic notion, which is not subject to contextualization or parametrization in this manner. What qualifies as a satisfactory explanation, they will say, may depend on our interests, needs, and predilections, but the causal relation itself cannot. On this traditional philosophical view, the 'cement of the universe' either connects events or it does not, no matter what our explanatory practices lead us to believe. But a truly naturalized ontology should treat causation like any other scientific or meta-scientific concept, which must earn its keep regardless of our traditional philosophical notions about its metaphysical status.[8] Many causal processes in the

special sciences suggest that causal compatibilism is truer to scientific practice than a blunt insistence that causation always operates exclusively at a single level.

Acknowledgements

I am grateful to two anonymous referees for this journal for very constructive criticism that led to major changes to this paper. I would like to thank the University Research Board of the American University of Beirut for two summer grants that enabled me to write this paper.

Notes

[1] Of course, properties do not cause other properties: it is property instances as manifested in events that are causally efficacious. In what follows, I will be speaking loosely.

[2] Those properties may themselves be predicates reducible to lower-order properties; I will ignore this complication in the rest of the paper. Also, in what follows, I will sometimes refer loosely to these predicates as 'properties', as Kim himself does on occasion.

[3] Lewis's functional reduction of psychological predicates builds on his earlier proposal for defining theoretical terms, as outlined in Lewis (1970/1983). He draws on Ramsey's method for expressing scientific theories, but in a twist on the theoretical–observational distinction, he postulates that the theory being interpreted contains two sets of terms, T-terms and O-terms, characterized as follows. A T-term is 'a theoretical term introduced by a given theory T at a given stage in the history of science', and an O-term is, by elimination, 'any *other* term, one of our *original* terms, an *old* term we already understood before the new theory T with its new T-terms was proposed' (Lewis 1970/1983, 79). Accordingly, in the above presentation, the T-terms are none other than X, Y, and Z. As applied to a psychological theory, the identification of O-terms with 'old' terms and T-terms with newly introduced terms is not entirely apt. The terms distinctive to psychological theorizing are often familiar mentalistic terms that have been in use for some time ('belief', 'desire', 'fear', 'pain', etc.), while the O-terms can be thought of as terms denoting various environmental stimuli and behaviors.

[4] In what follows, I will sometimes talk about the category or concept of *fear* (in italics), but I will also speak of a state or property of fear (no italics). In adopting the latter terminology, I do not mean to be prejudging the case for or against reduction; any mention of a property or state of fear can be paraphrased in terms of a category or concept of *fear* (which is denoted by the predicate 'fear').

[5] It should be noted that these behaviors, while typical and manifested by numerous individuals, were not universally exhibited in the three conditions; e.g. some monkeys froze in ST (see Buss et al. 2004, 585).

[6] See Dupré (1993) and LaPorte (2004), from which some of these examples are taken (though LaPorte would not agree with the position I am defending here). See also Khalidi (1993, 1998).

[7] This argument against functional reductionism is therefore what Ruphy (2005) considers a 'temporally qualified argument' against reductionism, since it depends on our explanatory and classificatory practices.

[8] Horgan argues that philosophically important concepts like *causation* need to accord not just with our intuitive judgments, but also with untendentious scientific knowledge, sociolinguistic purposes, and other types of data, all of which 'go into the hopper of wide reflective equilibrium' when determining the proper analysis of such concepts (Horgan 2001, 109).

References

Block, N. 2003. Do causal powers drain away? *Philosophy and Phenomenological Research* 67: 133–150.

Bontly, T. D. 2002. The supervenience argument generalizes. *Philosophical Studies* 109: 75–96.

Buss, K. A., J. R. Malmstadt Schumacher, I. Dolski, N. H. Kalin, H. H. Goldsmith, and R. J. Davidson. 2003. Right frontal brain activity, cortisol, and withdrawal behaviour in 6-month-old infants. *Behavioral Neuroscience* 117: 11–20.

Buss, K. A., R. J. Davidson, N. H. Kalin, and H. H. Goldsmith. 2004. Context-specific freezing and associated physiological reactivity as a dysregulated fear response. *Developmental Psychology* 40: 583–594.

Dupré, J. 1993. *The disorder of things.* Cambridge, MA: Harvard University Press.

Horgan, T. 2001. Causal compatibilism and the exclusion problem. *Theoria* 16: 95–116.

Kalin, N. H. 2002. The neurobiology of fear. *Scientific American,* special edition, 12: 72–81.

Kalin, N. H. 2003. Nonhuman primate studies of fear, anxiety, and temperament, and the role of benzodiazepine receptors and GABA systems. *Journal of Clinical Psychiatry* 63 (Supplement 3): 41–44.

Kalin, N. H. and S. E. Shelton. 1989. Defensive behaviors in infant rhesus monkeys: environmental cues and neurochemical regulation. *Science* 243: 1718–1721.

Kalin, N. H., S. E. Shelton, and L. K. Takahashi. 1991. Defensive behaviors in infant rhesus monkeys: ontogeny and context-dependent selective expression. *Child Development* 62: 1175–1183.

Khalidi, M. A. 1993. Carving nature at the joints. *Philosophy of Science* 60: 100–113.

Khalidi, M. A. 1998. Natural kinds and crosscutting categories. *Journal of Philosophy* 95: 33–50.

Kim, J. 1998. *The mind in a physical world.* Cambridge, MA: MIT Press.

Kim, J. 2003. Blocking causal drainage and other maintenance chores with mental causation. *Philosophy and Phenomenological Research* 67: 151–176.

LaPorte, J. 2004. *Natural kinds and conceptual change.* Cambridge: Cambridge University Press.

Lewis, D. 1970/1983. How to define theoretical terms. In *Philosophical papers,* Vol. 1, edited by D. Lewis. Oxford: Oxford University Press.

Lewis, D. 1972/1999. Psychophysical and theoretical identifications. In *Papers in metaphysics and epistemology,* edited by D. Lewis. Cambridge: Cambridge University Press.

Ruphy, S. 2005. Why metaphysical abstinence should prevail in the debate on reductionism. *International Studies in the Philosophy of Science* 19: 105–121.