The status of functional explanation in psychology: Reduction and mechanistic explanation

Raoul Gervais
Ghent University

Huib Looren de Jong
VU University Amsterdam

Abstract
The validity of functional explanations as they are commonly used in psychology has recently come under attack. Kim’s supervenience argument purports to prove that higher-level generalizations have no causal powers of their own, and hence are explanatorily irrelevant. In a nutshell, the supervenience argument forces us to either embrace epiphenomenalism of higher-level properties, or accept Kim’s specific brand of reductionism. However, with the current emphasis on mechanistic explanations, the literature on explanation in psychology has undergone some drastic changes. It could be argued, therefore, that Kim’s argument targets an outdated concept of functional explanations. In any case, these developments warrant a reassessment of the implications of his argument, which is the purpose of the present paper. First, we argue that the metaphysics behind the supervenience argument is incompatible with that of mechanisms. Second, we argue that Kim’s proposed brand of reductionism does not accurately describe the explanatory practices of cognitive science.

Keywords
functional explanation, mechanism, reductionism, supervenience, theory of mind

Conventional wisdom holds that the core business of the cognitive sciences is to provide intentional explanations: that is, explanations that refer to mental states (beliefs, desires, goals, and hopes) as the causes of behaviour (Davidson, 1980). These folk-psychological
explanations found their philosophical justification in functionalism, which interprets mental states as neurally implemented functions. This sense of “function” is not to be confused with the aetiological sense employed in evolutionary biology to explain the occurrence of the function bearers (Mayr, 1988). In this article, we use the term to indicate the causal role a given entity plays: for example, the systemic activity exhibited by physiological mechanisms (for a discussion of the various differences between aetiological and causal role functions, see Sober, 1990). Special emphasis was placed on those functions that are “the same” across many different implementations, the idea being that at a certain level of abstraction, two systems that are physically different can nevertheless be functionally equivalent (the so-called “multiple realizability” of functions). In cognitive psychology and philosophy of mind, functionalism is a thesis about the nature of mental processes and their realization in physical processes. Although the general tenets of functionalism are applicable to other fields of study as well (e.g., clinical and developmental psychology), this paper will be concerned with functionalism in its relation to cognitive science, as it is this domain that has seen its most articulated developments. Here, functionalism sought to combine the metaphysical position of materialism or physicalism with the autonomy of the special sciences (Fodor, 1974). The combination of these two claims amounts to the position of non-reductive physicalism, which is probably the dominant position in philosophy of mind today. The task of the cognitive scientists, then, was to explicate the intentional laws holding between mental states and behaviour. This explanatory strategy could be labelled functional, because it decomposes the system responsible for the behaviour that is to be explained in terms of the action of its parts or components. In traditional psychology, this type of functional analysis is used to great effect, explaining cognitive capacities underlying observable behaviour by showing how they work. It begins by specifying the capacity to be explained, and proceeds by decomposing this capacity into sub-tasks or -routines. Together these “make up” the original capacity or causal role that constitutes the explanandum (Cummins, 1980; Dennett, 1987).

However, this reliance on folk-psychological idiom has met with severe criticism lately, and the hegemony of non-reductive physicalism is being challenged. A highly influential proponent of reductionism today is Jaegwon Kim, who believes that the higher-level, functional generalizations (those generalizations involving precisely the kind of functions that are thought to be multiply realizable) are explanatorily irrelevant because the causal powers of the explananda and explanantia they refer to are pre-empted by the causal powers of their respective lower-level realizers (Kim, 1998, 2005). To give Kim’s example (Kim, 1998, pp. 21–22), the causal role of the functional property “dormitiveness” of a sleeping pill is nothing over and above the causal workings of the chemical properties that fill that causal role: it is those chemical properties that make one sleepy, not the property of dormitiveness. In a physically closed universe there seems to be no room left for intentional states or properties to be causally efficacious. The only way out, according to Kim, is to reject non-reductive physicalism and embrace reductionism instead. This dilemma is known as the supervenience argument; we will discuss it in the second section of this manuscript.

This metaphysical challenge to the causal status of intentional and functional properties has serious methodological implications. If reductionism is correct, then the traditional
functional analysis, with its decomposition of complex systems into operations of parts, is obsolete. As soon as we have identified the components that do the real causal work, any reference to intentional states seems nothing more than a red herring. To fill this gap, Kim has devised a special brand of reductionist explanation: functional reductionism (we will discuss this in the third section of this manuscript). Briefly, the idea is that in explaining why a system exhibits some or other higher-level property, we begin by providing a functional definition of this property (e.g., having a certain temperature can be defined in terms of its ability to cause wax to melt). Next, one identifies the lower-level realizers that fill this functional role (in the case at hand, it is molecular kinetic energy that causes wax to melt). In this way, a true explanation reveals the lower-level cause of the explanandum.

However, it could be argued that the supervenience argument targets a research strategy that has not remained unchanged over the years: new developments have been made in the area of functional explanations. Following Machamer, Darden, and Craver’s seminal paper, “Thinking about Mechanisms” (2000), the past decade has seen the rise of a substantial body of literature devoted to so-called mechanistic explanations in the life sciences. According to these authors, what actually happens in the cognitive sciences is that researchers explain a system’s behaviour by constructing a model of the underlying mechanism responsible for it. Although mechanistic explanations do involve decomposition of capacities into sub-routines, they differ from traditional functional explanation by requiring that one moves beyond mere abstract analysis of function and identifies the actual components that realize the (sub-)functions. Indeed, it is even argued that for a model to be genuinely explanatory, it does not suffice to give an abstract story about how a capacity could be realized; one also has to specify, at least tentatively, some of the parts and activities actually present in the mechanism itself (Craver, 2006). Given these developments, we think it is time to reassess Kim’s supervenience argument. Does it threaten mechanistic explanation to the same extent as it does traditional functional explanations (indeed, does it apply at all)? Do mechanistic explanations present those worried about the autonomous status of psychology with additional arguments to meet Kim’s challenge?

We will argue that the supervenience argument does not affect the legitimacy of mechanistic explanations. Specifically, we will make two claims regarding Kim’s position: (a) it gets the metaphysics wrong and (b) it gets the science wrong. Regarding the first claim, we will argue that the supervenience relation, insofar as it applies to mechanisms, involves a part–whole relationship (it is what metaphysicians call “mereological”). For this reason, Kim’s argument cannot get off the ground, since he restricts the supervenience relation to hold only between properties of the same entity, which is clearly not the case with part–whole relationships. Giving up this restriction, however, would invite additional metaphysical problems, namely that all higher-level causation is pre-empted. In short, the problem is that either Kim’s argument proves too much, or he must restrict the supervenience relation in a way that makes it incompatible with the relations holding between mechanisms and their parts. Opting for the second horn of the dilemma, however, one is forced to accept the causal efficacy of the functional states described by mechanistic explanations.

Regarding the second claim, we will argue that Kim’s functional reductionism does not mirror what is actually going on in mechanistic explanations. Kim conceives his reductive explanations as arguments, but mechanistic explanations consist of constructing models,
not arguments. Although such explanations do involve a reductive component, the mechanism itself produces behaviour that none of its components can produce. The components of a mechanism may be located on a number of different levels, so the resulting mechanistic model is a multi-level explanation—which Kim’s reductionism seems unable to account for.

Let us conclude this introduction with a brief overview of the paper. In the first section, we briefly introduce traditional functional explanation. In the second, we will give an account of Kim’s supervenience argument, indicate the reason why it poses problems for the traditional account, after which we present Kim’s own brand of reductionism as an alternative in the third. The fourth section then introduces mechanistic explanations and contrasts them with their functional predecessors/competitors. In the fifth and sixth sections the case is made for claim (a) and (b) respectively, where the latter is illustrated with the example of a mechanistic explanation of numerical cognition.

Traditional functional explanation and the autonomy of the special sciences

First, let us give some brief comments on what is traditionally meant by functional explanation. As a first approximation, a functional explanation can be said to focus on what something does, rather than what does it. It involves explaining a function or capacity by decomposing it into different sub-functions or -capacities. For example, a combustion engine has the capacity to convert energy into motion. This general capacity can be broken down into sub-capacities. In the example of a combustion engine, we can decompose the general function (to convert energy into motion) into four sub-routines (four-tact cycle): injection, ignition, combustion, and exhaust. Now the point is that from a purely functional point of view, how each of the individual sub-routines is carried out is irrelevant. In our example, the second part of the process, ignition, can be carried out externally or internally. At a certain level of abstraction (to the degree that they both exhibit the same, four-tact cycle), two combustion engines can be said to be functionally equivalent, even though the ignition part of the process is executed internally in the one and externally in the other.

This explanatory strategy was once widely popular in the cognitive sciences, where the behaviour of a system was described in terms of the functional role its components play (Bechtel & Richardson, 1993; Cummins, 1980, 1983). For example, in psycholinguistics, the activity of speech production was explained by decomposing that activity into sub-routines such as conceptualization, formulation, and articulation, while the sub-routine formulation was itself again divided into the even smaller (i.e., more local) sub-routines of lexicalization and syntactic planning (Levelt, 1989). The goal then was to show how the overall capacity or activity of speech production arises as a result of the sub-activities and their organization. Typically, a simple box diagram was used to this end, like the one shown in Figure 1.

All kinds of cognitive capacities were explained in this way, from spatial memory to face recognition. In short, the abstract, functional accounts of cognitive capacities were traditionally held to be at the very core of cognitive psychology (Cummins, 1980; Lycan, 1987). In the philosophy of mind, this explanatory strategy was married to functionalism
to argue against (classic) reductionism and for the autonomy of the special sciences (Fodor, 1974). In this way, philosophers sought to combine physicalism (roughly the view that all the physical “stuff” is all there is; we will give a more precise characterization later on) with the legitimacy of psychology, as the functional states to which psychological explanations refer (e.g., hunger, beliefs, desires) were construed as having abstract causal roles that do not map neatly onto physical states (Fodor, 1990). This is an important element in the discussion on the status of folk psychology, which explains behaviour in terms of beliefs and desires (Fodor, 1990). If functionalism is right, folk psychology has its own domain and cannot be reduced to neuroscience.

This combination of the functional research strategy and the metaphysical claim of physicalism, then, was, and indeed still is, the dominant position among theorists working in the cognitive sciences and the philosophy of mind. The vigour with which this conglomerate of metaphysical and methodological opinions has been argued for over the past decades reflects how deep the desire runs within the community of non-reductive physicalists to keep separate the explanatory research programmes of the life sciences (including psychology) and the so-called “hard sciences” (cf. Fodor, 1997).

Recently, however, this consensus is being challenged. At the heart of this new critique is Kim’s supervenience or exclusion argument, which purports to prove that non-reductive physicalism is an unstable position. More precisely, it aims to force a choice between epiphenomenalism, where functional properties are said to be without causal powers of their own and therefore lack explanatory power, on the one hand, and reductionism, which effectively means giving up the cherished autonomy of psychology as an independent discipline, on the other. Both options are disastrous for the status of functional explanation in psychology. In the next section, we will take a detailed look at Kim’s argument and its implications.

**The supervenience argument**

Over the past two decades, Kim has offered a number of different versions of the supervenience argument, which have sparked extensive debates (Ross & Spurrett, 2004; Walter, 2008). Here we will concentrate on his 2005 formulation. The argument revolves around a number of principles, each of them indispensable to proponents of non-reductive

![Figure 1. A functional analysis of speech production.](image)
physicalism, but which, when put together, conspire to make trouble. These are as follows (see Kim, 2005, pp. 33–43):

1. **The causal closure of the physical domain.** If a physical event has a cause at \( t \), then it has a physical cause at \( t \).
2. **Principle of causal exclusion.** If an event \( e \) has a sufficient cause \( c \) at \( t \), no event at \( t \) distinct from \( c \) can be a cause of \( e \), unless it constitutes a genuine case of causal overdetermination, where two independent causes appear simultaneously to produce a single effect (e.g., two assassins simultaneously firing their bullets into a person’s heart).
3. **Supervenience.** Mental properties strongly supervene on physical/biological properties. That is, if any system \( s \) instantiates a mental property \( M \) at \( t \), there necessarily exists a physical property \( P \) such that \( s \) instantiates \( P \) at \( t \), and necessarily anything instantiating \( P \) at any time instantiates \( M \) at that time.
4. **Irreducibility.** Mental properties are not reducible to, and are not identical with, physical properties.
5. **Causal efficacy.** Mental properties have causal efficacy—that is, their instantiations can, and do, cause other properties, both mental and physical, to be instantiated.

The goal of Kim’s supervenience argument is twofold: it purports to show that (a) these principles are not compatible with each other, and (b) that because of the ensuing contradiction, the principle of irreducibility is to be rejected. It comes in two stages: stage one is to show that mental-to-mental causation entails mental-to-physical causation, while stage two is to show that mental-to-physical causation is possible only if we reject irreducibility. To see this, we have to get a bit technical.

Let \( M \) be a mental property causing another mental property \( M^* \). Because we are physicalists, we have to postulate a physical property as supervenience base for every mental property we postulate, in accordance with the principle of supervenience described above, so let us postulate \( P \) as supervenience base for \( M \), and \( P^* \) for \( M^* \). But now there are two reasons why \( M^* \) is here: because it is caused by \( M \) and because it is realized by \( P^* \). Now \( P^* \) is by itself sufficient for \( M^* \), so there seems little for \( M \) to causally contribute. In fact, the only way \( M \) can make a meaningful causal difference to the occurrence of \( M^* \) is if it causes its supervenience base \( P^* \) (stage one is now complete: mental-to-mental causation entails mental-to-physical causation). This, however, is not an option, for then we have two causes for \( P^* \): \( M \) and \( P \)—one mental and one physical. The exclusion principle clearly states that an event cannot have two causes. Therefore, whenever a physical and mental property compete as possible causes, one of them has to go. Which one? The mental, for, as the principle of causal closure states, we cannot allow the physical domain to be causally underdetermined. Hence, if we do not want to give up mental causation, we had better become reductionists.

As Figure 2 shows, \( M \) does not really cause anything in itself; the real causal relationship is between \( P \) and \( P^* \). That is, \( M \)'s causal powers are pre-empted by \( P \).
It seems that non-reductive physicalism, conceived as the combination of the five metaphysical principles described above, is an unstable position. In fact, the first four of these principles exclude the fifth one: we cannot account for the causal efficacy of irreducible mental properties. If we want to save mental causation, we are advised to drop the irreducibility principle, and embrace reductive physicalism instead.

If correct, this means that functional explanations give us nothing over and above the underlying causal explanations. As mental properties are a subset of functional properties, and there seems no principled way to limit the scope of the argument to include only this subset, it would seem that to adopt this reductionist strategy is to discard not only mental causation, but functional explanations tout court as illegitimate, and it is at this point that Kim’s specific brand of reductionism (the subject of the next section) would step into the fray as an alternative to non-reductive physicalism. Talk about anything like “mental causation” or intentional generalizations (again, the generalizations that refer to the kind of multiply realizable functional states favoured by traditional functionalism) as they figure in psychology leads only to excessive explanatory liberalism, where any explanation can be invented ad hoc. It would seem then that Kim’s argument has led us to the following conclusion: if we want functional and intentional properties to be of any causal efficacy, we have to embrace reductionism. But what does this reductionism look like? Kim has a proposal.

**Kim’s functionalizing reductionism**

Kim’s proposal involves a rejection of the non-reductive component of non-reductive physicalism, but it is not a return to the kind of reductionism assumed in standard defences of functionalism. Instead, Kim (2005) offers his own specific brand of reduction as an alternative; a position which we might label *functionalizing reductionism*. It holds that for a given property to be reduced, it must first be defined “functionally”: that is, the causal task it is to perform must be explicated. Heat, for example, is that property which, when it increases, causes wax to melt, mercury in a thermometer to rise, and so on. Next comes the job of finding the realizers of this property; in the case of heat, the realizer is molecular kinetic energy. It is the energy associated with the movement of the...
particles which performs the causal task of melting wax. Finally, we need a theory to explain how the realizer of the functionalized property performs its causal task.

In effect, Kim has divided the functionalizing strategy in three distinct steps (Kim, 2005, pp. 101–102):

Step 1: Functionalization of the target property
Property $M$ that is to be reduced is given a *functional definition* of the following form: having $M =_{\text{def.}}$ having some property or other $P$ (in the reduction base domain) such that $P$ performs causal task $C$.

Step 2: Identification of the realizers of $M$
Find the properties (or mechanisms) in the reduction base that perform the causal task $C$.

Step 3: Developing an explanatory theory
Construct a theory that explains how the realizers of $M$ perform task $C$.

For example, let $M$ be a mental property—say, being in pain. First, we stipulate that being in pain is having a property that fulfils the role of causing wincing and groaning or whatever. Next we try to identify the property that fills this role, and we find it in neurophysiology: for example (to keep with a venerable but scientifically mistaken tradition in the philosophy of mind), wincing is caused by C-fibres firing, which in turn is caused by tissue damage and trauma. Finally, it is up to neurologists to construct a theory that explains how being in pain is in fact realized by C-fibres firing, at which point the reduction as Kim envisages it is complete: being in pain = C-fibres firing. As such, the causal powers of being in pain and C-fibres firing do not compete, therefore being in pain can be said to be causally efficacious. Thus, by rejecting the irreducibility principle, mental causation is saved—pain really does cause behaviour.

However, it does so only by equating mental causation with physical causation. Moreover, if the supervenience argument forces us to accept this specific kind of reductionism, then it is not just mental causation that is pre-empted, but functions in general: both are understood in terms of the causal role they play, and in both cases, these causal roles are filled by some underlying, physical property which pre-empts their causal efficacy. So what is at stake here is the explanatory status of functional properties in general. On this account, the properties to which the reduced theories refer are nothing but empty shells. Thus, Kim’s reductionism seems to have eliminative consequences: functional properties are mere higher-order properties that are, ontologically speaking, nothing over and above the physical properties that perform these functions. Although this does not mean that, according to Kim, all higher-level explanations are vacuous, it does undermine the causal force of the kind of functions to which the explanations of cognitive psychology refer. The traditional, functional paradigm described in section 1, where a capacity is decomposed into different sub-capacities or -functions, merely trades in vacuous explanations, and, consequently, the prospects for the autonomy of the special sciences, as envisaged by Fodor, seem bleak indeed. That is, they seem bleak *insofar they are tied to that particular explanatory paradigm*. As we will see, however, there are strong reasons to think that traditional functional analysis misconstrues the explanatory practices of the cognitive
Introducing mechanistic explanations

In psychology (although the point is made more generally for the life sciences, see Machamer et al., 2000), theorists frequently refer to underlying mechanisms to explain phenomena. These mechanistic explanations are functional in the non-aetiological sense, as specified in the introduction: they seek to explain the behaviour of a mechanism by decomposing it into sub-tasks carried out by the mechanism’s constituents and their operations. In focusing on mechanisms, they provide an alternative to the kind of causal explanations which traditional philosophy of science took to be central to the practice of scientific explanation. The past decade has seen the rise of a substantial debate about the nature of mechanistic explanations (Bechtel & Abrahamsen, 2005; Glennan, 2002; Machamer et al., 2000; Tabery, 2004).

So how do mechanistic explanations work? To explain a systemic activity of a system (e.g., a cognitive capacity in humans), one identifies the mechanism responsible for it. One level below that of the mechanism, we find its working components or parts. These parts may themselves be mechanisms, whose parts are on a still lower level. Here levels can be assigned only locally, with respect to a particular mechanism: they are developed for specific exemplars and there is no way or need to anchor them in a universal, basic level in nature (Bechtel & Abrahamsen, 2005). One of the reasons for this is that mechanistic explanations typically incorporate contextual and/or environmental data, which may vary from case to case. The point is that the mechanisms are realized via a constituency relation: the parts and their organization, on whatever level they might be, generate the mechanism.

Let us consider the example of Stricker and Verbalis’s mechanistic explanation of fluid homeostasis as cited by Craver (2007, p. 9). After eating salty food, or after sweating without replenishing the lost moisture, the level of plasma osmolality rises. This causes vasopressin to be released, which in turn helps the body to conserve water in a number of ways, and evokes the feeling of thirst. In this example, an explanandum (osmoregulation) is explained by decomposing it into the different sub-routines and operations of different mechanisms (e.g., the pituitary releasing vasopressin). These mechanisms are located at different levels, from behavioural to molecular.

As this example shows, mechanistic explanations assume a very local account of levels. The mechanistic explanation of fluid homeostasis “oscillates up and down in a hierarchy of mechanisms to focus on just the items that are relevant” (Craver, 2007, p. 10). For our purposes, it is important to note that there seems to be a determination relation between fluid homeostasis, which is an activity of the body, and several operations conducted by different parts of the body. These parts and their activities “fix” the activity of fluid homeostasis of the body as a whole. It seems that we have a case of interlevel explanation on our hands that cites parts and operations on different levels, from the molecular to the behavioural. Moreover, as the mention of parts already suggests, mechanisms are composed of hierarchically related entities and operations, and as such stand in a constitutive or mereological relation to their parts. In the fifth section of this article,
where we argue for the first (metaphysical) claim we made in the introduction, this mereological relation will become important.

How does this type of explanation relate to the traditional functional analyses described in the first section? For a start, there is a significant similarity: they both involve decomposition. Mechanistic explanations, too, typically start with decomposing the capacity or function to be explained into different sub-functions (in the example, fluid homeostasis is decomposed into sub-functions like “increasing plasma osmolality” and “releasing vasopressin”). However, as we have seen, a mechanistic explanation goes beyond that, in that it also specifies what entities perform those functions (in the example of fluid homeostasis, such entities are “plasma” and “the pituitary”). Unlike traditional functional analyses, then, mechanistic explanations do not remain silent as to what fulfils a certain (sub-)function. As we have seen, Craver even goes so far as to say that only to the extent that we have identified, to a varying degree, the entities responsible for a given (sub-)function is a mechanistic model truly explanatory with respect to the target function (Craver, 2006). Thus, one could say there are two types of functional explanations: those focusing solely on the causal roles a particular system exhibits, and those that also take into account implementational details of these roles. While Kim’s argument is primarily directed at the former, it is the latter that warrants a re-evaluation of the argument.

It seems that in scientific practice, researchers are not content to give a purely functional analysis of a cognitive capacity. At least some information about the entities involved in bringing about the overall behaviour of the mechanism should be included for the model to constitute a genuine explanation. In our view, these developments call for a reassessment of Kim’s supervenience argument. Exactly how do mechanistic explanations pose a problem for the supervenience argument, and for the ensuing functional reductionism that is presented as an alternative? At first glance, it might seem that mechanistic explanations, with their joining of operations with their respective realizers, provide just the identification that Kim’s functional reductionism relies on. As we shall see, however, there is a more than superficial difference, as the systemic activities or functions realized by mechanisms go beyond those of the realizing components. In what follows, we shall offer a mechanistically inspired critique of Kim’s argument. In the next section we will argue that the metaphysical picture that emerges from the recent discussion on mechanistic explanation withstands the supervenience argument, while in the final section, we shall make the case that Kim’s alternative of functionalizing reductionism gets the explanatory practice wrong.

The levels controversy: Supervenience or mereology?

What are the relata of the supervenience relation? In early versions of Kim’s supervenience argument, the relata were characterized mainly by examples. The targets were mental states, events, or properties (hunger, pain, hope, etc.). For reasons of clarity, let us stick to Kim’s preferred terminology and take the supervenience relation to hold between mental properties and their physical realizers: a certain instance of being-in-pain corresponds to a certain instance of having a physical property (like firing neurons). So far so good. However, though superficially clear-cut, this terminology hides a fundamental ambiguity.
This ambiguity was brought into the open when Kim defended his supervenience argument against so-called generalization problems raised in the literature. The gist of these objections was that the supervenience argument proved too much: it threatened the causal efficacy of not only mental properties, but macro-properties as such (Block, 2003; Bontly, 2002; van Gulick, 1992). After all, one might ask, what is so special about mental causation? What is there to stop the supervenience argument from generalizing to all instances of higher-level causation? The supervenience relation is generally thought to hold not only between mental and physical properties, but also between biological and chemical properties, between chemical and molecular properties, and so on. In all these instances, the causal efficacy of the properties at the level in question is excluded by the causal efficacy of the respective subvenient properties at the level directly below it, which in turn is excluded by the causal efficacy of the respective subvenient properties on the level below that, and so on. It would seem that “it is a consequence of the Causal Exclusion Argument that all macro level causation . . . is an illusion” (Block, 2003, p. 133).

On this account, the causal powers “drain away” to the bottom level. Only at the most fundamental level in physics is causation to be had. Even worse, if there is no bottom level, causation would not exist altogether! In the words of Block (2003):

> If there is no bottom level and there is endless subvenience, Kim’s causal exclusion argument would yield absurd results. The exclusion principle . . . is to blame and should be rejected. Even if there is subvenience only down to elementary particle physics, there is a problem of causal powers draining down that far. (p. 140)

Of course, this argument takes the form of a reductio, so it is only as strong as one’s commitment to the causal efficacy of macro-properties. Nevertheless, it is a hefty price to pay, and Kim is not prepared to bite the bullet on this one, as he makes clear in his Mind in a Physical World (1998): “[M]acroproperties can, and in general do, have their own causal powers, powers that go beyond the causal powers of their microconstituents” (p. 85). However, this leaves us with something of a conundrum. True, the five metaphysical principles we described in the second section of this paper are fundamentally at odds with each other, and Kim is to be credited for pointing that out, but the supervenience argument itself does not compel us to reject one of those principles rather than another (Walter, 2008, p. 684). Kim’s rejection of irreducibility is an interesting choice from a philosophical point of view, but it might be more advisable, in light of Block’s objection, to reject exclusion instead (or indeed one of the other principles).

So how can Kim have it both ways? That is, how can he maintain that the causal powers of mental and functional properties are pre-empted by those of their realizers, without also threatening the causal powers of macro-properties in general? Kim’s initial solution was to distinguish between levels and orders (1998). Basically, it is a distinction between levels, at which ontologically new properties with their own causal powers can emerge, and the orders that can be discerned within such levels. Different entities or objects are located at different levels, while within these levels there is a hierarchy of different properties belonging to the same object, related to each other by supervenience. First- and second-order properties (and, indeed, any properties still further up the order hierarchy) are properties of the same object. Because they are properties of the same object, they are
located at the same level in the micro/macro hierarchy. Thus, according to Kim, mental properties are second-order properties; the causal powers they have are not emergent, but are determined by the first-order physical property on which they supervene. The supervenience argument aims to show that second-order properties have no causal powers in addition to those of their first-order realizers. Therefore, the argument does not apply to higher-level properties: “[T]he realization relation does not track the micro–macro relation” (Kim, 1998, p. 82). Because they belong to different entities, higher-level properties do not supervene on lower-level properties, and so there is no danger of the causal powers of the higher-level properties draining to a lower level. To summarize these somewhat technical metaphysical considerations: in effect, Kim’s strategy is to admit that macro-properties are causally efficacious, but to deny that functional properties (including mental properties) fall into that category. Thus, it seems there is one class of properties which do have emergent causal powers, because they are realized by properties of another object (Kim, 1998, p. 84), and one class of properties that are realized by properties of the same object, and therefore have no causal powers of their own. The properties addressed in functional psychology belong to the latter category, and therefore cannot figure in genuine explanations.

It should be noted that in his 2005 book, Kim no longer explicitly invokes the distinction between levels and orders. Instead, he relies on the notion of “micro-based properties” to stop the supervenience argument from generalizing to all instances of macro-causation. This has led to a complicated discussion about the multiple composition of macro-properties and intra-level drainage, of which it is doubtful that it will resolve favourably for Kim (see Walter, 2008, for an extended discussion).

We feel, however, that Kim needs something like the levels/orders distinction if he wants to avoid inconsistency. Consider the following quote:

Indeed, the supervenience argument may be generalizable, but all that would show is that if there is biological causation, biological properties are, or are reducible to, physical or physico-chemical properties; it does not show that biological causation does not exist. The epiphenomenalist brunt of the argument is avoided if one is prepared, and is able, to choose the reductionist branch of the dilemma. (Kim, 2005, p. 55)

On this view, then, the following holds for any given property but the most fundamental: if it is causally efficacious, it must be reducible. Now Kim (2005) still thinks that “the supervenience argument does not eliminate all macrocausation” (p. 56). Hence, there must be some macro-properties that are reducible, but in order to be reducible, a property must have a functional role that can be specified; a role that is “filled” by its realizing property. In other words, if macro-properties are to be causally efficacious, and the only way for a property to be causally efficacious is to be reducible, then, assuming Kim’s brand of functional reductionism, these macro-properties must stand in exactly the same relation to their realizing properties as do mental properties. However, it is precisely the fact that mental properties stand in this relation to physical properties that prevented them from having any causal powers over and above those of their realizers in the first place, so it seems there is no room for macro-properties that “have their own causal powers . . . that go beyond the causal powers of their micro-constituents” (Kim, 1998, p. 85).
The upshot is this: If the causal efficacy of macro-properties is to be of a more robust kind than just inherited causation, Kim cannot allow the supervenience argument to generalize, and must restrict the scope of the supervenience relation so that it does not include macro-properties with their own emergent causal powers over and above the causal powers of their subvenient properties. The only viable way, it would seem, is invoking the distinction between levels and orders. Without that distinction, Block’s objection stands, and we must either give up reducibility, or concede to an all-encompassing epiphenomenalism. Thus, within Kim’s brand of reductionism, the awkward consequence of banning all robust macro-causation is best avoided by restricting the scope of the supervenience relation, so that it applies only within a level.

However, it is precisely here that we run into trouble with mechanistic explanations. As we have seen, mechanistic explanations assume a multilayered hierarchy of entities and activities that are related *mereologically*: that is, stand in a whole–part relation to each other. It would seem to us that Kim’s metaphysical account of levels fails to appreciate the subtle relations that can obtain between levels of organization in biological and psychological mechanisms. In fact, mechanistic explanations employ an alternative account of levels (Bechtel, 2007). Often, reductionists adhere to the “scientific disciplines” approach to levels, where psychology constitutes one level, below that we have biology, still further down is chemistry, and so on. This notion of levels cannot accommodate the entities referred to in mechanistic explanations. An alternative, more appropriate conception of levels in this context is to view them as designated by the different *sizes* of the entities involved, not by the discipline that studies them. In the case of fluid homeostasis, it is roughly the size of the entities involved that determines the level. In any case, the metaphysical picture emerging from mechanisms is one of mereology, where the mechanism as a whole is situated on a higher level than its parts.5 As Kim defines interlevel relation to hold between different objects, and parts and wholes are certainly different objects, then either we stick to the distinction between levels and orders, in which case mental and functional properties, as addressed by mechanistic explanations, are situated on different levels and thus not affected by the supervenience argument, or we reject the distinction, but then invite the problem of causal drainage.

**Why functionalizing reductionism doesn’t work: The case of exact arithmetic**

In the previous section, we have seen that the metaphysical picture that is assumed by Kim’s supervenience argument does not match up to the account of levels used in mechanistic explanations. But what about his functionalizing reductionism? It might be that Kim gets the metaphysics wrong, but that his reductionism still offers a sound description of scientific practice, and that in fact “most cases of interlevel reduction conform to the model I have just sketched” (Kim, 1998, p. 99). However, we fear that the actual explanatory practice of the cognitive sciences suggests a more dynamic and diversified account than Kim’s reductionism has to offer.

Clearly, mechanistic explanations involve *some kind* of reduction, as they refer to the components of a mechanism to explain its behaviour. However, they differ from standard accounts of reductive explanations, including Kim’s, in that this reductive side of
mechanistic explanation by itself is insufficient to explain the behaviour of a particular mechanism; one must also take into account both the way the parts of that mechanism are organized, and the (causal) relations between the mechanism and its environment. In other words, it represents an alternative model of interfield integration (Craver, 2007). This model calls for additional, autonomous investigation over and above the reductive explanation (Bechtel, 2007). Thus, as the example goes, although the process of yeast fermentation presumably takes place at the level of biochemistry, with yeast cells and enzymes interacting with oxygen to turn glucose into alcohol, this process requires the brewer to have met all kinds of environmental conditions (Bechtel, 2007). If the brewer does not create the right kind of environment, fermentation will not occur. Moreover, the knowledge of brewers to create these conditions predates by centuries the advances made in biochemistry that are needed to understand the process of fermentation.

Before we evaluate Kim’s functional model of reductionism, let us consider one example of a mechanistic explanation in somewhat more detail: the explanation of numerical cognition proposed by Nieder and Dehaene (2009; see also Carey, 2009). We will show that this is a good example of (interlevel) mechanistic explanation.

In their model, Nieder and Dehaene distinguish two basic types of arithmetic skills: approximate versus exact arithmetic. Evidence from behavioural and functional magnetic resonance imaging experiments suggests that approximate arithmetic relies on non-verbal visuo-spatial cerebral networks, whereas exact arithmetic requires a language-based format (Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999). Thus, basic numerical competence is not a cultural achievement, and is found in some animals, in infants, and in humans living in a language environment lacking number symbols. The precise enumeration and calculation skills of adult humans in more advanced cultures presuppose not only approximate arithmetic, but also appeal to semantic aspects of numerical quantity (Nieder & Dehaene, 2009).

In providing a mechanistic explanation for precise enumeration, we break this mental activity down into two sub-routines: approximate and exact numerical cognition—in this respect, the mechanistic explanation is similar to the traditional functional analysis described in the first section of this article. However, the model Nieder and Dehaene construct goes beyond that in that they identify the entities responsible for these operations: the two sub-routines approximate and exact numerical cognition are realized by operations in different parts of the brain—the former by activity in the prefrontal and posterior parietal lobes, the latter by activity of the intraparietal sulcus (Nieder & Dehaene, 2009). Thus, a mechanistic explanation of the systemic activity of precise enumeration or calculation fans out into lower levels, including neurophysiology.

With this concrete example in mind, can we say something about the accuracy of Kim’s functional model of reduction as a description of actual scientific practice? We think there are two problems.

First, Kim seems to believe that the first step of his model, the functional description of a property, forms a kind of a priori, conceptual prelude to the empirical handiwork. Before the “scientific work can begin in search of the realizers” (Kim, 2005, p. 164), the scientists sit down and muse a while over what causal tasks are to be performed by the target property. This is not very likely. In the example of numerical cognition, the point is not that a property like “being able to represent numbers” *cannot* be functionally
defined in Kim’s sense, but rather that it simply isn’t. To be sure, Nieder and Dehaene (2009) begin by giving a characterization of numerical cognition as a capacity which enables us to achieve scientific and technological advancements and so on (p. 186), but these are merely remarks to underline the scientific importance of the explanandum, not premises in an argument. In fact, the only type of psychological explanations that often defined the target property or function in terms of the causal tasks they performed are behaviouristic, where a belief that it is about to rain is defined in terms of the disposition to run for shelter, open umbrellas, and so on. But as behaviourists explicitly choose to remain silent about what goes on between input and output, they never advance to step 2 (the realizer-identification stage). Instead of an argument beginning with a functional definition of a target property, mechanistic explanations target an activity or capacity of a system that is presumed as given, and proceed to explain it by providing a model of the underlying mechanism.

Second, the next step in Kim’s functionalizing process, which consists of finding the realizer that performs the causal tasks defined in step 1, glosses over an important difficulty. Consider again that mechanistic models explain a certain property or capacity exhibited by a system in terms of the underlying mechanism: that is, its parts, their activities, and their organization (Bechtel, 2007). To be true to the model, we would have to say that the combination of parts, activities, and their organization that make up the mechanism fill the causal role that is the overall behaviour of the mechanism. Yet while there is no problem with identifying the parts responsible for certain operations, since both are located at a lower level than the explanandum, the level at which the organization of these parts and activities is to be found (information about which is necessary for the mechanistic explanation to be successful, see Bechtel, 2007, p. 183) is not so clearly characterized. In the example of numerical cognition, while the activities’ approximate and exact numerical cognition and their respective neural realizers can without controversy be said to be located at a lower level than general numerical cognition, it is far less clear at what level the organization of these parts and operations is situated. Even worse, a genuine mechanistic explanation also invokes the causal interactions between the mechanism and its environment (Bechtel, 2007, p. 183) and might be discovered to be part of another, still larger mechanism. In numerical cognition, important environmental influences include cultural and linguistic education, natural selection, and so on. Though these factors reside on a level higher than that of the original explanandum of numerical cognition, it can be necessary to incorporate information about them in the mechanistic explanation.6 This means that mechanistic explanations do not always proceed downwards, which clearly presents a problem for Kim’s functional model of reduction, considering the requirements to which genuine reductive explanations have to adhere:

Not only must the explanans of a reductive explanation of F not refer to F, but it also must not refer to any other property at the level of F—or, equivalently, a reductive explanation of F may refer in its explanans only to properties at levels lower than that of F. (Kim, 2005, p. 106).7

Clearly, this means that Kim’s model of functionalizing reductionism is at a severe disadvantage compared to mechanistic explanations, which typically involve “Looking down, around and up” (Bechtel, 2009). Therefore, it seems that this functional model of
reduction cannot accommodate important aspects of a type of explanation that features prominently in the life sciences.

At heart, the problem is that Kim’s model takes the form of an argument, with a definition among its premises and an identity as a conclusion, and this format is not suitable to capture mechanistic explanations. Mechanistic explanations, because they are a type of functional explanation, do not subsume the explanandum in question under a general rule; rather, they identify and provide a model of the mechanism responsible for it. They explain why by showing how. To this end, they involve the functional decomposition of the mechanism in question. Though this decomposition strategy is in a sense reductive, its reference to the way the parts are organized, together with the fact that these parts are located at different levels, vindicates the autonomy of these explanations (Bechtel, 2007).

Conclusion

Functional explanation, understood as explaining a capacity by decomposing it into several sub-tasks, has for a long time played an important role in cognitive psychology. However, among theorists in psychology, there is an increasing tendency to focus on mechanistic explanations, which differ from traditional functional analyses in a number of important respects. In this article, we have explored whether this change poses any new problems for Kim’s position on reductionism and functional explanation in psychology. We have identified two important challenges.

First, the metaphysical framework underlying the supervenience argument, which for Kim constitutes the main reason to adopt reductionism in the first place, is not compatible with the framework of mechanistic explanation. In order to avoid awkward consequences like Block’s causal drainage, the scope of the supervenience relation has to be restricted, but of course not so much that it becomes insignificant or trivial. It is walking the tightrope between proving too much and proving too little. In any case, the restriction Kim has made by distinguishing between levels and orders excludes the systemic behaviour of mechanisms, since this behaviour depends mereologically upon the mechanism. As mental causation was the original target of the supervenience argument, and mechanistic explanations are now being used in the cognitive sciences to explain mental capacities or functions, this is something that should worry the reductionist à la Kim.

Second, even if we ignore the metaphysical difficulties and embrace reductionism, Kim’s alternative model does not work. It is not the case that scientists first provide a functional characterization of the phenomenon they wish to explain, nor is it true that the entities or activities the explanans refers to are always found on a lower level than the original explanandum.

To what extent is this bad news for Kim’s programme? To be sure, the foregoing analysis has not shown the supervenience argument to be invalid, nor the functional model of reductionism to contain inconsistencies. However, with regards to our main question, namely how Kim’s challenge to functional explanation in psychology fares in light of the new developments in the form of mechanistic explanations, we conclude that Kim’s position runs into metaphysical problems, and falls short of empirical adequacy.
Acknowledgements
The authors thank the members of the section Theoretical Philosophy of the Faculty of Philosophy, at the VU University Amsterdam for their comments on an earlier version of this paper.

Funding
Raoul Gervais acknowledges support by the Research Fund Flanders (FWO) through grant nr. G.0031.09 for this article.

Notes
1. That is to say, functional state types do not map onto physical state types, hence classic reductionism is rendered impossible from the outset; this is the non-reductive part of non-reductive physicalism. However, the physicalism part requires that functional state tokens do map onto physical state tokens. Thus, a particular mental state token may be realized by one particular brain state at time $t$, and by another particular brain state at $t_2$, but there is nothing to guarantee that these brain states are of the same type, so reductionism does not get off the ground. However, both mental state tokens may play the same functional role, by virtue of which they are thought to feature in psychological explanations.

2. Similar functional models of reduction were proposed by Chalmers (1996) and Jackson (1998).

3. Again, this causal-role sense of functional explanation is different from the aetiological sense. In the example above, an aetiological explanation would explain the presence of the function bearer, in this case the mechanism, by referring to its function: osmoregulation. Alternatively, if we want to fix osmoregulation as the explanandum, the aetiological explanation would explain its presence by providing a story about the function that osmoregulation plays in the survival of the organism. However, we are trying to draw attention to the fact that the systemic behaviour of the system (fluid regulation) is explainable in terms of the causal role/functions that are filled by the parts of the mechanism, where one of the interesting points is that these parts are located at different levels.

4. Compare also Bechtel and Abrahamsen’s 2005 article, where they say: “Although researchers might frequently differentiate component operations before linking them with parts, or identify component parts without yet knowing what operations they perform, the ultimate goal is to link operations with parts” (p. 434).

5. Recall that, as we noted previously, in mechanistic explanations, levels can be assigned only locally: that is, with regard to a specific mechanism. Of course, if we understand levels in a mereological sense, this same conclusion should hold, as indeed it does: What counts as a part, and what as a whole can be decided only relative to the particular mechanism under investigation. The point then is not that a functional property may or may not generalize across different species with different realizing mechanisms (humans, Martians, octopuses, or whatever), but that in a given mechanism, the entities are of different sizes and hence cannot be related through supervenience as Kim understands it, and are thus not epiphenomenal. Whether and how such functional states generalize across different realizers is not at issue here.

6. Of course, in stating laws or generalizations about the actual role phenomena play in a mechanism, such environmental information may not be required. That is, there might be a loose sense of “reduction” which is not precluded by these environmental factors. However, Kim’s functional brand of reductionism, with its exclusively downward direction, is precluded as a general research strategy.

7. In this quote, Kim uses the word “level” in the ordinary, unspecified sense (not as opposed to “order”).
8. But is it not a limitation of mechanistic explanations that they do not generalize? They *do* generalize, but in a less straightforward way than via laws. Generalizations of mechanistic explanations typically involve model cases, or exemplar theories. A particular case is chosen (for often pragmatic reasons—see Bechtel & Abrahamsen, 2005) which then serves as a model. Next, one identifies the differences between a new case and the model, and assigns a level of importance to these differences. So, rather than having every instance of a law exemplify that law in equal measure, as the deductive-nomological model would have it, we end up with a degree to which a certain mechanism exhibits typical traits (Bechtel & Abrahamsen, 2005).

References


**Author biographies**

Raoul Gervais is a Ph.D. student in the Centre for Logic and Philosophy of Science at Ghent University. Currently, his research focuses on pragmatic aspects of scientific explanation, in particular in the disciplines of psychology and biology. Address: Centre of Logic and Philosophy of Science, Faculty of Arts and Philosophy, Ghent University, Blandijnberg 2, 9000B, Ghent, Belgium. Email: Raoul.Gervais@UGent.be

Huib Looren de Jong is Assistant Professor of theoretical psychology at the Free University of Amsterdam. His interests include foundations of empirical research in cognitive and neurosciences, philosophy of mind, reduction, and levels of explanation. Address: Department of Theoretical Psychology, Free University of Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands. Email: h.loorendejong@vu.nl