As I see bus no. 29 approaching, I raise my arm. The bus stops, I take a few steps and get on it. This happens because the driver, having seen my arm raised, interpreted the gesture as a conventional expression of my wish to get on the bus. If it had been bus no. 17, I would not have raised my arm because I know that that bus follows quite a different route. This is not, of course, the end of the story. I might also mention the reasons why I really wanted route 29, and so on. Fortunately, there is no need to go into the details. The important point is that in the explanation of this trivial set of actions I can quite naturally appeal to a number of mental states: perceptions, beliefs, interpretations, desires, etc. The previous story assumes, for instance, that

> my seeing bus no. 29 approaching (causally) explains my arm raising,

and also that

> the driver's perception of my raised arm (causally) explains her stopping the bus.

In general, examples like this invite the following principle:

> *The causal efficacy of the mental:* The contents of an agent's mental states causally explain her actions.

At first sight, it is hard to think how this principle could be false. For we conceive of ourselves as agents whose actions are guided by the content of our beliefs, perceptions, commitments, desires, etc., and have no idea of how it could be otherwise. There is some reason to claim, however, that the causal efficacy of the mental clashes with some deep physicalist convictions. The problem is that, if the conflict turned out to be really inevitable, some people would surely tend to stick to the latter (allegedly grounded on the explanatory success of natural sciences), and discredit mental causation as illusory, as mere appearance. But what are, roughly speaking, those physicalist convictions?

Agents are part of nature. After all, we act, think, and perceive because we have a body. It is certainly thanks to my eyes that I can see bus no. 29 approaching and, when I subsequently raise my arm, this is surely the result of some nerve impulses. My arm raising appears then as a physical event that can be causally explained by a number of antecedent physical features of my body. In other words, it sounds quite obvious that

> there are a set of physical features of my body that causally explain my arm raising

and, similarly, that

> there are a set of physical features of the driver's body that causally explain her stopping the bus.

The plausibility of these statements is associated with an overall physicalist conviction, namely:

The causal closure of the physical world: every physical event has a complete causal explanation in physical terms.

It follows that, insofar as our actions are physical events, a physical explanation must underlie any mental or intentional explanation that we may provide for our actions. Yet, the combination of the two metaphysical principles so far introduced, implies that actions will have two different causal explanations, namely: a mental and a physical one. The problem is that this proliferation of causes may conflict with some fundamental intuitions about the individuation of causal processes and, relatedly, about the nature of causal explanation. In fact, this volume will focus on one such intuition, which Jaegwon Kim describes as follows:

Causal exclusion principle: "(...) Two or more complete and independent [causal] explanations of the same event or phenomenon cannot coexist." (Kim 1989, p. 89).

We often mention partial explanations of an event. Thus, I explained my arm raising as a result of my seeing bus no. 29 approaching, but I could also have explained it by appeal to my desire to go up to a certain point in route 29. We would thereby have two causal explanations of the same event, although this duality causes no trouble precisely because they are just two partial explanations of the event at stake.

There are, indeed, cases where a single event has two or more complete and independent explanations, like when ten darts simultaneously hit a balloon causing its explosion, or when my launching a stone against a window pane coincides with an earthquake. The effect is, in such cases, overdetermined, since each of the darts alone would have caused the balloon explosion, and my launching a stone or the earthquake alone would also have guaranteed the shattering of the glass. But we tend to regard situations like these as exceptional, as coincidences. We are, in other words, reluctant to individuate causes in such a way that this kind of coincidence would abound, that overdetermination were a massive phenomenon. But if we do not want overdetermination to proliferate, it seems that we must accept Kim's causal exclusion principle and, therefore, assume that, in general, whenever we are confronted with two causes of a single event, they are either partial or interdependent. What are, though, the implications of this upshot for the causal efficacy of mental contents?

It follows from the causal closure of the physical world, that each action that has a mental explanation must also have a physical one. But, if one assumes

that massive overdetermination ought to be averted, one cannot allow that each case of mental causation could be regarded as a case of overdetermination. Mental causes must hence be either partial or interdependent. Trivially, mental causes cannot be partial in the relevant sense: they cannot be an ingredient of the physical complete cause, since the latter is perforce exclusively physical. Hence, the only option is to regard the mental explanation as dependent upon the physical one.[1] But what sort of dependence?

Type-type identity comes up as the closest link. From this viewpoint, each type of mental state ought to be identical to a certain type of physical state. This approach, though, is excessively parochial, since we do not want to deny that there could be intelligent agents with quite different physical settings, that is, agents whose psychological capacities and states would rest on physical properties quite disparate from each other. We can thus expect that, for instance, pain would depend on different types of physical states in human beings and in some other intelligent agents, including individuals of some other biological species. It seems then that we must allow for distinct instances of the same type of mental state to depend upon different kinds of physical states. This is why we should look for a type-type relation weaker than identity in order to specify the way mental and physical properties relate.

At this stage, strong supervenience emerges as a promising alternative, since it is consistent with the plurality of physical implementations of mental properties and seems to impose a reasonable modal constraint upon the correlation between mental and physical properties:

> Mental properties *supervene* on physical properties, in that necessarily, for any mental property *M*, if anything has *M* at time *t*, there exists a physical base (or subvenient) property *P* such that it has *P* at *t*, and necessarily anything that has *P* at a time has *M* at that time (Kim 1998, p. 9).

There are, however, several reasons to think that strong supervenience is, despite its name, too weak.[2] Firstly, supervenience is not a constitutively asymmetric dependence. The fact that properties of type B strongly supervene on properties of type A, does not rule out that properties of type A would also strongly supervene on properties of type B. Secondly, strong supervenience just picks up a covariance between types of properties, but does not ensure the existence of a dependence relation between them. The covariance between A-properties and B-properties could be the outcome of their common dependence upon properties of a third class. This is, indeed, a manifestation of a more general concern: strong supervenience leaves us in the dark as to what are the grounds for the covariance between A-properties and B-properties.

In any case, even if strong supervenience does not successfully apprehends the sort of dependence that some physicalist intuitions call for, it represents at

least a minimal physicalist commitment. The worry is that this minimal demand already threats the causal efficacy of the mental. For, insofar as each action that has a mental cause must also have a physical one, we may derive the impression that the physical cause alone is the one that is actually doing the causal work while mental contents are after all idle or superfluous. One might come to think, in other words, that the causal efficacy of the physical excludes or undermines the causal relevance of mental contents. This is precisely the problem of causal/explanatory exclusion which the five essays in this volume intend to investigate. Sabatés provides, in this respect, a careful characterization of the different problems arisen around exclusion, as well as a chart of the available answers. In particular, Yablo and Horgan seek to solve the crucial metaphysical problem by outlining two significantly different compatibilist proposals, while Pineda and Vicente raise a number of serious objections against the most promising compatibilist strategies. But, before going into details, let me elaborate the problem a bit further.

As we have seen, strong supervenience does not ensure the existence of a genuine metaphysical dependence between properties of type A and B, but a covariance between them. This concern only makes sense if one assumes that there are metaphysical relations that are deeper than mere covariance. On this assumption, it sounds reasonable to claim that causation is at least one of those deeper metaphysical links, i.e., that causation does not reduce to covariance. Consider, for instance, the correlation existing between parents having dark eyes and their offspring having dark eyes too: nobody thinks that the former causes the latter. The uniformity in the colour of their eyes is instead explained by the existence of a common factor, namely: some transmitted genes. We thus regard the relation between genes and dark eyes as causal while the correlation between the eye colour in parents and offspring appear as metaphysically shallower.

In the light of this distinction, we would prima facie assert that the link between my seeing bus no. 29 coming and my arm raising is causal and not a mere covariance. But can we consistently sustain this claim? Schematically, we have seen that, for each action A which has a mental cause M, there must be a physical property P such that:

(a) M causes A, and
(b) P causes A

such that M strongly supervenes upon P. The causal exclusion principle suggests that, in general, causes must be counterfactually necessary for their effects. Hence, (b) generally involves the truth of the following counterfactual:

(C/b) If P had not taken place, A would not have occurred either.

But what happens with (a)? Does it generally involve the truth of the corresponding counterfactual (C/a)?

(C/a) If M had not taken place, A would not have occurred either.

We could say that it does because, *ex hypothesi*, M strongly supervenes upon P and, hence,

(C/c) If M had not taken place, P would not occurred either.

And, indeed, the combination of (C/b) and (C/c) certainly imply (C/a). It is clear, however, that the truth of (C/a) depends on the truth of (C/b), but not the other way round. This intimates that (C/b) apprehends a causal counterfactual, a deep metaphysical relation, while (C/a) just picks up a covariance to be explained by means of some other metaphysical link. Like in the dark eyes case, the intermediate place that (C/b) occupies between (C/a and (C/c) suggests that P is the common source of both M and A, whereby the link between M and A is not genuinely causal, but a case of covariance to be explained by the combination of the links expressed by (C/b) and (C/c). In other words, the causal efficacy of the physical seems to exclude the causal relevance of the mental. Mental contents should then be regarded by the physicalist as causally idle, as mere epiphenomena, unless a solution is worked out for the problem of causal exclusion.

We are thus confronted with a serious metaphysical problem which, according to Marcelo Sabatés, constitutes the hard problem of exclusion. For, as he points out, those who apparently advocate for an explanatory (and, therefore, epistemical) principle of exclusion infallibly appeal to causal considerations as they try to motivate it. Sabatés argues, in any case, that no principle of exclusion is actually required in order to generate the exclusion problem. The causal closure of the physical world by itself suffices to rule out the possibility of regarding mental causation as a putative case of overdetermination. It follows from that physicalist principle that the mental-to-physical causal chain cannot be regarded as independent of the physical-to-physical one. For, otherwise, the former could take place without the latter and, thereby, we could have a physical effect without a physical cause and this trivially conflicts with the physicalist principle at stake.

Sabatés also insists on distinguishing the exclusion problem itself from its unwelcome possible consequences, such as: mental irrealism, explanatory irrelevance, and the inefficacy of functional properties. The paper closes with a map of answers which are mainly divided into compatibilist and incompatibilist strategies. As Sabatés suggests, compatibilist proposals are prima facie

more attractive but often fail to really solve, and even address, the hard problem. The present volume focuses on a number of interesting compatibilist approaches, which are both proposed and challenged.

In general, those compatibilist strategies seek some sort of relation, other than strong supervenience, to characterize the way the mental is supposed to depend upon the physical. Such a relation should hopefully overcome the weaknesses that we have detected in the notion of strong supervenience: they must pick up a constitutively asymmetric relation, they must go beyond covariance, they must satisfy certain explanatory constraints and, finally, they must handle the problem of causal/explanatory exclusion. The notion of realization and the relation between a determinable and its determinates come up as the two main alternative proposals. The notion of realization treats mental properties as second-order properties:

> Having a mental state M is having a physical property P that satisfies a certain causal role R.

On this view, distinct instances of M may be realized by different sets of physical properties. The physicalist intuition is, however, that the causal powers of each instance of M are determined by the causal powers of the corresponding realizer Pi. One may then have the impression that no specific causal task is reserved for M, that the causal work is, on each occasion, done by the corresponding realizer Pi. In fact, since M is multiply realizable, its different realizers P(1...n) constitute a disjunction of wildly heterogenous causal powers and, therefore, we can hardly make sense of the idea that all instances of M have some causal powers in common which would explain the production of the corresponding effect.

David Pineda's paper focuses on the notion of realization and explores the possibility of retaining the causal efficacy of the mental by weakening the previous definition of a second-order property. He is thus reluctant to identify the instantiation of a functional property with the instantiation of one of its realizers, and proposes to regard the instantiation of the latter just as a condition for the instantiation of the former. This scheme surely opens the door to the possibility of picking up the causal powers of an instantiation of a functional property in a way that does not reduce them to those of its realizer. But how should such autonomous causal powers be identified? Pineda stresses that any attempt to answer this question by appeal to a causal role falls into a rather serious problem, namely, the existence of a conceptual link between a functional property and the properties mentioned in its causal role. For there seems to be a tension between the fact that two properties A and B are conceptually linked

and the possibility that the tokening of A could causally explain the tokening of B. Pineda investigates different modes in which this tension could be released. He distinguishes, for instance, between direct and indirect conceptual links, and insists that only the former are at odds with causal connections. Yet, as Pineda himself points out, this is of no avail to our case because the link between a functional property and the properties mentioned in its causal role is conceptually direct.

Stephen Yablo suggests that the problem of causal exclusion derives from a misconception of the conditions under which causes are individuated. The causal efficacy of a determinable, like red, is not necessarily screened off by that of a given determinate, say scarlet. The individuation of causes must respect a principle of proportionality which ensures that, in some circumstances, red may be more proportional to the effect than scarlet. For instance, the redness of the cape is more proportional to the bull's charge than the scarlet shade. Coherently, Yablo claims that pain may in some circumstances be more proportional to the effect that the determinate physical state upon which that instance of pain depends.

But, as Yablo points out, this solution to the exclusion problem may clash with the extrinsicness of mental contents, since it could be argued that the intrinsic surrogate of mental states is always more proportional to the acting than the extrinsically individuated mental content. Yablo replies, however, that this line of reasoning crucially relies on a principle stronger than proportionality and manifestly implausible, namely: the principle of superproportionality. Yablo ends up by outlining some reasons why, in the relevant circumstances, the extrinsically individuated mental states may be more proportional to the acting than its intrinsic surrogates.

Agustin Vicente doubts, however, that the determinable/determinate distinction may help to vindicate the causal efficacy of the mental. He views that distinction as a case of the genus-species relation, where the species is identified as a complex of genera plus differences. It follows that distinct species of the same genus have some internal similarities, and they have some causal powers in common precisely because they have that genus as a part. This explains, according to Vicente, why on some occasions the genus is causally efficacious in detriment of the species: it is that part of being scarlet which is being red that causes the bull's charge. It is clear that there is no such commonality among the different realizations of the same functional property: they needn't be internally similar to each other. Hence, we cannot assimilate the idea of realization to the genus-species relation in this crucial respect and, therefore, the latter cannot help us to vindicate the causal efficacy of the men-

tal, at least insofar as mental properties are construed as second-order properties. This is Vicente's main sceptical point.

Terence Horgan, like Yablo, is convinced that the problem of causal exclusion hangs on a poor understanding of the nature of causal explanation. In particular, Horgan argues that this problem only arises if one disregards a crucial feature of both causation and causal explanation, namely: that although causal explanations (and causation) rest upon objective dependence-patterns among properties, what counts as the relevant dependence-pattern is implicitly contextual, perspective-relative. Philosophical discussion places mental and physical explanations on the same level and easily generates the illusion that the individuation of causes is context-free. Once we are aware of this powerful illusion, we find no metaphysical obstacle in recognizing that the same phenomena can be explained in a variety of ways, that it can be explained by appeal to the properties that figure in distinct patterns of counterfactual dependence. It is important to remark that, according to Horgan, this plurality of causal explanations does not involve overdetermination because the latter is an intra-level notion, whereas distinct patterns of counterfactual dependence would be placed at distinct ontological levels. Such ontological levels are connected by inter-level supervenience relations. Horgan maintains, however, that this supervenience link neither gives rise to a new version of the exclusion problem nor entails explanatory irrealism.

Notes

[1] Interdependence is trivially excluded by the causal closure of the physical world. For the latter guarantees that physical properties by themselves, independently of any other property, completely explain any physical event.

[2] There are, indeed, some other worries which have to do with the opposite intuition, namely, that strong supervenience is too strong. For instance, those which are concerned with the causal relevance of extrinsically individuated properties.

BIBLIOGRAPHY

Kim, J.: 1993, *Supervenience and Mind. Selected Essays*, Cambridge, Cambridge University Press.

Kim, J.: 1989, *Mind in a Physical World*, Cambridge, Mass., The MIT Press.

Josep E. CORBI
Departament de Metafísica i Teoria del Coneixement
Universitat de València
Blasco Ibáñez 30 - 46010 València
E-mail: Josep.Corbi@uv.es

THEORIA

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA
FUNDADA EN 1952 - FUNDADOR: Miguel SANCHEZ-MAZAS (†) - SEGUNDA EPOCA

REDACCION / EDITORIAL OFFICE / REDACTION

ESPAÑA/SPAIN: CALIJ-THEORIA, Alcalde José Elosegui, 275, E 20015, Apartado 1.594, 20080, San Sebastián, España. Tel.: (34 943) 29.17.25 / 28.84.00. Fax: (34 943) 28.06.23. E-mail: theoria@sf.ehu.es
EXTRANJERO/FOREIGN COUNTRIES: Asociación Cultural España-THEORIA, Case 2.730, 1211 Genève-2, Suisse.

DISTRIBUCION / DISTRIBUTION

Para suscripciones, números atrasados y cambios de dirección:/For subscriptions, back volumes and changes of address: Servicio Editorial, Universidad del País Vasco, Apartado 1.397, E 48080 Leioa, España.
Tel.: (34 94) 601.51.26. Fax: (34 94) 480.13.14. E-mail: luxedito@lg.ehu.es

THEORIA es una revista cuatrimestral (sale en Enero, Mayo y Septiembre). Los trabajos que aparecen en esta Revista son registrados y clasificados en las siguientes fuentes bibliográficas: Bulletin Signalétique 519, ICYT e ISOC del Centro de Información y Documentación Científica, Mathematical Reviews, Current Mathematical Publications, MathSci, Philosopher's Index, Répertoire bibliographique de la Philosophie. Las instituciones coeditoras de esta Revista y su Consejo Editor no se identifican necesariamente con los juicios expresados en los trabajos publicados en ella.
Véanse en las últimas páginas, informaciones para colaboradores y subscriptores.

THEORIA is a four-monthly journal (issues in January, May and September). The contents of this Journal are indexed and compiled in the following bibliographic sources: Bulletin Signalétique 519, ICYT e ISOC del Centro de Información y Documentación Científica, Mathematical Reviews, Current Mathematical Publications, MathSci, Philosopher's Index, Repertoire bibliographique de la Philosophie. The publishers and the Editorial Board assume no responsability for any statements of fact or opinion expressed in the published papers.
See at the end of this issue informations for contributors and for subscribers.

¡NUEVO / *NEW*! **www.sc.ehu.es/theoria**

THEORIA

REVISTA DE TEORIA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

SUMARIO / *CONTENTS*

Vol. 16/1, Nº 40, pp. 1-202, Enero/*January* 2001

ISSN 0495-4548

SECCION MONOGRAFICA:
Mental Causation and the Exclusion Problem
Guest Editor: Josep E. Corbí

ARTICULOS / *ARTICLES*

RECENSIONES / *BOOK REVIEWS*

CRONICAS Y PROXIMAS REUNIONES /
NOTICES AND ANNOUNCEMENTS

SECCION MONOGRAFICA

MENTAL CAUSATION AND THE EXCLUSION PROBLEM

Guest Editor: Josep E. CORBI