Functions and Emergence: When Functional Properties Have Something to Say

Abstract

In a recent paper, Alexander Bird (2007) has argued that some higher-order properties – which he calls "evolved emergent properties"– can be considered causally efficacious in spite of exclusion arguments. I have previously argued in favour of a similar position. The basic argument is that selection processes do not take physical categorical properties into account. Rather, selection mechanisms are only tuned to what such properties can do, i.e., to their causal powers. This picture seems ultimately untenable in the light of further exclusion problems; but at the same time, it meets our explanatory demands. My purpose is therefore to show that there is a real antinomy with regard to evolved emergent properties. I develop a physicalist exclusion argument and then I go on to consider an argument that seems to establish that evolved emergent properties *are* causally efficacious, and propose a compatibilist solution. Finally, I very briefly consider what the proposed model may imply for the issue of mental causation.

1. Introduction

In a recent paper Alexander Bird (2007) has argued that some higher-order properties can be considered causally efficacious in spite of exclusion arguments. I have argued for a similar position in author (2002, 2004a). The basic idea developed in those three papers is that selection processes do not take physical categorical properties into account. Rather, selection mechanisms are only tuned to what such properties can do, i.e., to their causal powers. This is made explicit in Larry Wright's (1973) old definition of teleological function, according to which: that *x* has the teleological property *F* means first that *x F*s and further that *x* "is there" because it *F*s, that is, because it has a certain functional property. So, if there is a categorical/functional distinction, then so far as selection is concerned, it is functional properties which can claim causal efficacy. Another way to explain this point is: functional properties are said to be causally inefficacious because their putative effects can be causally explained by the action of categorical properties[1]. Thus the cause of my falling asleep is not my ingesting a sleeping pill, but my ingesting whatever chemical realizes the property *being a sleeping*

[1] See especially Prior, Pargetter and Jackson (1982).

pill in this particular case. However, when a causal process is part of a selection process, things are reversed: a heart is not selected for because it has this or that physical property *P*, but because it has the functional property *being a blood pump*, which is realized by *P*[2]. In a nutshell, functional properties are causally efficacious when they are converted into *teleological* properties.

This idea is what makes Bird claim that emergent properties are *evolved* emergent properties. The existence of selection processes implies the involvement of functional, i.e., higher-order, properties in the world's causal web. Emergentism claims that not all causally efficacious properties in the world are identical to, or reducible to, physical properties. So, if evolution has been blind to categorical properties, and having a certain functional property is what explains the "being there" of a certain entity, then there are emergent properties: namely, all those functional properties that have developed a role in evolution. Minimally, these constitute the realm of biology. Furthermore, insofar as psychological properties are the result of a process of selection, there are also emergent psychological properties. Moreover, there are emergent psychological properties regardless of evolutionary processes as long as we concede that neural structures can be the object of within-individual selection[3].

My purpose in this paper is to argue for this basic picture after first acknowledging that it does, however, seem untenable in the light of exclusion arguments. I want to argue that there are good explanatory reasons for adopting the picture I defend, and that such explanatory reasons overrule the conclusions drawn from the discussion of exclusion arguments. Thus, I will try to convince the reader that there is an antinomy with regards to evolved emergent properties and that the sounder way to go, based on our explanatory commitments, is to adopt a compatibilist stance regarding exclusion arguments. In Section 2, I discuss the reasons we have to reject the idea that functional

---

[2] In what follows, I will assume that the realizers of functional properties, i.e., their categorical bases, are physical properties, and I will not distinguish between physical properties and categorical properties or between physical processes and processes that involve only categorical properties unless strictly necessary. I will also assume that dispositional properties should be understood as second-order functional properties. As I am going to argue that functional properties can be causes (in selection or recruitment processes), I may occasionally refer to them simply as emergent properties. However, even though I think that the context will make it clear what I refer to in each case, I have to note that not every occurrence of "emergent properties" in this paper refers just to functional properties.

[3] Teleological accounts of content such as Dretske's and Van Gulick's (see below) hold that the neural structures of an individual subject can be selected on account of their causal powers. Ruth Millikan's (1993) theory disputes this idea. According to her account, talk of teleology makes sense only when whole species are considered.

properties have anything to say in causal matters. In Section 3, I consider an argument that seems to establish that evolved emergent properties *are* causally efficacious. Once we arrive at the antinomy, I will argue for a compatibilist solution regarding the exclusion problems discussed in Section 2. I close by very briefly considering what the model I defend may imply for the issue of mental causation.

2. The argument from exclusion

There is a problem with the general picture I just summed up. The picture responds to an exclusion problem between functional properties and their categorical bases. Usually, categorical bases are said to screen off functional properties. The emergentist answer is that in selection processes, the direction of this screening off relation is reversed: selection processes are concerned with functions, not with their realizers. Some effects are caused by the instantiation of functional properties (i.e., a heart "is there" because it pumps blood). However, this is not the end of the story, for the physicalist can pose a new and different exclusion problem that can be summarized as follows:

(a) the heart's being there is a physical effect;

(b) according to the principle of the causal completeness of the physical (CCP) every physical effect (i.e., caused event) has physical sufficient causes;

(c) the heart is there because it instantiates the functional property *being a blood pump*;

(d) the heart's being there is not causally overdetermined.

The idea is that even if the action of the categorical base does not exclude the action of the functional property, surely there is another physical cause that excludes it. As a matter of fact, things do look that way: the heart's being there is the end point of a long causal story which can be told in purely physical terms. The existence of a particular heart can be said to be due to the physical causes that account for its development, and this would make the functional explanation idle.

In the late eighties, Fred Dretske (1988, 1993) proposed an account of mental causation according to which, *carrying the information that F* was regarded as a "structuring cause" of a given behaviour *B*. For instance, suppose that a neural structure carries the information that there is a tiger over there; such a neural structure may be

recruited to inform about the presence of tigers, connected to motor neurons and ultimately to a fleeing movement. Thus, the neural structure's property of *carrying the information that there are tigers here* can be regarded as a cause of the fleeing movement (it was connected to such a movement because it carried that information). However, one problem arose for Dretske's theory when Terry Horgan (1991) pointed out that the recruitment process by which the neural structure acquired the function of carrying a certain information *and* of ultimately causing a given behaviour did have a physical explanation: "[I]n principle, the recruitment process also is explainable in more fundamental terms, and at lower levels the correlation [carrying the information that F] will play no explanatory role" (p. 84). I think this same problem applies, *mutatis mutandis*, to the present account of higher-order causation: there is a lower-level explanation of an effect, such as a heart's being there, and the functional property does not figure in it.

The literature on the exclusion problem for mental causation is huge; however, despite all the attempts to solve it, the problem persists[4]. Is the situation any better when it comes to confronting the problem spelled out above? At first sight, it seems that we are no better off. Counting the heart's being there as a physical effect is controversial: it is not clear what can be considered as physical in any event[5] and it is even more difficult to explain what is to be considered physical so that the physical world is indeed causally closed. However, the burden of proof here is on the person who denies that the existence of a certain heart is a physical effect or is reducible to a physical effect. After all, the development of a heart amounts to an investment of energy in structure. If that is not a physical effect, what is it? Something along the same lines can be said about the CCP: its truth cannot be taken for granted, despite the fact that many philosophers claim that it is a truth of contemporary physics (see author, 2006, for discussion). Yet, we seem to be on firmer ground if we assume it to be the case than if we do not. In any case, the point is that *there is* a physical explanation for the heart's being there[6]. Finally,

---

[4] This is clearly a controversial claim. Philosophers such as Bennett (2003) and Menzies and List (2009), among the authors recently developing new approaches, would think otherwise. The focus of this paper is not mental causation, but it is possible to claim that neither Bennett's compatibilism, which draws on Yablo's (1992) insights, nor Menzies and List's counterfactual approach, is a satisfying solution. Both approaches tend to ignore the fact that not all asynchronous counterfactual dependencies are causal dependencies, and that the relation between functional properties and their realizers is different from the relation between, e.g., determinables and their determinates (more on this in Section 3).

[5] See Hempel (1980), Crane and Mellor (1990) and Montero (1999).

[6] Towards the end of the paper, I will argue that such a physical explanation might not be complete.

it is safe to exclude the idea that the heart's existence has two complete and independent causes that can co-exist; one couched in functional terms and one a physical explanation. Failing to exclude such a possibility would go against the assumption that nature is simple.

However, there is a difference between this exclusion problem and the exclusion problem for mental causation. Exclusion problems are generated by what Jaegwon Kim (1993) called "the principle of causal-explanatory exclusion", which says that there cannot be two complete and independent causes for a given effect. When we are talking about mental causation, mental and physical causes look like irreconcilable rivals since both of them seem to suffice to bring about the effect and there is no dependency relation between them that can be seen as a way out of the exclusion problem. Supervenience, realization and determination were offered as candidates, but none of them succeeded[7]. The lesson to be extracted from these failures, I think, is that what the exclusion principle asks for is not a relationship between properties, but one between causes. That is, what needs to be shown is not that mental properties depend on physical properties or vice-versa but that mental causes, or mental causation, to put it more clearly, depends on physical causation, or vice versa. To put it in yet other words, the action of mental properties must depend on the action of physical properties, or the other way around (see Kim, 1993).

In contrast to this, the confrontation between the functional explanation and the physical explanation summed up in (a-d) may be said to be only apparent. The functional and the physical explanations are *not* two independent causal explanations. Rather, the functional explanation explains the physical explanation, for it explains why the physical process takes place. This kind of functional explanation is able to tell us why a certain process takes place in the way it takes place and, as a result, why e.g., hearts grow and endure, while a lot of other things are eliminated. We cannot answer these questions without mentioning functional properties. We have a causal story couched in physical terms, but still we can ask ourselves why such a causal process took place. We cannot do that on most occasions: things just happen the way they happen. However, at least since living organisms appeared in the world, there are some causal processes whose occurrence can be explained; answers to "why questions" that go

---

Whether it is sufficient or not depends precisely on our understanding of what it is for an explanation to be sufficient. Perhaps an explanation can be considered sufficient in one context, but not in another.

[7] See author (2001) and Walter (2007).

above and beyond the chain of local causes, and that explain why such a chain exists in the way it exists. As I said, this is not the usual case: there is no such explanation for the movement of the planets, or for the collapse of wave-functions. But there are also processes of selection in the world, and the occurrence of those processes can be explained only by mentioning functional properties, since it is in virtue of the functional properties that a particular instantiates that it is selected—or its categorical physical properties are selected. I take it that we cannot really understand how life evolves if we do not take functional properties into account. We could have an adequate account of the order (even the causal order) of events and a good record of the types of things that endure and the types of things that disappear; but we would lack an explanation of why things happen that way.

So suppose the physical explanation for a heart's being there covers the development and ongoing existence of a heart from a physical point of view. The idea is that we can ask why such a process takes place at all, and answer that it does so because the heart is a blood pump, and having a blood pump is good for an organism. In general, we can say that the living world is the way it is because natural selection has shaped it that way, and that in selection processes what is important is what entities can do, regardless of their constitution.

However, a question springs to mind, namely: does the action of physical causes really depend on the functional properties of anything? It seems that the CCP excludes such a possibility. The process that we say has to be accounted for in terms of what entities can do must have, and indeed has, a triggering cause which is physical; for such a process is nothing but a chain of physical causes and effects which has as its origin a physical cause. That is, if the response described above makes sense and constitutes a plausible solution to the exclusion problem (a-d), then it seems we have not got very far; in fact, we have taken a step backwards, so to speak. Let me explain.

We started with an exclusion problem: that between functional properties and their realizers. We said that in selection processes categorical properties are not the causally relevant properties: a heart is there not because it has these or those categorical physical properties but because it is, or can be, a blood pump. However, this response was not entirely satisfactory, because even if categorical properties are not *the* cause of a heart's being there, there must be a physical explanation for its existence: the heart is there because it is the end point of a chain of physical causes and effects that account for its

existence. So we have a new exclusion problem. I said that this new exclusion problem admits a solution: the physical explanation is in turn explained by the functional explanation. In general, physical processes involved in selection processes occur because selected entities have certain functional properties. If those entities had not had such functional properties, the physical story of the world would have been different. Yet if this is the claim, then we have a new problem, for the physical story of the world is simply a chain of physical causes and effects with no place for external influences. Thereby, any discrete piece of that chain has a causal antecedent that is physical, and this is contrary to what is suggested by saying that some physical processes (those involved in selection processes) have functional causes.

So it seems that we arrive at a dead end. It is the same kind of dead end that haunts other contemporary emergentist proposals. For instance, in a series of work, Charbel Niño El-Hani and others[8] have struggled to come up with a notion of downward causation which nonetheless respects the CCP. The idea proceeds as follows. Exclusion arguments for physicalism are enthymematic arguments. One of the premises left implicit is that all causation in the world is of the same kind, that is, that causality is a homogeneous kind. To my knowledge, only Tim Crane (1995) has made this assumption explicit in the context of the mental causation debate. However, Emmeche, El-Hani and some other philosophers of biology, influenced by the work of Kauffmann, hold that we should distinguish between at least two kinds of causation. Following Aristotle, they have labelled these, "efficient" and "formal" causation, and their idea is that some effects can have both kinds of causes: efficient and formal. Thus, physical properties can *efficiently* cause their effects, whereas emergent properties *formally* bring about those same effects.

According to this account, emergent properties are non-physical properties that self-organizing complex wholes have. For instance, a living cell, an organ such as a brain and an individual organism are self-organizing self-preserving (*autopoetic*) complex wholes. They are ultimately composed of physical entities, but they have a peculiar tendency for self-preservation and to arrange their constituent parts so that the whole functions in a particular way. In turn, the notion of formal causation comes to this: emergent properties, which according to this account are properties of these self-

[8] See El-Hani and Pereira (1999) and El-Hani and Emmeche (2000).

organizing wholes, impose constraints on the way physical laws are applied. Thus, the behaviour of molecules in a cell, which is a physical matter that obeys physical laws, cannot be fully explained if we do not take into account that such behaviour is constrained by the molecules' being elements of the cell. The same account can be applied to individuals and communities. Suppose you want to account for the behaviour of the members of a soccer team. In principle, players do what they do because they have some causal powers of their own. However, we cannot fully explain their behaviour if we forget that they form a team which has its own dynamics. This dynamics imposes constraints on what players do, such that it is not possible to fully explain their behaviour by focusing solely on them as individuals. Roughly: their causal powers are activated or inhibited depending on the dynamics of the team.

Now, the problem is this: a physicalist may agree that constituents of a whole exhibit peculiar behaviour in virtue of their being constituents of such a whole. However, she may add that the constraints that emergentists talk about are themselves explainable in physical terms[9]. The cause of the peculiar behaviour of molecules within the cell is due only to the interplay of the causal powers of the molecules and of whatever other components of the cell, or extracellular elements, have to be taken into account. Which physical laws rule the behaviour of the constituents is not a matter of high-level constraints: the physical laws that are applied are the physical laws that rule the behaviour of those constituents when they are all together in that kind of situation. Similarly, the cause of the peculiar behaviour of the players in the team is solely due to the interaction of the individuals that compose it: the activation of the causal powers of one individual may inhibit the activation of the causal powers of another (which would be active in another scenario) but this does not mean that the possessors of causal powers are any other than the individuals themselves. In conclusion, certain patterns emerge in so-called self-organizing self-preserving wholes, but they have low-level causes: there is no formal, or indeed any other kind of downward, causation.

Before moving on to discuss this issue under different premises, I want to consider another possible solution to the exclusion problem (a-d). What I have proposed above is that the functional explanation explains the physical explanation, so that the two kinds

[9] See Davies (2006) for further consideration of this position.

of explanation are not independent. Another way to interpret what I have suggested is that I was arguing for a dual-*explanandum* strategy. This strategy, when applied to the exclusion problem for mental causation, says that there is no conflict between mental and physical causes because physical causes bring about bodily movements whereas mental causes cause actions that are not reducible to bodily movements[10]. The answer to the present problem would be: the proper *explanandum* for the functional explanation is not the heart's being there, as claimed in (c), but the whole physical process whose end point is the heart's being there. As the physical process in question comes to an end with the heart's being there, we also explain why the heart is there, but we do so derivatively.

Now, it is possible to take a different dual-*explanandum* approach. We may say that what we explain with the functional explanation is not this heart's being there, but the existence of the kind: *heart*. That is, what we explain when we invoke the property of being a blood pump is why a certain kind of thing exists at all; and as we explain why a certain kind of thing exists as such we also explain, though this time derivatively, why an instance of such a kind "is there". That is, we explain why this heart exists because this heart belongs to a class, or a kind, whose existence as such we can explain.

This response depends on a distinction that not all philosophers would be willing to accept. However, I think it can be argued that there is a distinction between a kind and its extension or the set of individuals that are examples of such a kind. Now, if this distinction is conceded, then it is possible to say that a physical explanation is insufficient to account for the existence of, e.g., the kind: *heart*. A physical explanation may well explain why all the individuals that belong to such a kind exist, but it cannot account for why the kind as such exists. The sort of explanation that accounts for the existence of kinds is the explanation that mentions the functional properties of hearts. If we ask ourselves: "why does a kind such as *heart* exist in the world?" we are forced to answer "because its instances have certain functional properties, in particular, because they pump blood". That there is this sort of explanation becomes apparent when we think of, "why this instead of that" questions. Why is there the *heart* kind instead of the kind: *x*? It is not satisfactory to say that it is because you know that the story of the world is such that hearts have survived and *x*s have gone out of existence. Rather, the

[10] See Hornsby (1997) and Marras (1998). For a discussion of this view see author (2004b).

answer is that instances of the *heart* kind were replicated because the kind as such (hearts as such) was adaptive, and that it was adaptive because hearts are blood pumps. In contrast, *x*s as such were not adaptive because, for instance, they had no dispositions that were beneficial to the organism. So, even if the plurality of hearts exist, or have existed, because they are the end points of different chains of physical causes and effects, the kind of thing exists because it is an adaptive kind.

Now, whatever the plausibility of this distinction between explaining the existence of a kind as such and explaining the existence of its instances, the account does not work. The reason is that we want the model to cover all cases where a disposition is converted into a teleological function and this kind of response would only cover the case of t-functions generated by natural selection. To see this, let us go back to Dretske's theory of mental causation. Dretske argued that a neural structure, N, can be recruited to indicate F (so that it acquires the t-function of indicating F) in virtue of its indicating F. That is, N acquires the function of firing when Fs are around because it has fired in that situation in the past. We would then say that the disposition to fire when F has been converted into the t-function of firing when F. Now Dretske claims that the main recruitment engine is not natural selection but learning, so that N does not have any function until a learning process recruits it for firing when F. So N fires when F because N itself, i.e., this particular group of neurons, fired when F. There is nothing here going on at the level of kinds: the selection –or recruitment– process proceeds at a nominalistic level, so to speak.

Robert van Gulick (1993) developed an account of mental causation that resembled Dretske's. The main difference was that while in Dretske's account activated neurons had backward-looking dispositions (indicating F, i.e., being caused by F), in Van Gulick's they had forward-looking dispositions (causing this or that). That is, according to van Gulick, it is the forward-looking dispositions of neural structures that make them cause what they cause. In his view, a brain, being a dynamic self-preserving and self-reproducing organ, recruits physical structures according to their causal capacities, and because they have such causal capacities, and then *selectively activates* them. Take an injured brain, says van Gulick: if the neurons that have been injured were previously displaying an important function for the organism, the brain will *recruit*

some other neurons to take on, if possible, the causal role that was temporarily lost[11].

This account, then, includes the fact that brain structures have properties of the sort *having the function of having such-and-such a causal role* (i.e., they have a mental property iff they have been recruited to play a certain causal role). According to Wright's analysis this means: (i) brain structures fulfil such-and-such a causal role; (ii) they are "there" (instantiated or active, in this case) because they can fulfil that role. In a nutshell, there has been a process of selection (here recruitment) in which the functional property acted as a causal factor. Now, what is important at this point is to look at the differences between this kind of process and selection processes in evolution. The main difference is that whereas it can be argued that evolution recruits kinds (not this heart or that heart, but *heart* as a kind –this kind of thing) in an injured brain it is *these* neurons, or *this* brain structure, that is recruited. This means that the functional explanation explains exactly what a physical explanation would explain, namely why these neurons fire now. Physics provides two kinds of explanation for this fact: first, these neurons fire because they have been activated; second they fire because there was a certain process (described in non-physicalist terms as a recruitment process) which resulted in a connection between these neurons and the neurons they activate. The second explanation competes with the functional explanation, and it is not possible to avoid the clash by saying that the functional explanation (neurons fire because they had the disposition to fire) has a different *explanandum*.

3. Explanation and Causation

From the preceding, we have seen that there is no room for the causal efficacy of dispositional properties: they cannot cause a physical effect; they cannot cause the physical process that ends in such an effect; and they cannot cause anything that is not physical. In spite of all this, from now on I want to argue that explanations couched in dispositional terms are irreducible, necessary *and causal*. An explanation is causal, according to many authors, iff it informs about the cause of the *explanandum*. This can be read either in a liberal way, such that an explanation is causal as long as it contains some information or other about the cause of the *explanandum*, or in a strict way, so that for an explanation to be a causal explanation it must mention the cause of the

[11] The brain is constantly recruiting neurons, making them grow new tendrils in order to establish some new connections and thus realize new functions. It is said that the brain can be rewired in this way in about twenty minutes (see Bennett and Barden, 2001).

explanandum. Under the first reading, a causal explanation can in principle be replaced by another, more informative explanation, with the most informative being the one that mentions the cause directly. Under the second reading, a causal explanation cannot be replaced by another that is more informative. Thus, we have a kind of criterion here: if we have a good, true, explanation which informs about the cause of the effect and cannot be replaced by another that is more informative, then, *prima facie*, such an explanation is mentioning the cause of the *explanandum*. As I want to say that functional explanations are not replaceable by more informative explanations, we arrive at a contradiction: dispositional properties should be causes, but they cannot be (as indicated by all that has been said above in relation to exclusion problems). Now we could be content with the antinomy but it is preferable to resolve it in one way or another. My proposal is to give credit to our explanatory demands.

In what follows, I want first to show that the stated, intuitive, criterion for causal efficacy works (to repeat: that if we have an explanation which (a) informs about the cause of the *explanandum* –i.e., it is a causal explanation – and (b) cannot be replaced by a more informative one, then that explanation must be mentioning the cause of the effect to be explained[12]). This will make me revise what other philosophers, notably Frank Jackson and Phillip Pettit, have to say about it. Then, I will move back to remind us what kind of explanations functional explanations of the sort considered here are, and why they cannot be replaced by a physicalist explanation. This re-states the problem of exclusion we have been struggling with. However, the pressure to find a compatibilist solution to this problem grows once we grant that functional explanations are causal explanations; that is, once we cannot simply exclude functional causes or claim that the work of functional properties is screened off by the work of physical/categorical properties. Thus, in the last part of this section I will argue for a compatibilist solution to the exclusion problems developed here.

The criterion I have proposed for causal efficacy states that a certain event is causal if and only if it is mentioned in a causal explanation which cannot be replaced by another without a loss of information. Many philosophers have made use of an

[12] "Causal explanation" is used here in its liberal sense, meaning an explanation which provides information about the cause. This use entails that there can be causal explanations that do not pick out causes directly. In turn, this entails that the statement of the criterion: "a certain event is causal iff it is mentioned in a causal explanation which cannot be replaced by another without a loss of information," is not circular. It would only be circular if the intended meaning of "causal explanation" were the strict one, according to which an explanation is causal only if it cites the cause directly.

explanatory criterion (to little avail) in order to argue for the causal efficacy of higher-level properties (see, notably, Fodor 1990). So the first task to be undertaken is to show that the explanatory criterion proposed here is different from the one used by such philosophers. In particular, I will try to show how their criterion and the criterion I propose identify different things as being causally efficacious. Then I will argue that the results of using my new criterion match our intuitions.

Fodor (1990) argued for an explanation-driven identification of causes (in particular for an explanation-driven identification of higher-level causes) along the following lines: special sciences have laws that are perfectly predictive and explanatory. The best explanation that we have for the predictive success of a theory is its being true; but that a theory is true means that the entities it postulates exist and that the causes it invokes are the causes of the events it explains. It follows then that the explanations of the special sciences identify the causes of their *explananda* directly, that is, that when we explain, e.g., a certain piece of behaviour in terms of a belief-desire pair, we are providing the causes of such a piece of behaviour. If we further take it that the predicates of the special sciences pick out functional properties, this means that functional properties are causally efficacious regardless of what exclusion problems tell us.

Fodor's criterion can be summed up as the claim that a law is strictly causal (i.e., it cites causes directly) insofar as it is predictive and explanatory. However, this account is not right: a law can be predictive and explanatory even though it cites causes only indirectly. This was first shown by Block (1990) and also made clear by Jackson and Pettit (1990). According to those authors, the explanatory and predictive success of a law or theory (or a family of either) are compatible to their providing only incomplete information about the causes of their *explananda*. That is, it is not true that the only explanation for the explanatory success of a theory is its identifying causes directly.

Block, for instance, explained how a functional explanation can be seen as a partially informative explanation that is completed by a categorical explanation. According to him, the use of a functional predicate brings with it the information that one of its categorical bases brought about the effects to be explained, thus excluding the intervention of any other property. This view of functional explanations can reconcile the putative fact of the causal inertness of dispositions with their illuminative use in explanations: the functional explanation can be considered as a pointer to the

categorical explanation, i.e., the one that cites directly the cause of the effect to be explained. An explanation such as "the glass broke because it was fragile" is true and causal under the first, liberal, reading described above: it informs about the cause of the breaking of the glass. However, it does not pick out the cause of the breaking directly. Rather, what an explanation such as this says is that the glass broke because it instantiated one of the realizers of the functional property *fragility*.

Hence, Block's account of functional explanations rebuts Fodor's explanation-driven identification of causes: not all successful explanations are causal on the strict reading of what it is for an explanation to be causal. However, it does not rebut the criterion for identifying causes proposed here. For, according to that model, functional causal explanations can always be replaced by more informative explanations, namely by those which cite the cause of the *explanandum* directly, i.e., by categorical explanations.

Thus, the explanatory criterion for identifying causes that I have proposed is different from more liberal criteria such as Fodor's. Moreover, it can be said that some of the arguments that count against the latter seem to give support to the former, given that it delivers the right results, so to speak (namely, that the usual functional explanations do not mention causes). However, the proposal needs more argumentation. It has been said that Jackson and Pettit (1990) also developed a rebuttal of Fodor's argument for the causal efficacy of higher-order properties. Theirs resembles Block's except in one point: Jackson and Pettit explicitly claim that functional explanations provide some information that is lost when they are substituted by categorical explanations (which, on the other hand, they assume are the only *real* causal explanations). Now, if they were right, I would have to say that functional explanations are causal, since functional explanations cannot be replaced by any other explanations without a loss of information. This would be highly problematic, for what I wish to argue is that *only some* functional explanations (those which explain selection processes) are causal. So let me discuss Jackson and Pettit's account in order to reinforce the plausibility of the explanatory criterion for the identification of causes.

Jackson and Pettit (1990) put forward a model of high-level explanation called "program explanation". The general idea is that a high-level property can "program" for the presence of an efficacious low-level property in the same way that a computer

program is relevant to a change in a computer screen, namely, by ensuring the occurrence of some mechanical process within the computer. In virtue of such programming, high-level explanations acquire causal relevance (though not efficacy). For instance, the water's being at boiling temperature is causally relevant for the breaking of its container because the water's boiling temperature programs for the occurrence of a low-level process, e.g., a molecule having a certain momentum, which brings about the effect. This notion of programming looks ontologically loaded, as if Jackson and Pettit were speaking about a specific kind of causation whose *relata* are high-level events on the one hand and low-level processes on the other. However, their use of the notion is epistemic, and what they argue is that explanations of the sort just mentioned provide information that "someone in possession of the process account may lack" (J&P, 1990: 130). A program explanation tells us not only that a certain low-level process brought about the effect, but also that if things had been somewhat different, the effect would have taken place all the same. Mentioning that the water was at boiling temperature provides the information that a certain molecule or group of molecules cracked the container *and* that the container would have broken all the same even if those molecules had not cracked it (because the boiling temperature ensures that there are molecules with the momentum required to crack the container). An explanation couched in a low-level vocabulary cannot provide that kind of information, so we are led to conclude that the higher-level explanation cannot be replaced by any other.

Thus, even God would need to know that the water was at boiling temperature in order to know that the container would have cracked even in those worlds where none of the molecules that actually did the cracking have enough momentum to do it. God can predict that if the molecules that broke the container had not broken it, then some others would have, provided the action of the former molecules is pre-empting the action of the latter. However, having the information that the water was at boiling temperature ensures more than this: it ensures that the container would have cracked even if such molecules were not pre-empting the action of others. It tells you that the average momentum is high enough; thus it tells you that if the molecules that had a high momentum in this world had not had it, then the ones that had a low momentum would have had such a high momentum.

Now, this can be debated. The water in the container receives an input of energy, and the energy received is such that we know that some molecules will increase their

kinetic energy in such a way that they will strike some molecular bonds of the container. In the actual world, molecule *a* increases its kinetic energy in the way described and hits a molecular bond at place *e*. If we change the initial conditions slightly, then it is molecule *b* that strikes a molecular bond at place *e'*. So it is true that someone in possession of the process account only (the container broke because molecule *a* hit a molecular bond at *e*) lacks some information. However, the information she lacks is why such a molecule struck the container. It might have happened randomly or, as was the case, because the water's energy was increased. Once she is informed that the hitting took place for the second reason, she also knows how the counterfactuals go and, in particular, that the container was going to break anyway. So the knowledge that is relevant is not knowledge of a higher-level property which programs for the cause of the breaking, but knowledge of the causal antecedent of such a cause.

Jackson and Pettit use other examples to illustrate their point. For instance, suppose that a lift has been built so that it halts whenever its load weighs more that 65 kg, and that it is loaded with 75 kg and halts. What is the explanation for its halting? That it halted because it was loaded with 75 kg or because it was loaded with more than 65 kg? The first explanation points to the actual cause. However, the second cannot be replaced by the first, for it provides counterfactual information lacking in the first. Again, this interpretation of what goes on is debatable. For one thing, it is not clear that determinates screen off the causal efficacy of determinables[13]: a red traffic light makes you stop regardless of what shade of red the light is; or having mass makes you subject to some physical laws, regardless of what mass you happen to have. Why do we have to concede that the cause of the lift's halting is that it was loaded with 75 kg? I think that that position is not compulsory, and that it can be argued, with at least the same conviction, that the cause of the halting was that the lift was loaded with a body that weighed more than 65 kg. Thus, if you only have the process information you do lack some information. But I would say that what you do not know is the cause.

According to Jackson and Pettit, functional explanations are a species of program explanation. However, they say that "[a] program explanation tells you about how that history [the history as it actually went] might have been" (1990: 130). This is so in the case we have just considered: the lift would have halted in all and only those worlds

[13] See Stephen Yablo's (1992) famous discussion on the determination relation.

where what went in it weighted more than 65 kg. The same can be said about a bull's charging against a red cape. It does not matter what shade of red the cape is: the bull would have charged anyway, provided the cape was red. However, a functional explanation cannot provide that kind of information. The explanation "the glass broke because it was fragile" does not tell you what would have happened if the molecular structure that realizes fragility in this world had been absent. It only tells you that the glass would have broken in all those worlds where some other realizer of *fragility* was present. Note that what this means is that the glass would have broken in all those worlds where another categorical cause of the breaking was present. This is not informative. God would know that much without having the concept of fragility[14].

So I conclude that Jackson and Pettit fail to show that there are non-replaceable causal explanations which do not mention the cause of the *explanandum*. Their examples do not form a unified class; but what is important is that they do not exemplify cases where (a) effect *e* is brought about by cause *c*, while (b) there is a non-dispensable explanation of *e* that does not mention *c* directly. In particular, they do not show that our usual functional explanations are not replaceable by categorical explanations. If you have the categorical explanation, then you know the cause and the causal path followed by the cause. The functional explanation is less informative than that, because it only provides vague information about the cause ("it is one of my realizers") and tells you about the causal path. However, it does not really tell you how things might have been had that cause been absent.

I proposed above that we could adopt the following principle: if an explanation informs about the cause of the *explanandum* and it is not replaceable by any other, then it is a strictly causal explanation. Program explanations looked like counterexamples to such a principle because, allegedly, a program explanation does not pick out the cause

[14] In author (2001) I argue that the determinable/determinate relation is fundamentally different from the functional property/realizer relation. One of the differences has to do with causally relevant information: a determinable concept, say, *RED*, is causally informative in that it may give you an idea of where (i.e., in which worlds) a certain causal relation holds –it holds wherever the causal event instantiates the property of being red, regardless of what shade of red it instantiates. However, a functional concept is informatively vacuous in this respect. The concept of being a sleeping pill is the concept of having a property which causes one to sleep. This means that the explanation "I fell asleep because I took a sleeping pill" only tells you that you would have fallen asleep provided you had ingested a pill which had something that would have caused you to sleep. That is, it only tells you that the worlds where the causal relation holds are the worlds where it holds, and that is it.

of the effect directly, although it contains some modal information that the strictly causal information lacks. I have tried to show that they are not, in fact, counterexamples. So I think it is safe to adopt my principle at least as a default principle; and in particular, I think we can adopt it in with regards to functional explanations.

Now, are functional explanations such as, "the heart is there because it pumps blood" non-replaceable? What kind of information do they give us that cannot be given by non-functional explanations? It may seem that they give us modal information: *ceteris paribus*, the heart would be there no matter what it was made of, as long as it (or rather, its ancestors) had the disposition to pump blood. (Contrast this with: the glass would have broken no matter what it was made of, as long as it was fragile, i.e., as long as it was made of something that could cause its breaking.) This is informative: it tells us that we will find the heart in all those worlds where its ancestors have a certain disposition. God would not know this if he only had categorical information. Ironically, he must know that there are selection processes in the world and that they pick out entities by what they can do. This applies to *heart* as a kind: we can expect to find instances of the *heart* kind in all those worlds where there have been selection processes and blood pumps have been adaptive. It applies differently if we want to explain the existence of a particular heart. In that case, this heart could be slightly different from the way it is, and it would still be there as long as its ancestors were efficient blood pumps[15].

[15] The modal information functional properties provide in these cases can be seen as analogous to the modal information provided by determinables when determinables are causally efficacious. If a pigeon pecks at red things, and we find it pecking at a scarlet circle, we know that it would have pecked at that circle in all the worlds where the circle had been of a different tonality, as long as it had been a tonality of red. That is, the determinable demarcates the sphere of worlds in which we find the effect occurring. Analogously, when we know that a heart is there because it pumps blood, we also know that the heart would have been there even if it had been physically different and even if the physical process that generated it had been different as well (and also, even if the fundamental physics of the world had been different). The analogy breaks down, however, if we take into account that the heart's being a blood pump does not by its own guarantee its existence. For a heart to exist, for instance, it is necessary that it is in a body that does not reject it, that its ancestors belonged to organisms that were, in general, well adapted to the environment, and that they belonged to organisms for which blood pumping was beneficial. That is why we can only say that the heart would be there as long as its ancestors pumped blood, *ceteris paribus*. This means that functional properties cannot demarcate the boundaries of the sphere of worlds in which the effect takes place as neatly as determinables do. In any case, I do not think that this argument from modal information is *the* argument for the causal efficacy of functional properties. The main reason to believe that functional explanations are causal is the one developed in the next paragraph, namely, that they give us a kind of information that we would otherwise miss (that is, regardless of whether this information has, or does not have, the modal implications that I think it has).

So I think there is some modal information one would miss even if one knew all the causal history of the world. However, the important point is this: one would never be in a position to completely explain why certain things endure or do what they do while other things disappear or do some other thing. We might look at the world and be able to track its causal history. We would thus see that some types of things have gone out of existence while others have been replicated and endured. We would also see that some structures start doing things that they did not do before. We might think that there is nothing else to be explained: physical laws just drive the world and its entities in one definite direction. Yet, the fact is that there is something else to be known. We can find a more illuminating explanation of why some things endure and others disappear, and such an explanation is couched in functional terms: adaptive entities –i.e., entities that can perform certain functions– tend to endure. Entities whose causal powers are not of interest to selection or recruitment processes tend to disappear while many other entities do what they do because they were able to.

Summing up: explanations of the sort "the heart is there because it pumps blood" are non-replaceable. This means that they are causal explanations, at least according to the principle proposed above, i.e., that if a good causal explanation is non-replaceable, then it must count as causal. As noted before, this contradicts the conclusion of our discussion of exclusion problems. We could accept the contradiction and leave it at that, in the hope that it can be solved in the future in some way or other. However, it is obviously better to have something positive to say about this antinomy.

In the mental causation debate, some have argued for a compatibilist position (see Horgan, 2001, Bennett, 2003, Burge, 2007). It is likewise possible to try and adopt a compatibilist stance here. In order to argue for such a stance, I propose to let the explanatory facts guide us. Causes are discovered in the process of explaining why certain things happen. That is, our metaphysics follows our epistemology. So, if we have a non-replaceable explanation of an event, we have identified its cause. However, what if such an event is shown to have another explanation, i.e., another cause? Well, insofar as this new explanation is not replaceable by the first one, we should adopt it as well. This is actually what happens with functional explanations of selection processes and their physicalist counterparts: functional explanations cannot be replaced by physicalist explanations without a loss of information, but physicalist explanations cannot be replaced by functional explanations either. Physicalist explanations provide

one kind of information –mechanistic, let us say– that is absent from the explanation couched in functional terms. In a nutshell, the two kinds of explanation must coexist because each throws a different, unique, light on the events to be explained. This being so, we have to conclude that there are two different kinds of causes operating here.

This is only part of what we have to do, for the conclusion amounts to a re-statement of the exclusion problem[16]. The trip has not been in vain, though; exclusion problems are usually seen as contests where one of the participants loses. Given that we cannot have two causes for one event, one of the putative causes must go. This is the idea we have been assuming all along: some functional explanations seem to be strictly causal explanations, but they cannot be, since their putative effects have another causal explanation. So, functional properties must be inert. Now we have to see things in a different way: if the argument from the non-replaceability of functional explanations works, then the exclusion problem cannot be solved by ruling out the causal efficacy of functional properties. That means that we have to look for another solution.

It may seem that if we cannot exclude either of the two participants in the contest, we cannot but conclude that there is causal overdetermination. However, this is not the only possibility. There are two other paths a compatibilist may try to follow; one successful and the other not. First, the compatibilist may deny that there actually is an exclusion problem after all. As noted above, exclusion problems are generated by the principle of causal-explanatory exclusion. In turn, this principle is grounded on our explanatory practices: we do not accept two causal explanations that offer alternative complete explanations for the same event. The metaphysical principle which claims that there cannot be two complete sufficient causes for one event (except in cases of overdetermination) is derived from the explanatory principle. Now, given that, it seems that we can still accept that there are two complete sufficient causes for one event, *e*, in some cases, namely, in those where we are prone to accept two alternative explanations for *e*. Going back to our case: functional explanations explain why a certain entity is there in a way physical explanations do not and cannot. Do we feel that we have to exclude one of these two alternative explanations? Well, leaving feelings to one side, we can see that the possible exclusion problem is certainly not as pressing as it is in the mental causation debate, for instance. So, suppose we are convinced that we lack the

[16] I thank an anonymous referee for pressing this point.

epistemic pressure required to generate an exclusion problem between functional and physical causes. Then, it seems, we could keep both.

This path is unsuccessful, however, because claiming that both functional and physical properties are causes means accepting that they are overdetermining their effects. This result should not be surprising, given that the principle of causal-explanatory exclusion can be read as a definition of causal overdetermination. In effect, the principle simply tells us what causal overdetermination consists of, namely, having two independent complete causes that do not exclude each other[17]. Thus, if we concede that we have two complete and independent causes and we insist that they are both causes, we are thereby claiming that we are facing a case of causal overdetermination.

The second path towards a compatibilist response is more promising. It amounts to denying that the explanations are indeed alternative complete explanations. If I am right, the functional explanation provides a kind of information that the physical explanation does not provide. The reverse is also true: the physical, mechanistic explanation gives us a kind of information that the functional explanation cannot. Thus, it can be said that functional and physical explanations are complementary –and not alternative– explanations, and that the one completes the other. The principle of explanatory exclusion tells us to exclude one of two explanations when these are both complete; by denying that functional and physical explanations are both complete explanations, the strictures of that principle no longer apply. This, I suggest, is the best way to avoid having to choose between functional and physical causes, without eventually saying that they overdetermine their effects.

Does this mean that the CCP is false, after all? That depends on how we state it. If the CCP is construed as the claim that every physical effect has a *complete* physical cause, then my position contradicts it. However, the CCP can be construed in a weaker way; as the claim that every physical effect has a physical cause. This construal of the CCP allows for the existence of complementary causes. If, once again, we construe our metaphysics in the light of our epistemology, this weak reading of the CCP looks more plausible, for if my proposal makes sense, there are physical explanations that are not

[17] The principle of causal-explanatory exclusion says: there cannot be two independent and complete causes/causal explanations for a given event, except in cases of overdetermination. What I say is that, when read right to left, this is a definition of overdetermination: there is overdetermination if and only if there are two complete and independent causes for a given event.

complete[18]. Furthermore, I do not think that the rejection of the strong reading of the CCP is far-fetched. It is difficult to show any variant of the CCP to be true, at least when "the physical" is taken to be characterized by (current or future) physics itself, and not simply defined as "the non-mental" (see Montero, 2003, author, 2006, forth.). So, in principle, the task is even more complicated if the variant of the CCP we want to argue for is the more ambitious version.

As an anonymous referee points out, this proposal is still a bit too programmatic. One would surely want to have a more definite idea as to how basic physical causes and functional causes complement each other. From what has been said, the basic physical explanation and the functional one seem to be just parts of a complete causal explanation. This would mean that the two kinds of cause are parts of a complete cause; but what exactly does that mean? It cannot be that they are parts of a cause in the sense that a body *a*'s weighing 10 kg is only a part of the cause of two bodies', *a* + *b*, weighing 20 kg. So, in what sense are basic physical and functional causes parts of a complete cause? I am afraid I have not worked out an answer to this question. I have tried to convince the reader that functional and physical causes complement each other. However, going forward implies having some idea about how causes of apparently different sorts may work together, and, as I say, I lack an adequate answer to that problem.

In the context of the mental causation debate, Tyler Burge (2007) explores a position which may bear some similarities to the one advanced here, though his proposal is much more programmatic. There (pp. 380-382) he claims that perhaps we should clarify the notion of mental causation in order to explain how mental and physical causes are not in competition, but work in concert and non-coincidentally[19]. However, he claims, "we do not have a satisfying understanding of mental causation, or indeed any causation" (2007: 381). There is a methodological point here that I want to subscribe to. Burge recognizes that one of the tasks for our theories of causality is to explain how mental and physical causation are compatible. That is, he proposes to take mental causation as a fact, and to construe our theories of causality (and of mentality) in

[18] We can even concede that every physical effect has a sufficient physical cause, as long as sufficiency does not entail completeness. Such a link from sufficiency to completeness can be broken if we take sufficiency to be context-dependent. An explanation would be sufficient insofar as it is sufficient in one context or another. Thus, the CCP can be read as the claim that for every physical explanation there is a context where that explanation is sufficient. This construal of the CCP allows for the existence of complementary causes, for it permits there to be contexts where the physical explanation is not sufficient.

[19] He adds (p. 381): "The two types of causation clearly operate together in some systematic way. What that way is remains to be understood".

view of that fact. The general point, then, is that perhaps our notions of causality are ill-developed, at least in the sense that they do not help us to explain some things we would like to have explanations of. According to what I have been saying, one of these things is that functional and physical causes complement each other (in the cases I have dealt with). This, I have tried to argue, is a fact. Thus, my proposal, if right, could be taken to be an *explanandum* for a theory of causality. Unfortunately, I do not have a theory of causality that can account for it.

Now, let me close by taking a very brief look at the problem of mental causation, for it may seem that the picture I have been developing could shed some new light on it. My own view is pessimistic: there is no new insight on offer. If mental properties are functional properties, and mental properties compete with their realizers for the same effects, then there is no hope that they can be causes. Dretske (1988) argued that mental properties are structuring causes: what they explain is why some neurons fire when they fire, and not directly why this body moves the way it does. This would make mental properties causal, at least according to the argument I have developed here. However, Dretske's theory was heavily criticized in the nineties. One of the most telling criticisms was that according to Dretske's view, mental properties were merely past-operative causes while we want them to be here-and-now causes of behaviour (see Horgan, 1991). Now, if this is so, that is, if we are really and deeply convinced that the mind's causal work consists of triggering behaviour, then mental causes cannot but compete with physical causes; and probably lose the contest. The compatibilism I have argued for cannot be exported to the mental causation debate (despite what Horgan, 2001, or Bennett, 2003, hold). For, what seems to be the case is that mental explanations *can* be replaced by physicalist explanations: that is at least the conclusion to be extracted from the discussion about functional explanations of the sort "the glass broke because it was fragile" in this section. We had better not want mental properties to be functional properties and at the same time to trigger behaviour: we will end up suffocating them[20].

References

[20] Burge, in the aforementioned paper, seems to coincide with this diagnosis, given that he denies that mental properties are functional properties. Considering all that has been said, perhaps that is a good starting point if you want to defend a compatibilist solution to the mental causation problem. However, to examine Burge's proposal here would take me beyond the scope of the current paper, given that its topic is not mental causation but the causal efficacy of functional properties.

Bird, A. (2007) "Causal Exclusion and Evolved Emergent Properties", in *Revitalizing Causality: Realism about Causality in Philosophy and Social Science* (ed. Ruth Groff), Routledge.

Bennett, K. (2003) "Why the exclusion problem seems intractable, and how, just maybe, to tract it", *Noûs* 37: 471-497.

Bennett, M.R. and Barden, J.A. (2001) "Ionotropic (P2X) receptor dynamics at single autonomic varicosities", *NeuroReport* 12: A91-A97.

Burge, T. (2007) "Postscript to Mind-Body Causation", in *Foundations of Mind, Philosophical Essays, vol. 2*. Oxford: Oxford University Press.

Crane, T. and Mellor, D.H. (1990) "There Is No Question of Physicalism", Mind 99: 185-206.

Crane, T. (1995) "Mental Causation", *Proceedings of the Aristotelian Society*, 69: 211-36.

Davies, P. (2006) "The Physics of Downward Causation", in *The Re-Emergence of Emergence*, Davies,P, and Clayton, P. eds., Oxford: Oxford University Press.

Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dretske, F. (1993) "Mental Events as Structuring Causes of Behavior", in Heil and Mele (eds.) *Mental Causation*, Oxford: Clarendon Press.

El-Hani, C. N., and Pereira, A. M. (1999) "Understanding Biological Causation", in *Where biology meets psychology*, edited by Valerie G. Hardcastle. Cambridge: MIT Press.

El-Hani, C. N., and Emmeche, C. (2000) "On some theoretical grounds for an organism centered biology: Property emergence, supervenience and downward causation". *Theory in Biosciences* 119: 234–75.

Fodor, J. A. (1990) "Making Mind Matter More", in *A Theory of Content and Other Essays*. Cambridge: MIT Press

Hempel, C.G. (1980): "Comments on Goodman's Ways of Worldmaking", *Synthese* 45: 193-200.

Horgan, T. (1991) "Actions, Reasons, and the Explanatory Role of Content", in McLaughlin, B. (1991) *Dretske and his Critics*. Cambridge, MA.: MIT Press.

Horgan, T. (2001) "Causal Compatibilism and the Exclusion Problem", *Theoria*, 40: 95-116.

Hornsby, J. (1997) *Simple Mindedness: In Defense of a Naive Naturalism in the*

Philosophy of Mind. Cambridge, MA: Harvard University Press.

Jackson, F. and Pettit, P. (1990) "Program Explanation: A General Perspective", *Analysis* 50: 107-117 (reprinted in Jackson, Pettit and Smith, *Mind, Morality and Explanation*, Oxford University Press, 2004).

Kim, J. (1993) *Supervenience and Mind*. Cambridge: Cambridge University Press.

Marras, A. (1998) "Kim's Principle of Explanatory Exclusion'", *Australasian Journal of Philosophy* 76: 439-451.

Menzies, P. and List, C. (2009) "The Causal Autonomy of the Special Sciences", in *Emergence and causation*, Oxford University Press.

Millikan, R. (1993) *White Queen Psychology and Other Essays for Alice*. Cambridge: Cambridge University Press.

Montero, B. (1999) "The Body Problem", *Nous* 33: 183-200.

Montero, B. (2003) "Varieties of Causal Closure", in Walter, S. and Heckmann (eds.) *Physicalism and Mental Causation*, Imprint Academics.

Prior, E.W., Pargetter, R. and Jackson, F. (1982) "Three Theses about Dispositions", *American Philosophical Quarterly*, 19: 251-7.

Van Gulick, R. (1993) "Who's in Charge Here? And Who's Doing all the Work?" in (eds.) *Mental Causation*, Oxford: Clarendon Press.

Walter, S. (2007) "Determinables, Determinates, and Causal Relevance", *Canadian Journal of Philosophy* 37: 217-43.

Wright, L. (1973) "Functions" *The Philosophical Review*, 82: 139-168.